Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology & Evolution doi: 10.1111/j.2041-210X.2010.00012.x

Simple means to improve the interpretability of regression coefficients

Holger Schielzeth*

Max Planck Institute for Ornithology, Eberhard-Gwinner-Str. 5, 82319 Seewiesen, Germany; and Department of Evolutionary Biology, Evolutionary Biology Center, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden

Summary

1. Linear regression models are an important statistical tool in evolutionary and ecological studies. Unfortunately, these models often yield some uninterpretable estimates and hypothesis tests, especially when models contain interactions or polynomial terms. Furthermore, the standard errors for treatment groups, although often of interest for including in a publication, are not directly available in a standard linear model.

2. Centring and standardization of input variables are simple means to improve the interpretability of regression coefficients. Further, refitting the model with a slightly modified model structure allows extracting the appropriate standard errors for treatment groups directly from the model.

3. Centring will make main effects biologically interpretable even when involved in interactions and thus avoids the potential misinterpretation of main effects. This also applies to the estimation of linear effects in the presence of polynomials. Categorical input variables can also be centred and this sometimes assists interpretation.

4. Standardization (z-transformation) of input variables results in the estimation of standardized slopes or standardized partial regression coefficients. Standardized slopes are comparable in magnitude within models as well as between studies. They have some advantages over partial correlation coefficients and are often the more interesting standardized effect size.

5. The thoughtful removal of intercepts or main effects allows extracting treatment means or treatment slopes and their appropriate standard errors directly from a linear model. This provides a simple alternative to the more complicated calculation of standard errors from contrasts and main effects.

6. The simple methods presented here put the focus on parameter estimation (point estimates as well as confidence intervals) rather than on significance thresholds. They allow fitting complex, but meaningful models that can be concisely presented and interpreted. The presented methods can also be applied to generalised linear models (GLM) and linear mixed models.

Key-words: confidence intervals, generalized linear models, interaction terms, null hypothesis testing, partial correlation coefficients, partial regression coefficients, standard errors, standardized effects sizes

Introduction

Data analysis in ecology and evolution is largely based on the use of linear regression models such as ANOVA, ANCOVA, multiple regression, GLM or mixed models (Quinn & Keough 2002; Faraway 2005; Bolker *et al.* 2009). Linear models involve a response of interest and a set of predictors, possibly including some interactions among input variables. Such models can be used to test null hypotheses about the significance of individual predictors and to estimate effect sizes and their standard errors (Nakagawa & Cuthill 2007; Stephens, Buskirk, & del Rio 2007; Garamszegi *et al.* 2009). In this article, I will advertise the use of centred predictors as a simple means to greatly improve the interpretability of parameter estimates. Furthermore, I will advocate looking at estimates for standardized input variables that are valuable as standardized effect size estimates for between-study comparisons. Standardized estimates are frequently used in other fields but are not in widespread use in ecology and evolution. For simplicity, I will introduce my suggestions for general linear models, but the methods easily generalize to GLM and mixed models (see 'Extensions').

Although the estimation and interpretation of at least some of the estimates and associated P values is possible on the original scale and without centring, centring and scaling of input

*Correspondence author. E-mail: holger.schielzeth@ebc.uu.se

variables and responses has several advantages. First, in the presence of interactions, it enables the interpretation of main effects, which are biologically meaningless otherwise (Engqvist 2005). Second, it enables the estimation of curvature and synergistic effects of continuous predictors that can be interpreted independent of the main effects. Third, they facilitate the interpretation and comparison of the relative importance of predictors within models by looking at the estimates rather than the *P* values (Gelman & Hill 2007). Fourth, they can serve as standardized effect size estimates for between-study comparisons. Furthermore, I will present an easily applied method to extract group mean and group slope estimates and their appropriate standard errors from linear models. Although these points are not new and are well treated in the statistical literature (see, e.g. Aiken & West 1991; Neter *et al.* 1996; Gelman & Hill 2007), they are surprisingly little used in the study of ecology and evolution. The aim of this paper is to encourage the use of standardizations and to give a guideline for parameter interpretation.

Some of the points I raise are indeed a matter of convenience and preference. Whether predictors are interpreted on the original scale or on the standardized scale will depend partly on the system of study. In some contexts, unstandardized effect size estimates may be more easily interpreted than standardized effect size estimates, because the latter depend on the phenotypic variation in each study population. Other points, like the centring of input variables that are involved in interactions, are also not strictly necessary, but are very advisable, since they safeguard against potential misinterpretations. Main effects are not biologically interpretable if involved in interactions without centring the input variables and the same is true for linear terms in the presence of quadratic terms. In many cases, centring will circumvent the need for model simplification, since parameter estimates in complex models can be directly interpreted. Therefore, centring of the input variables will avoid several critical issues and will thus allow fitting complex, but meaningful models. At the same time, it helps putting the focus on parameter estimates rather than *P* values.

Phenotypic standard deviations

Throughout the paper, I will make an important distinction between input variables and predictors. Input variable are the variables that were measured (possibly transformed), while predictors are the terms that are entered in the model (Gelman & Hill 2007). Hence, predictors encompass the main effects, but also polynomials of input variables and interaction terms. Note that one should always transform the input variables and not the predictors (Gelman 2008).

I will refer to 'centring' as subtracting the sample mean from all input variable values. The mean of the centred variable is zero, but the units are still on the original scale. Centring input variables will also result in centred polynomials and in centred interaction terms if the data points are distributed symmetrically around their mean. However, care should be taken in cases of skewed distributions. I will refer to 'scaling' as dividing the input variables by their sample standard deviations. Although scaling can be done without centring, usually scaling will be combined with centring. Hence, I will assume scaled variables to be centred and will refer to these as scaled or standardized input variables.

Standardization converts the original units to units of phenotypic standard deviations. If the sample is representative for the population studied, this is a meaningful measure that is comparable across studies. In the case of approximately normal distributed input variables, *c.* 95% of the values will be within ± 2 units. Hence, standardized variables will typically range from −3·0 to 3·0. Phenotypic standard deviation (and estimates derived from models using them) can easily be reconverted to original units if the means and

Table 1. An example for coding and centring of categorical predictors in a case with four groups and two units per group ($A_{1\text{-}2}$, $B_{1\text{-}2}$, $C_{1\text{-}2}$, $D_{1\text{-}2}$). Columns I_1, I_2, I_3 and I_4 show indicator variables (often called dummy variables) that are used as predictors in the model, while the 'Int' columns (zero for all indicators) are estimated as the intercept. (a) Coding that is implicit when fitting categorical predictors in a linear model with treatment contrasts. Coding can also be done manually and the indicators can be fitted in the model without changing the parameter estimates. The parameters are estimated as the mean of the references group (M_A, estimated as the intercept) and three treatment contrasts to the reference group ($C_{B\text{-}A}$, $C_{C\text{-}A}$, $C_{D\text{-}A}$). (b) After manual coding and centring within indicators, the model estimates one intercept (at an imaginary mean category M_0) and three contrasts that are identical to the implicit coding model ($C_{B\text{-}A}$, $C_{C\text{-}A}$ and $C_{D\text{-}A}$). (c) When removing the intercept, implicit coding results in four indicator variables and effects are estimated as four groups means (M_A, M_B, M_C, M_D). Note that the group means can easily be retrieved from the intercept removed model, while the calculation of group means requires combining intercepts and contrasts in the other two models

