Women have Relatively Larger Brains than Men: A Comment on the Misuse of General Linear Models in the Study of Sexual Dimorphism

WOLFGANG FORSTMEIER*
Max Planck Institute for Ornithology, Eberhard-Gwinner-Str., Seewiesen, Germany

ABSTRACT

General linear models (GLM) have become such universal tools of statistical inference, that their applicability to a particular data set is rarely questioned. These models are designed to minimize residuals along the y-axis, while assuming that the predictor (x-axis) is free of statistical noise (ordinary least square regression, OLS). However, in practice, this assumption is often violated, which can lead to erroneous conclusions, particularly when two predictors are correlated with each other. This is best illustrated by two examples from the study of allometry, which have received great interest: (1) the question of whether men or women have relatively larger brains after accounting for body size differences, and (2) whether men indeed have shorter index fingers relative to ring fingers (digit ratio) than women. In depth analysis of these examples clearly shows that GLMs produce spurious sexual dimorphism in body shape where there is none (e.g. relative brain size). Likewise, they may fail to detect existing sexual dimorphisms in which the larger sex has the lower trait values (e.g. digit ratio) and, conversely, tend to exaggerate sexual dimorphism in which the larger sex has the relatively larger trait value (e.g. most sexually selected traits). These artifacts can be avoided with reduced major axis regression (RMA), which simultaneously minimizes residuals along both the x and the y-axis. Alternatively, in cases where isometry can be established there are no objections against and good reasons for the continued use of ratios as a simple means of correcting for size differences. Anat Rec, 294:1856–1863, 2011. © 2011 Wiley-Liss, Inc.

Key words: analysis of covariance; geometric mean regression; line-fitting methods; model 2 regression; ratios; reduced major axis; standardized major axis; 2D:4D

The application of general linear models (GLM) has become such a widely accepted and universal tool for statistical inference, that many readers will be surprised to hear about problems associated with it. The purpose of this commentary is to remind the reader about a major limitation of GLMs, which are based on ordinary least square regression (OLS, also known as Model I regression: Sokal and Rohlf, 1995). The limitations of OLS regression are well described in the specialized literature (e.g., Warton et al., 2006; Arnold and Green, 2007; Smith, 2009; Peig and Green, 2010), but there is still a lack of awareness among researchers as evidenced by the frequent misuse of OLS regression in the empirical literature (see e.g., Green, 2001; Peig and Green, 2010). The problem I want to highlight is best illustrated with examples from the study of allometry (i.e., relative size or body shape). I will demonstrate that GLMs can lead to erroneous conclusions in the study of sexual dimorphism, but the problem is more general and

*Correspondence to: Wolfgang Forstmeier, Max Planck Institute for Ornithology, Eberhard-Gwinner-Str., Seewiesen 82319, Germany. Fax: +49-8157-932400. E-mail: forstmeier@orn.mpg.de

Received 7 February 2011; Accepted 26 April 2011
DOI 10.1002/ar.21423
Published online 28 September 2011 in Wiley Online Library (wileyonlinelibrary.com).

extends to a wide range of cases, namely it arises whenever GLMs contain two predictors that are correlated (e.g., sex and body size in a size dimorphic species) and at least one of the predictors contains considerable amounts of statistical noise (which I will define later). Hence, my examples will focus on the issue of how to deal with body size differences (e.g., between sexes or between groups of species), but those examples illustrate a much more general issue.

The study of relative size differences (allometry) is a minefield of possible mistakes to be made. For instance, Smith (2005) presents a very insightful review of 2 decades of research literature that was often mislead by a misunderstanding, namely the failure to distinguish between (1) the study of relative size by means of ratios or reduced major axis regression (RMA, explained later) and (2) the statistical control for differences in size by means of OLS regression.

1. The study of relative size asks whether body proportions remain the same as body size increases (isometry). This has frequently been studied by looking at ratios, for example, brain mass divided by body mass. This is a very natural way of looking at body shape because it directly relates to the Intercept Theorem of elementary geometry, which deals with the equality of ratios of distances. However, the use of such ratios got criticized very heavily for a variety of reasons (e.g., Packard and Boardman, 1988; Ranta et al., 1994). First, many traits scale allometrically rather than isometrically to size, and ratios cannot account for this allometry. Second, ratios still show a (typically negative) correlation with the denominator (e.g., body mass), which seemingly they mean to control for. However, such remaining correlations with measures of size are not always problematic or unwanted. For instance, relative body fat content (mass of fat divided by total mass) only remains a sensible measure of fatness as long as it remains correlated with body mass (here positively, i.e., a 300 kg person scoring higher than a 40 kg person of the same height).
2. Statistical control for differences in body size is achieved by taking residuals from an OLS regression of the trait of interest over body size, or by fitting body size as a covariate into a GLM (or specifically ANCOVA). This approach has the statistically convenient property of removing any residual correlation with the particular measure of size that was controlled for. However, as I will show later, its apparent success in rendering measurements independent of size is illusionary (see also Green, 2001).

THE DISTINCTION BETWEEN OLS AND RMA REGRESSION

For a better understanding, it first needs to explain the critical distinction between the very widely used OLS regression and the relatively unknown reduced major axis (RMA) regression (also named standardized major axis regression or geometric mean regression: McArdle, 1988; Sokal and Rohlf, 1995; Warton et al., 2006; Smith, 2009).