	(a) Implicit coding				(b) Centred coding				(c) Intercept removed			
Unit	Int	I_1	I_2	I_3	Int	I_1	I_2	I_3	I_1	I_2	I_3	I_4
A_1	0	0	0	0	0	−0.25	−0.25	−0.25	1	0	0	0
A_2	0	0	0	0	0	−0.25	−0.25	−0.25	1	0	0	0
B_1	0	1	0	0	0	0.75	−0.25	−0.25	0	1	0	0
B_2	0	1	0	0	0	0.75	−0.25	−0.25	0	1	0	0
C_1	0	0	1	0	0	−0.25	0.75	−0.25	0	0	1	0
C_2	0	0	1	0	0	−0.25	0.75	−0.25	0	0	1	0
D_1	0	0	0	1	0	−0.25	−0.25	0.75	0	0	0	1
D_2	0	0	0	1	0	−0.25	−0.25	0.75	0	0	0	1
Estimate	M_A	$C_{B\text{-}A}$	$C_{C\text{-}A}$	$C_{D\text{-}A}$	M_0	$C_{B\text{-}A}$	$C_{C\text{-}A}$	$C_{D\text{-}A}$	M_A	M_B	M_C	M_D

standard deviations of the original variables are known and reported in a paper.

Categorical input variables

Often some input variables are categorical and at first glance it seems impossible to transform them. However, coefficients of categorical predictors are estimated as slopes in linear regression models and, with treatment contrasts, are thus implicitly coded as 0 and 1 (Table 1). It is possible to manually code categorical input variables without changing the results. Manual coding is done by constructing $k - 1$ indicator or dummy variables (where k is the number of levels of the categorical predictor) that take a value of 1 if the record belongs to a specific category and 0 otherwise (see, e.g. Quinn & Keough 2002; Gelman & Hill 2007 for more details on indicator variables). One category is set to 0 for all indicators and hence serves as a reference category that is estimated as the intercept (hence $k - 1$ and not k indicator variables).

After coding, indicator variables can be centred like any continuous predictor (Table 1). Whether are not centring of indicators assists interpretation depends on the study design and research interests (see 'Estimating group means appropriately' for examples, where binary inputs are left uncentred). Centring is particularly advantageous if the levels of categorical input variables can be thought of as random selection from a larger subset of possible categories. Centring indicators will lead to other main effects and the intercept being estimated at this predictor being zero, i.e. at an imaginary mean category (Table 1, Gelman & Hill 2007). Although this seems inappropriate, since there are no observations at this value of the indicator, this makes sense if we are interested in the average effect across different levels of the categorical predictor (Gelman & Hill 2007). For example, the breeding success of some bird species might have been measured under a few environmental conditions that represent a sample of a larger number of possible environments. In this case, the average breeding success might be of more general interest than the estimate for a particular environment. As long as the difference in values for the two categories of an indicator is 1 (e.g. −0·25 and 0·75 as in Table 1), the estimates for this indicator still express the expected change in mean values compared with the reference category (but not the difference to the intercept if the reference category is not coded as zero).

Indicator variables should not be standardized, since this would change the difference between the categories, which will no longer be equal to unity. This would render the estimates almost impossible to interpret. In the following, I will refer to estimates for categorical input variables alternatively as treatment effects.

Interactions between categorical and continuous input variables

There are well-justified warnings against the interpretation of main effects when the input variables are involved in interactions (Aiken & West 1991; Engqvist 2005; Gelman & Hill 2007). This is because main effects are estimated where all other predictors are zero, but in many cases zero is not a meaningful point that lies outside the range of the data (imagine testing for sex differences at a body size of zero). In the presence of interactions, treatment main effects and slopes are usually negatively correlated with each other (if the range of the predictor values is all-positive). This is why treatment main effects often become either spuriously significant when involved in significant interactions (Fig. 1a) or become non-significant even though there is a clear main effect (Fig. 1b).

It has sometimes been argued that the removal of interaction terms is necessary to interpret main effects (Engqvist 2005). However, centring of input variables offers an easy solution to this issue: Centring effectively removes the correlation between slopes and intercepts and makes treatment main effects meaningful independent of the slopes (Fig. 1). This holds true for the interpretation of the estimates and for the t or F tests (i.e. statistical significance) on main effects and interaction. For example, it is possible to conclude that some covariate affects males and females differently *and* that males and females differ on average in their response values. Note, however, that a significant interaction means that main effects are not constant across the whole range of the covariate. Thus, although it is possible to conclude that groups differ *on average*, they do not necessarily differ across the whole range of the covariate (the

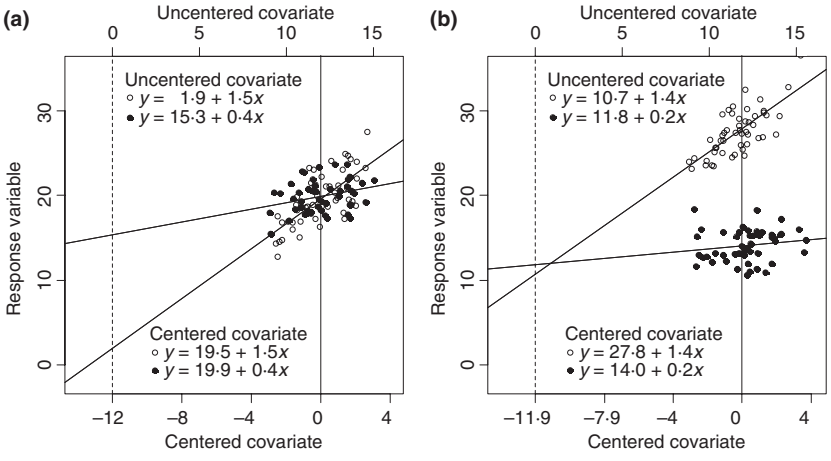

Fig. 1. Two examples of how the main effect estimates for a categorical predictor (filled and unfilled circles) dependent on the values of the covariate x in a simple linear model with an interaction term. The solid and dashed vertical lines show where the main effects are estimated with and without centring of the covariate, respectively. Note that if the covariate is centred, the group main effect becomes meaningful as it is estimated at the average value of the covariate.

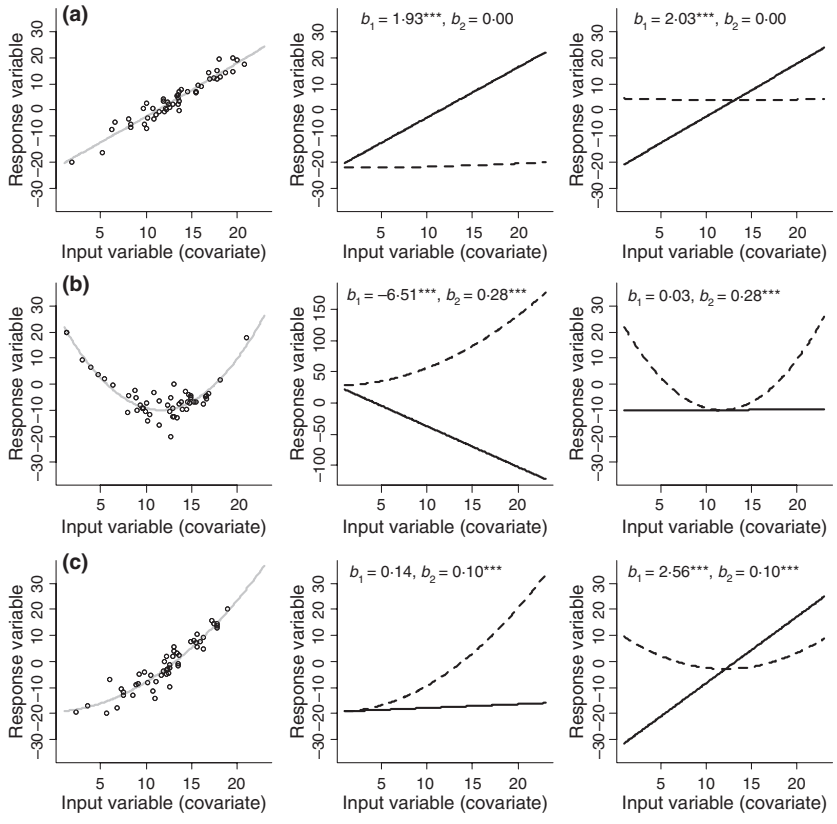

Fig. 2. Three examples of slope estimates for the linear and the quadratic term with and without centring of the input variable. The left column of plots shows the full models with their best-fitting lines (that are identical with and without centring). The middle and right columns show estimates and fitted lines from models without and with centring of the input variables, respectively. Solid lines show the predicted values based on the linear term (b_1), dashed lines those for the quadratic term (b_2). Note that without centring the linear and quadratic term are often negatively correlated with each other (as in the middle row) and the linear term is uninterpretable (e.g. it does not capture the positive trend in the lower row).

Johnson–Neyman procedure can be used to identify regions of significant differences, Johnson & Neyman 1936; Quinn & Keough 2002). Further complications arise if groups differ substantially in their mean covariate values.

Polynomials

Polynomials are interactions between continuous input variables and themselves (Aiken & West 1991). Unlike dedicated functions that will return orthogonal polynomials, squaring raw input variables will usually not result in uncorrelated predictors. The squared and non-squared values usually correlate with each other and the correlation will be very strong if the input variable is all-positive or all-negative (e.g. $r = 0{\cdot}99$ for an input variable drawn at random from a normal distribution with mean = 12 and SD = 4). As a consequence, the linear and the quadratic term will be confounded with each other (as can be seen for example by a high variance inflation factor). The estimate for the linear term is uninterpretable, since the estimate tends to be correlated with the estimate for the quadratic term (negatively so if the range of values is all-positive). Furthermore, the collinearity between the two predictors produces unstable parameter estimates and large standard errors for the linear term (Bowerman & O'Connell 1990; Quinn & Keough 2002; Tabachnick & Fidell 2006).

With a focus on parameter estimation and inference, the main motivation for including polynomials should be to test for curvature in addition to linear effects and it is desirable to estimate linear and curvature effects in the same model (see Arnold & Wade 1984 for the analogous case of linear and nonlinear selection differentials). Independence of the two terms can be easily achieved by centring input variables before squaring them (Gelman 2008; e.g. $r = 0{\cdot}04$ after centring for the same input variable as above). After centring, the estimate for the linear term will express the linear effect (e.g. higher predictor values yield higher response values), while the estimate for the quadratic term estimates if extremes of the distribution elicit higher (positive slopes) or lower (negative slopes) response values on top of possible linear relationships (Fig. 2). Hence, both estimates have a clear interpretation independent of each other. If both terms are positive, the estimated best-fitting curve is a slope that increases in steepness (Fig. 2c). Note, however, that a significant quadratic term is indicative of curvature in the response even without centring of the input variables. In fact, the estimates and significance tests for the quadratic term are identical for the two models (Fig. 2). Hence, it is the interpretability of the linear term that benefits from centring the input variable.

Interactions between continuous input variables

Interactions between continuous input variables are very much like quadratic terms, because the interaction is estimated as the slope of the product of the two input variables (Aiken & West 1991). This can be seen from the basic structure of a linear model with an interaction

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i,$$

where y_i is the observed response value at the ith occasion, x_{1i} and x_{2i} are the values of the two continuous

explanatory input variables at the *i*th occasion, β_0 is the intercept, β_1 and β_2 are regression coefficients for the two main effects, β_3 is the regression coefficient for the interaction and ε_i is the error associated with the *i*th measurement.

Including two continuous input variables and their interactions without centring the input variables will result in an interaction predictors that is collinear with the main effects (e.g. the correlation between raw input variables and the interaction term were $r = 0{\cdot}70$ and $r = 0{\cdot}54$ for two input variables drawn at random from normal distributions with means = 12 and SD = 4). As in the case of non-orthogonal polynomials, the collinearity produces correlated estimates that are essentially meaningless for the main effects, unstable parameter estimates and inflated standard errors for the main effects (Quinn & Keough 2002; Tabachnick & Fidell 2006).

Centring the input variables before fitting the model will largely remove this correlation (e.g. $r = -0{\cdot}12$ and $-0{\cdot}05$ for the same input variables; Aiken & West 1991; Neter *et al.* 1996; Quinn & Keough 2002). This allows testing for interaction effects, i.e. if combinations of extremes produce differential responses on top of what is explained by the sum of the main effects. Purely additive effects will result in significant main effects and a non-significant interaction term, while multiplicative effects will result in the interaction being significant (possibly on top of significant additive main effects). For example, we might want to know if offspring growth in a species with biparental care is influenced by the combination of male and female traits within a pair. The estimate for the interaction term will indicate if the combination of the parent's traits matters for offspring growth on top of the individual effect of the male's and the female's trait value that are estimated as main effects. Without centring, a significant interaction term will also indicate an interaction effect, but the two main effects will not be interpretable, very similar to the non-interpretable linear term in a model with non-centring squared terms.

Comparing the importance of predictors

In linear models with multiple predictors, it is often of interest to judge the importance of individual predictors (Healy 1990; Chao *et al.* 2008). There has been some controversy about how to measure the importance of parameters in linear models, because with correlated predictors there is no unique way to partition the variance in the response (Bring 1996; Johnson 2000). Among the several methods that have been proposed are dominance analysis (Budescu 1993; Azen & Budescu 2003, 2006), Johnson's relative weight (Johnson 2000) and the Pratt's product measure (see Chao *et al.* 2008 for a review). These methods are advantageous if the predictors are correlated. I will here propose standardization of input variables as a simple, but efficient alternative to compare the unique explanatory value of predictors in a linear model that are often measured on different scales (see also Gelman 2008). This is clearly advantageous to compare *P* values, since it puts the focus on effect sizes rather than significance (Gelman 2005; Nakagawa & Cuthill 2007). I will assume the model structure to be known, i.e. the fitted model will contain all influential effects and all predictors included in the model are meaningful and of interested to the researcher. This is a necessary precondition, since the estimates and their standard errors in a linear model are always conditional on the fitted model (Burnham & Anderson 2002).

Regression coefficients in linear models are usually not comparable, because the estimates depend on the variances and these usually differ between input variables. These differences in variances arise for example from input variables being measured in very different units. Hence, given the same predictive value, predictors with low variances (narrow range of values) will have large absolute point estimates whereas predictors with high variance (wide range of values) will have low absolute estimates. A simple way to make continuous predictors comparable within models is to standardize their units to units of standard deviations by scaling all continuous input variables. The effect of every input variable is then measured in units of phenotypic standard deviations of the input variable.

However, categorical indicator variables are not directly comparable with continuous predictors even after standardization of continuous input variables. Their effects will appear large (in absolute values) compared with standardized continuous input variables. This is because the standard deviation of a binary input variable with equal numbers of observations in both groups has a standard deviation of 0·5 (Gelman 2008), whereas standardized continuous input variables have a standard deviation equal to unity. Therefore, Gelman (2008) recommended to standardize continuous input variables by dividing two standard deviations (instead of one) by default. The estimates can then be compared directly between standardized continuous and unstandardized categorical predictors. The standard deviation of binary predictors is close to 0·5 only if the two groups have roughly equal number of observations. If one of the groups has far fewer number of observations, the standard deviation of the binary predictor is substantially less than 0·5, and the regression coefficient will thus appear relatively large, emphasizing the effect of a rare group (Gelman 2008). Nevertheless, the coefficient still has a clear interpretation (change in mean values between the two groups), although a direct comparison of the magnitude of different effects is ambiguous and will depend on whether the rare groups is naturally rare in the population or rare only in the sample that is analysed.

Whether or not one wants to standardize continuous predictors by one or two standard deviations will depend on personal preference and whether or not binary predictors are involved. Given the widespread use of standard deviations as a descriptive statistic in ecology and evolution, I suggest a division by one standard deviation, while keeping in mind that continuous and binary input variables will not be directly comparable. Notably, however, the coefficients for continuous predictors are the same, if continuous input variable and the response are standardized in the same way (by one or two standard deviations). In contrast, the effect of indicators variables will express the change in units of two standard deviations when Gelman's

transformation is applied to the response. This might be harder to interpret than the more familiar unit of one standard deviation.

Standardized slopes as standardized effect sizes

Most textbooks on statistics make the important distinction between correlation and regression (Zar 1999; Quinn & Keough 2002). Correlations are expressed as correlation coefficients r that take values between −1 and 1, while the regression estimates are expressed as slopes b that can take much larger or much smaller values. Importantly, the correlation coefficient is scale-independent whereas the slope is scale-dependent. The scale-independence of r makes it a widely used standardized effect size that can be compared across studies (Nakagawa & Cuthill 2007).

In the presence of multiple predictors, there are two alternative estimates that can be used as standardized effect sizes: the partial correlations coefficient (PCC) $r_{xy|z}$ and the standardized partial regression coefficient (SPRC) $b^*_{xy|y}$. They are both conditional on the other predictors (hence the notation $xy|z$, where z could be one or more covariates). The PCC has been proposed as a standardized effect size for biological research (Nakagawa & Cuthill 2007), whereas SPRCs are used as path coefficients in path analysis (Wright 1918; Sokal & Rohlf 1995; Shipley 2000). The two estimates PCC and SPRC are often very similar and I will discuss the differences below. I argue that the SPRC is often the more interesting value and is better suited as a standardized effect sizes at least when the input variables are uncorrelated with each other. However, before discussing multiple regression analysis it is worth looking at a simple linear regression with one continuous predictor to show that the correlation coefficient r and regression coefficient b can be made numerically equivalent.

In simple linear regression, the slope is estimated as

$$b = r_{xy} \cdot \frac{\sigma_y}{\sigma_x},$$

where r_{xy} is the correlation coefficient between predictor x and response y and σ_x and σ_y are the standard deviations of the predictor and the response, respectively. Therefore, if $\sigma_x = 1$ and $\sigma_y = 1$ as in a regression with standardized input variables and response then b equals r. Like correlation coefficients they can take values between −1 and 1. Since the two estimates are identical, one is as good as the other as a standardized effect size estimate. This is also true for multiple regressions if the predictors are uncorrelated (Bring 1996). In this case, standardized slopes are equal to the bivariate correlation (but not to the partial correlations, see below).

The situation is usually more complicated in the case of multiple regressions. Standardized slopes estimate the effect of the predictor of interest x on the response y, while all other predictors z in the model are being controlled for (i.e. statistically held constant). The resultant estimates are the SPRC that are sometimes called beta coefficients or beta weights and are denoted $b'_{xy|y}$ or $b^*_{xy|y}$ (Mayer & Younger 1976; Sokal & Rohlf

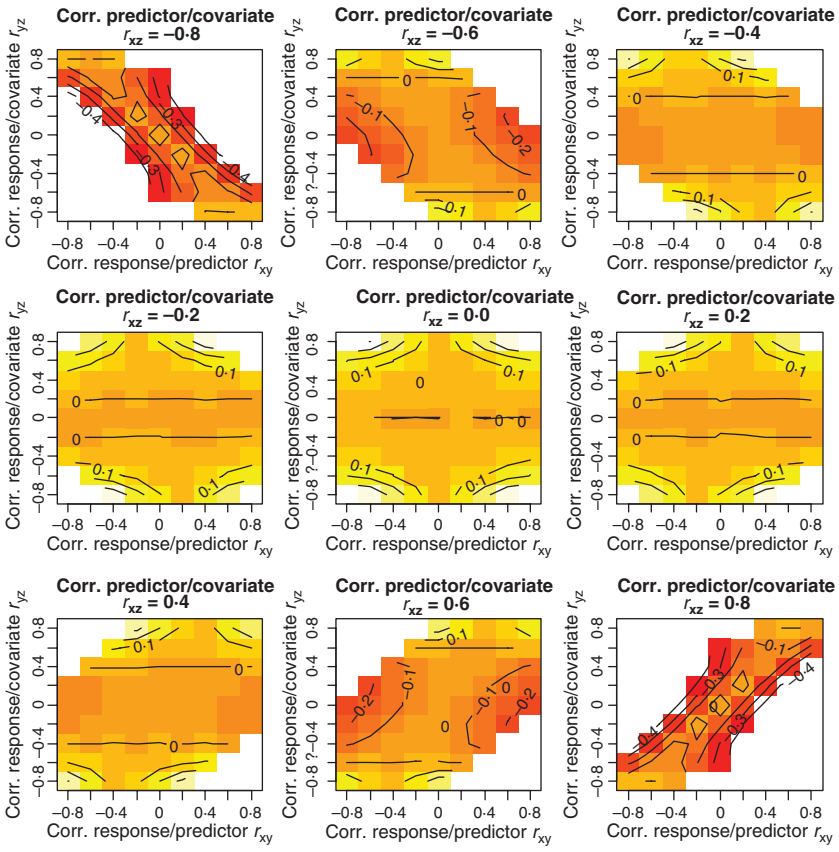

Fig. 3. The absolute difference between partial correlations coefficients (PCC) and standardized partial regression coefficients (SPRC) for the effect of predictor x on the response y estimated from a multiple regression with two continuous predictors (x and a confounding covariate z). Yellow areas show regions in parameter space where the SPRC are smaller (in absolute values), red areas show regions in parameter space where SPRC are larger (in absolute values) compared with the PCC. Note that when x and z are uncorrelated (central plot), SPRC are similar to the PCC over a larger range of values (but smaller when z has a large effect on y), whereas the SPRC tend to be larger than PCC when x and z are correlated.

1995). SPRC are a special case of semi-partial correlation coefficients (Nakagawa & Cuthill 2009) and are not numerically identical to PCC, although the difference between the two estimates is often small. They can be converted into each other by the formula (Sokal & Rohlf 1995) :

$$r_{xy|z} = b^*_{xy|z}\frac{\sqrt{1-r^2_{xz}}}{\sqrt{1-r^2_{yz}}}.$$

As can be seen from this equation, the difference between the two estimates depends on the correlation r_{xz} between the predictor of interest x and confounding covariate z and on the correlation r_{yz} between the response y and z (Fig. 3). If the predictor of interest x and the covariate z are independent (i.e. uncorrelated with each other, hence $r_{xz} = 0$), then the SPRC are always smaller (or equal) in absolute values compared with the PCC (Fig. 3, central plot). This is because the PCC measures the effect of x on y after controlling for the effect of z on x and of z on y (PCC can be calculated as the correlation between the residuals of a regression of x on z and the residuals of a regression of y on z, Nakagawa & Cuthill 2007). In contrast, SPRC measures the effect of x on y, controlling for the effect of z (i.e. statistically holding z constant), but does not statistically remove variation in y explained by z. With correlated predictors and for a given r_{yz}, the PCC will be larger in absolute values compared with the SPRC if $|r_{xz}| < |r_{yz}|$ and smaller in absolute values if $|r_{xz}| > |r_{yz}|$.

The fact that SPRC estimate the expected change in the response in units of the full variation in the response constitutes an advantage of SPRC over PCC. For example, we might be interested in the association between a bird's condition (measured as its body mass or body mass adjusted for body size) and its chick rearing qualities (measured as the weight of its chick). We know that the time of breeding also affects the weight of chicks, but is unrelated to the parent's condition. A SPRC of 0·3 for body condition and 0·6 for time of breeding tell us that changing parent condition by one population standard deviation increases offspring weight by 0·3 population standard deviations and another 0·6 standard deviation for every population standard deviation change in time of breeding. These estimates are equivalent to the bivariate correlations with chick weight (note that it is still worth to include time of breeding in the model, since it reduces residual variation and therefore the standard error for the estimate).

The equivalent PCC of 0·375 would tell us that the correlation between parent condition and chick weight is 0·375 after controlling for time of breeding (i.e. holding time of breeding constant). But since in a natural population there will be variation in breeding time (also in a second population that we might want to compare our results to), it seems more intuitive to interpret the SPRC rather than the PCC. This might be different if variation was artificially induced, e.g. if we had increased the variance in chick weights by offering a range of different food supplies. In this case, another population will not have experienced artificially increased variance in food supply and therefore it will be more interesting to compare PCC, i.e. the correlation after removing excess variation caused by the food supply treatment.

Standardized slopes (used in the comprehensive sense encompassing univariate standardized regression coefficients and SPRC) can be easily converted to raw-scale slopes, if the standard deviations of the raw input variables are known:

$$b_{\text{raw}} = b^* \cdot \frac{\sigma_y}{\sigma_x},$$

where b_{raw} is the slope on the original scale, b^* is the standardized slope and σ_y and σ_x are the raw standard deviation of the input variables. This equation can be rearranged to calculate standardized slopes from unstandardized slopes. Since the standardization is a monotonic transformation, hypothesis tests will be identical independent of whether applied on standardized or unstandardized predictors (Quinn & Keough 2002).

Some objections have been raised against standardized slopes. One important criticism is that they are sample-specific and differences between samples might arise because of differences in the strength of the relationship between x and y or because of differences in the sampling variation (King 1986). The sample-specificity of standardized estimates should not hamper the used of standardized slopes, but should remind us that the presentation of means and standard deviations is necessary for all standardized input variables, since only this information will allow reconverting the estimates to unstandardized slopes. It is also be possible to standardize input variables not by the sample standard deviation, but by some other standard deviation that was estimates from another (larger) sample or derived from theoretical considerations. Whether or not this is sensible will depend on the specific research question asked.

The second major criticism was laid out by Bring (1994). He criticized that the slope estimates in a linear model are conditional on the other predictors in the model, while the standardization is done by dividing by the sample standard deviation that is unconditional on the predictors in the model. He proposed to standardize by the conditional standard deviations, which makes these standardized slopes very similar to PCC (Bring 1994). As I have argued above, the expected change in the response caused by some change in the predictor is often of more interest with respect to the full (unconditional) variation in the response rather than the conditional variation. Therefore, it is an advantage of SPRC that they express the change in the unconditional standard deviation of the response that makes SPRC so easily interpretable.

Estimating group means appropriately

In a linear model with categorical predictors an initial purpose is usually to test for differences between groups. For display, however, we often want to show the individual group means and their appropriate standard errors. This is straightforward for a simple one-way ANOVA, since the means and their standard errors can easily be calculated from the raw data. If we want at the same time control for some confounding effects, we might

want to get the estimates from a linear model (ANCOVA-type with categorical predictors and one or more covariates).

Since in a standard linear model with treatment contrasts, one factor level is used as a reference (and estimated as the intercept), the other factor levels are expressed as contrasts (or differences) to the reference category (Table 1). It is easy to calculate the appropriate group means by adding the contrasts to the intercept for all non-reference groups. However, the standard errors for the contrasts given by the model output are not the appropriate standard errors for the group means. Indeed they are too large, since they depend on the uncertainty in the estimate for the reference category and on the uncertainty in estimating the contrast. The appropriate standard errors for the group means should be calculated as (Sokal & Rohlf 1995):

$$\mathrm{se}(b_j) = \sqrt{\mathrm{se}(b_j - b_1)^2 - \mathrm{se}(b_1)^2},$$

where $\mathrm{se}(b_j)$ is the desired standard error of the estimate for category j, $\mathrm{se}(b_j - b_1)$ is the standard error for the difference of non-reference category j to the references category and $\mathrm{se}(b_1)$ is the standard error of the reference category (the intercept). The values for $\mathrm{se}(b_j - b_1)$ and $\mathrm{se}(b_1)$ are given in the model output.

Alternatively to this calculation, one could fit a second model, this time with the same set of predictors, but with the intercept removed (Gelman & Hill 2007; in R syntax, e.g. lm($y \sim x - 1$), R Development Core Team 2009). If the intercept is removed from the model, standard software will by default make k indicator variables instead of the $k - 1$ contrast indicators (Table 1). Therefore, we will get the groups means and, importantly, the appropriate standard errors for the group means directly from the model output (Table 2). Note, however, that the F test will be meaningless in this case, since it is not testing if there are differences in means between groups, but if the population of group means differs from zero (which it usually will unless the response is centring to zero).

This simple solution can also be used to extract slope estimates and their appropriate standard errors for cases where a categorical predictor interacts with a continuous predictor (Table 2). In a linear model with the two main effects and an interaction, we will get the slope estimate for one reference category and the slope contrasts for the non-references groups. Again, the standard error of the slope contrasts is not the same as the standard error of the group slopes. By removing the main effect of the covariate, we will get slope estimates for each factor level of the categorical predictors and the appropriate standard errors for these slopes (Gelman & Hill 2007). Evidently, the F test for this interaction without the main effect should not be interpreted, since it is testing if the population of slopes differs from zero rather than if treatment groups differ in their slopes.

Worked example

To illustrate the main points of this paper, I will use an empirical data set on reproductive success of zebra finches under aviary conditions with a focus on phenotypic selection (caused by variation in fertilization success) on tarsus length in males. The analysis is simplified for the sake of illustration, a more detailed analysis, including a separation into genetic and environmental selection differentials is in preparation (E. Bolund *et al.* unpublished data). In short, zebra finches were allowed to breed in nine aviaries for a period of about 3 months. There were three sex ratio treatments (SR) with three replicates of each treatment: a female-biased

Table 2. An simulated example for estimating group means (upper row of tables) and group slopes (lower row of tables) from linear models. The left column shows the standard model, with an estimate for a reference category (Group A) and two contrasts. The right column shows a modified model without an intercept (for the group means, upper right) or without the main effect of the covariate (for group slopes, lower right). The data were generated with unequal sample sizes ($n = 30$ for Groups A and B, $n = 20$ for Group C). Note that point estimates and standard errors in the left column estimate the difference to the reference group or reference slope (Group A) and the standard errors for the differences, whereas point estimates and standard errors in the right column estimate the respective group values. Furthermore, note the larger standard errors for the contrasts in the standard model compared with the standard errors of the group means and slopes in the modified model

Standard model

	Estimate	SE	t	P
(Intercept)	9·86	0·40	24·86	< 0·001
groupB	−1·18	0·56	−2·10	0·039
groupC	0·92	0·63	1·47	0·145

	Estimate	SE	t	P
(Intercept)	9·81	0·37	26·89	< 0·001
groupB	−0·82	0·51	−1·61	0·111
groupC	1·80	0·58	3·11	0·003
covar	0·87	0·08	10·81	< 0·001
groupB:covar	0·31	0·11	2·78	0·007
groupC:covar	−0·24	0·12	−1·99	0·051

Modified model

	Estimate	SE	t	P
groupA	9·86	0·40	24·86	< 0·001
groupB	8·68	0·40	21·89	< 0·001
groupC	10·78	0·49	22·20	< 0·001

	Estimate	SE	t	P
(Intercept)	9·81	0·37	26·89	< 0·001
groupB	−0·82	0·51	−1·61	0·111
groupC	1·80	0·58	3·11	0·003
groupA:covar	0·87	0·08	10·81	< 0·001
groupB:covar	1·18	0·08	15·73	< 0·001
groupC:covar	0·63	0·09	7·17	< 0·001

Table 3. Three models (a, b, c) fitted to estimate linear selection differentials on tarsus length (caused by variation in relative fertilization success) in zebra finches held under three different sex ratio (SR) treatments. Significance levels are shown as asterisks for illustration (†$P < 0{\cdot}1$, ***$P < 0{\cdot}001$). For details see text

	(a) Tarsus length uncentred		(b) Tarsus length centred			(c) Tarsus length and SR treatment centred		
	Estimate	SE	Estimate	SE		Estimate	SE	
(Intercept)	−7·29	8·59	1·08	0·23	***	1·00	0·12	***
SR (male-biased)	−2·84	12·09	−0·12	0·29		−0·12	0·29	
SR (female-biased)	0·72	11·79	−0·10	0·31		−0·10	0·31	
Tarsus	0·48	0·50	0·48	0·50		0·53	0·28	†
SR (male-biased) : tarsus	0·16	0·69	0·16	0·69		0·16	0·69	
SR (female-biased) : tarsus	−0·05	0·68	−0·05	0·68		−0·05	0·68	

treatment with nine females and six males in one aviary, an equal sex ratio treatment with six females and six males and a male-biased treatment with six females and nine males. I will here focus on only one breeding season for simplicity (Bolund *et al.*, unpublished data, have analysed two breeding seasons), where every male participated in only one sex ratio treatment (randomly assigned, 63 males in total). Fertilization success was measured as the number of eggs sired by a male and was converted to relative fertilization success (RFS) within aviaries to calculate selection differentials (Arnold & Wade 1984; Brodie, Moore, & Janzen 1995). For brevity, I will refer to tarsus length as TL and the treatment by tarsus interaction as SR : TL. Models are described by the simplified syntax response ~ predictor(s). I explicitly include the intercept (*I*) whenever it is included in the model and will refer to sex ratio treatment contrasts by SRC. Centring input variables are marked with a subscript *C* (TL_C or SRC_C).

In a first step, we test for differences in linear selection differentials between treatments. We fit three alternative models: (a) a standard model with TL and SR contrasts uncentred (RFS ~ *I* + SRC + TL + TL : SRC), (b) a model with TL centred, but SR contrasts uncentred (RFS ~ *I* + SRC + TL_C + TL_C : SRC), and (c) a model with TL and SR contrasts centred (RFS ~ *I* + SRC_C + TL_C + TL_C : SRC_C). Table 3 shows the parameter estimates for these models. Notably, the estimates and standard errors for the interaction are identical in all models. They are all estimating the change in the slope of RFS on TL from equal sex ratio to male-biased and female-biased SR, respectively. For models (a) and (b), the estimate for the TL main effect refers to the SR reference level (equal SR in this case). This is interpretable, since the equal SR treatment is a meaningful level in the analysis. In model (a), however, the estimates for the SR main effects are basically meaningless, since they refer to the effects at TL = 0 (which is out of the range of the data and out of the range of any tarsus length sample). In model (b), the estimates for the main effects of SR are meaningful, since they estimate the effects at a typical (average) tarsus length. All estimates are close to unity, which is not surprising, since RFS is standardized within aviaries. In model (c), we learn something subtly different: the expected change in RFS per millimetre change in TL is 0·53, which is

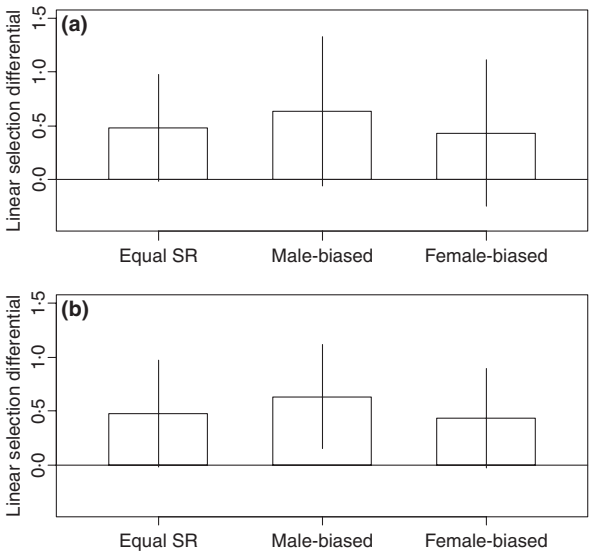

Fig. 4. Linear selection differentials on tarsus length in zebra finches held under three sex ratio (SR) treatments. (a) Incorrect standard errors as extracted from model fitting one main effect and two contrasts. (b) Correct standard errors as extracted from a model fitting no tarsus length main effects, but three tarsus length by sex ratio treatment interaction terms. For details see text.

the expected change at an average SR. Note that the two slope contrasts (interaction terms) still give the expected change between equal SR and male-biased and female-biased as before. Similarly, the intercept now estimates the effect at an average SR, while the SR main effects still estimates the change between equal SR and male-biased and female-biased, respectively. The intercept and the tarsus main effect estimates of model (c) are informative, if we consider the three treatments as a (representative) subset of a larger number of possible treatments and are interested in the average selection differentials. (In this particular example, it might be sensible to centre the SRC at 1/3 rather than at their sample standard deviations, if one wants to give all SR equal weights.)

In a second step, we want to display the selection differentials in a bar chart. The intuitive, but inappropriate way is to fit a model (a) or (b) and to add the contrast to the effects and display them together with their standard errors (Fig. 4a). The

Table 4. Three models (e, f, g) fitted to estimate linear and nonlinear selection differentials on tarsus length in zebra finches held in a female-biased sex ratio treatment. Significance levels are shown as asterisks for illustration (†$P < 0{\cdot}1$, ***$P < 0{\cdot}001$). For details see text

	(e) Tarsus length uncentred		(f) Tarsus length centred			(g) Tarsus length uncentred		
	Estimate	SE	Estimate	SE		Estimate	SE	
(Intercept)	−49·56	91·89	1·01	0·12	***	−6·57	3·70	†
Tarsus (linear)	5·38	10·57	0·41	0·22	†	0·43	0·21	†
Tarsus (non-linear)	−0·14	0·30	−0·14	0·30				

Table 5. A full model fitted to estimate linear and nonlinear selection differentials on tarsus length in zebra finches held under three different sex ratio (SR) treatments. For details see text

	Tarsus length standardized			
	Estimate	SE	*t*	*P*
SR (equal)	1·07	0·27	3·94	0·0002
SR (male-biased)	1·00	0·24	4·24	0·0001
SR (female-biased)	1·01	0·27	3·76	0·0004
SR (equal) : tarsus (linear)	0·22	0·30	0·73	0·47
SR (male-biased) : tarsus (linear)	0·27	0·21	1·28	0·21
SR (female-biased) : tarsus (linear)	0·17	0·21	0·84	0·40
SR (equal) : tarsus (nonlinear)	0·01	0·19	0·07	0·95
SR (male-biased) : tarsus (nonlinear)	−0·06	0·21	−0·30	0·76
SR (female-biased) : tarsus (nonlinear)	−0·03	0·12	−0·21	0·83

appropriate alternative is to calculate the standard errors according to the formula given above or to fit a model (d) with the tarsus main effect removed (RFS $\sim I$ + SRC + TL : SRC). The estimates and standard errors can be directly used for display (Fig. 4b). As can be easily seen from this comparison the standard errors are too large for the two contrast categories when model (a) is used without appropriate standard error calculation.

In a third step, we want to calculate nonlinear selection differentials in addition to the linear differentials. Nonlinear selection differentials estimate curvature effects (Arnold & Wade 1984; Brodie *et al.* 1995). I will illustrate this for the female-biased treatment only. A simple model (e) will include tarsus and tarsus squared (RFS $\sim$ TL + TL^2), while an improved model (f) will include tarsus and tarsus squared after centring (RFS $\sim$ TL_C + $TL_C{}^2$). Table 4 shows the estimates from these two model as well as from a model (g) with only the linear effect (RFS $\sim$ TL_C). Note that the two estimates for the nonlinear effects are identical, but model (f) estimates for the linear effect are not biologically interpretable, while the estimates from model (g) and model (h) are very similar for the linear effect. Hence, with prior centring it is possible to estimate linear and nonlinear effects simultaneously, which is also the standard method in selection analysis (Arnold & Wade 1984; Brodie *et al.* 1995).

Finally, standardization allows the estimation of standardized selection differentials that are comparable across studies (e.g. for species with different tarsus lengths). In this case, relative fertilization success is naturally standardized (though not centred to zero), so that we standardize only TL. Table 5 shows the results of a comprehensive model that estimates standardized linear and nonlinear selection differentials for all SR treatment levels. This shows the overall very similar selection differentials in all treatments with linear selection differentials estimated positive in all treatments (although large standard errors show that they are not significantly different from zero) and very low and non-significant nonlinear selection differentials.

Extensions

Most of the points made for general linear models generalize easily to linear mixed models (LMM) and mostly also to GLM and GLMM. These models differ from the general linear models addressed so far by their link functions and error distributions (GLM and GLMM) and/or the presence of random effects (LMM and GLMM). Centring remedies the issue of interpreting main effects in the presence of interactions independent of these changes and scaling of input variables also makes the estimates comparable within models for GLM, LMM and GLMM.

However, for LMM, the use of standardized slopes as standardized effect sizes might be slightly different from what I have described above. This is because the SPRC expresses the change in the response variable relative to the total variation in the response. Depending on whether the predictor of interest is a between-group or a within-group predictor (see van de Pol & Verhulst 2006; van de Pol & Wright 2009 for methods to separate between- and within-group variation), it might be more interesting to rescale the variance in the response to the between- or within-group variance. This can be done by first fitting a model with only group-specific random intercepts. The required between-group or within-group (residual) standard deviation (σ_α or σ_ε, respectively) can be extracted from this model and can be used to rescale the response. In a second step, a full model can be fitted with the response standardized by its within-group standard deviation (within-group predictor) or by its between-group standard deviation (between-group predictor). Notably, the estimates are no longer standardized partial regression coefficients but will appear larger than SPRC, since the response is scaled to a standard

deviation larger than unity. However, slope estimates will be qualitatively equivalent to SPRC calculated from group means (standardization by the between-group standard deviation) or deviations from group means (standardization by the residual standard deviation). I am not aware of any study that has done so, but if a clear separation of within- and between-group predictors becomes more popular, this might be an efficient way to obtain comparable standardized effect size estimates.

If the focus is on within-group predictors in a LMM, there might be between-group variation in slopes that makes it necessary to fit random-slope models rather than random-intercept models (Schielzeth & Forstmeier 2009). To make the random-slope variance comparable with the random-intercept variance, it has been suggested to use standardized input variables, which implies to estimate the variance on the scale of standardized slopes (Nussey, Wilson, & Brommer 2007; Schielzeth & Forstmeier 2009). The between-group random-slope variance of standardized slopes can be interpreted as the distribution of individual slopes around the population mean slope (which is the fixed effect estimate of the slope). Hence, the use of standardized input variables and responses make the random-slope variance scale-independent and comparable across studies. This standardization of random-slope variances constitutes yet another advantage of using standardized input variables.

Acknowledgements

I am grateful to Elisabeth Bolund, Wolfgang Forstmeier, Roger Mundry, Shinichi Nakagawa and to an anonymous referee for their critical comments and discussion. Elisabeth Boland provided empirical data.

References

Aiken, L.S. & West, S.G. (1991) *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications, Newbury Park.

Arnold, S.J. & Wade, M.J. (1984) On the measurement of natural and sexual selection: theory. *Evolution*, **38**, 709–719.

Azen, R. & Budescu, D.V. (2003) The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, **8**, 129–148.

Azen, R. & Budescu, D.V. (2006) Comparing predictors in multivariate regression models: an extension of dominance analysis. *Journal of Educational and Behavioral Statistics*, **31**, 157–180.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, **24**, 127–135.

Bowerman, B.L. & O'Connell, R.T. (1990) *Linear Statistical Models: An Applied Approach*, 2nd edn. Duxvury Press, Belmont, CA.

Bring, J. (1994) How to standardize regression coefficients. *American Statistician*, **48**, 209–213.

Bring, J. (1996) A geometric approach to compare variables in a regression model. *American Statistician*, **50**, 57–62.

Brodie, E.D., Moore, A.J. & Janzen, F.J. (1995) Visualizing and quantifying natural selection. *Trends in Ecology & Evolution*, **10**, 313–318.

Budescu, D. V. (1993) Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, **114**, 542–551.

Burnham, K.P. & Anderson, D.R (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer, Berlin.

Chao, Y.C.E., Zhao, Y., Kupper, L.L. & Nylander-French, L.A. (2008) Quantifying the relative importance of predictors in multiple linear regression analyses for public health studies. *Journal of Occupational and Environmental Hygiene*, **5**, 519–529.

Engqvist, L. (2005) The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. *Animal Behaviour*, **70**, 967–971.

Faraway, J.J. (2005) *Linear Models in R*. Chapman & Hall/CRC, Boca Raton, FL, USA.

Garamszegi, L.Z., Calhim, S., Dochtermann, N., Hegyi, G., Hurd, P.L., Jørgensen, C., Kutsukake, N., Lajeunesse, M.J., Pollard, K.A., Schielzeth, H., Symonds, M.R.E. & Nakagawa, S. (2009) Changing philosophies and tools for statistical inferences in behavioral ecology. *Behavioral Ecology*, **20**, 1376–1381.

Gelman, A. (2005) Analysis of variance: why it is more important than ever. *Annals of Statistics*, **33**, 1–31.

Gelman, A. (2008) Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, **27**, 2865–2873.

Gelman, A. & Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK.

Healy, M.J.R. (1990) Measuring importance. *Statistics in Medicine*, **9**, 633–637.

Johnson, J.W. (2000) A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, **35**, 1–19.

Johnson, C.R. & Neyman, J. (1936) Tests of certain linear hypothesis and their appplication in some educational problems. *Statistical Research Memoirs*, **1**, 57–93.

King, G. (1986) How not to lie with statistics: avoiding common mistakes in quantitative political science. *American Journal of Political Science*, **30**, 666–687.

Mayer, L.S. & Younger, M.S. (1976) Estimation of standardized regression coefficients. *Journal of the American Statistical Association*, **71**, 154–157.

Nakagawa, S. & Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, **82**, 591–605.

Nakagawa, S. & Cuthill, I.C. (2009) Corrigendum. *Biological Reviews*, **84**, 515.

Neter, J., Kutner, M.H., Nachtsheim, C.J. & Wasserman, W. (1996) *Applied Linear Statistical Models*, 4th edn. Irwin, Chicago, IL, USA.

Nussey, D.H., Wilson, A.J. & Brommer, J.E. (2007) The evolutionary ecology of individual phenotypic plasticity in wild populations. *Journal of Evolutionary Biology*, **20**, 831–844.

van de Pol, M. & Verhulst, S. (2006) Age-dependent traits: a new statistical model to separate within- and between-individual effects. *American Naturalist*, **167**, 766–773.

van de Pol, M.V. & Wright, J. (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. *Animal Behaviour*, **77**, 753–758.

Quinn, G. P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, UK.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schielzeth, H. & Forstmeier, W. (2009) Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*, **20**, 416–420.

Shipley, B. (2000) *Cause and Correlation in Biology: A User's Guide to Path Analysis, Strutural Equations and Causal Inference*. Cambridge University Press, Cambridge, UK.

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd edn. W.H. Freeman and Company, New York, NY, USA.

Stephens, P.A., Buskirk, S.W. & del Rio, C.M. (2007) Inference in ecology and evolution. *Trends in Ecology & Evolution*, **22**, 192–197.

Tabachnick, B.G. & Fidell, L.S. (2006) *Unsing Multivariate Statistics*, 4th edn. Allyn & Bacon, Boston, MA, USA.

Wright, S. (1918) On the nature of size factors. *Genetics*, **3**, 367–374.

Zar, J.H. (1999) *Biostatistical Analysis*, 4th edn. Prentice Hall, Upper Saddle River, NJ, USA.

Received 16 November 2009; accepted 4 January 2010
Handling Editor: Robert P. Freckleton