OLS regression is the type of regression that everyone is familiar with: a regression line is fitted to the data such that the deviation of data points along the *y* axis (dependent variable) from the line is minimized. This method of line fitting assumes that all deviations of data points from the regression line are due to statistical noise in the dependent variable *y*, whereas the *x* axis is free of such noise (Sokal and Rohlf, 1995). The violation of this assumption is unproblematic when regression is only used to measure the strength of a relationship (in terms of r^2 or *P*), and also when our purpose is to predict values of *y* from measured values of *x* (Warton et al., 2006). However, when the slope (*b*) of the regression line is important for statistical inference, then it is easy to be misled (see also Smith, 2009; for a detailed discussion). In studies of allometry, the *x* axis is a morphometric trait that shows both measurement error (1) and, more importantly, biological noise (2). (1) Measurement error refers to imprecise measuring of individuals, or, when data points represent species, this additionally includes sampling error, that is, noise due to picking one individual but not another. (2) Biological noise refers to the sum of all genetic and environmental effects that are specific to an individual (or species) and that have uncorrelated effects on *x* and/or *y*. It is the sum of all these effects that make individual data points deviate from a regression line of *y* over *x*. In studies of allometry, the *x* and *y* variables share much of the same properties and can be exchanged arbitrarily. If it seems probable that the deviation of data points from the regression line is due to equal amounts of statistical noise in *x* and *y*, then major axis (MA) regression or reduced major axis (RMA) regression are the methods of choice for line fitting (Legendre and Legendre, 1998; Warton et al., 2006; Smith, 2009). If the amount of statistical noise in *x* and *y* is equal in absolute amounts then MA is preferable, whereas if they are equal in terms of making up equal proportions of the total variations in *x* and *y*, then RMA is preferable. Both methods fit the regression line such that the deviations of data points from the line are simultaneously minimized along both the *x* and the *y* axis.

Note that the two components that make up the statistical noise have different properties. Although measurement error in *x* and *y* can be quantified by taking repeated measurements, the relative amounts of biological noise in *x* and *y* are often impossible to know (Smith, 2009). Hence, when *x* and *y* have very similar properties, as in the below example of the lengths of the 2nd and 4th fingers, then the assumption of RMA of equal proportions of noise in *x* and *y* seems most justified. If, in contrast, the *x*-variable is free of biological noise but only contains a quantifiable amount of measurement error, then there are more specific methods to account for this (see Warton et al., 2006; Freckleton, 2011).

OLS regression lines coincide with RMA regression lines only when all data points are exactly on the line ($r^2 = 1$). With increasing scatter around the line ($r^2 < 1$) OLS regression lines become increasingly shallower than RMA lines (smaller slopes *b* in absolute values), a phenomenon that Galton (1886) termed “regression back to the mean,” an expression that gave regression its name. The reason for this behavior of OLS regression lines lies in the fact that only the deviation along the *y* axis is minimized and that, as scatter increases, the best prediction of *y* from values of *x* approaches the population mean of *y*. The shallow regression line of an OLS regression produces very large residuals along the x axis. A RMA regression, in contrast, minimizes the squared residuals along both axes and this is best achieved by a steeper regression line. More specifically, the slope of a RMA regression line is simply the ratio of the standard deviations of the data along the y and the x axis, respectively (b = SDy/SDx), whereas in OLS regression this is multiplied by the correlation coefficient [b = r * (SDy/SDx)]. Note that slopes of RMA regressions become meaningless in cases where x and y are unrelated ($r \approx 0$), which may also bear a danger of leading to misinterpretation (Smith, 2009).

The failure to distinguish between OLS and RMA regression has probably lead to numerous mistakes in the scientific literature of which I would like to point out two very prominent examples that will hopefully stick in the mind of the reader and lead to a reduction in the number of such mistakes in the future. Both examples are cases of misuse of OLS regression because the widespread use of OLS methods makes such cases abundant. RMA regressions may also be misused or misinterpreted (Smith, 2009), but since the less conventional method is rarely applied without careful consideration, such cases are probably exceedingly rare.

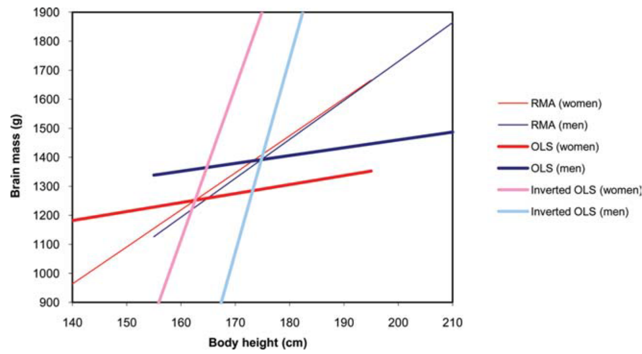

Fig. 1. Different regression lines relating brain mass to body height in the two sexes. OLS regression lines (bold red and bold dark blue) show the shallow slopes characteristic for low r^2 values. When body height is regressed over brain mass, OLS regression lines are also shallow, and hence they become very steep when plotted in this figure where the axes are exchanged (inverted OLS lines in bold pink and bold light blue). RMA regression lines (thin lines) have slopes of intermediate steepness (geometric mean of the two OLS slopes). Note that for each sex, the three regression lines pass through the respective centre of data (mean body height and mean brain mass).

EXAMPLE 1: INTRODUCING SPURIOUS DIMORPHISM

Back in 1992, Davison Ankney, who apparently had read Packard and Boardman's (1988) critique on the use of ratios, adopted their OLS regression methodology (ANCOVA) to reanalyze data on relative brain sizes of men versus women. The data had been collected by Ho et al. (1980) and analyzed in terms of the calculation of ratios (e.g., brain mass divided by body mass). From the results of the new analysis (which was inappropriate due to the use of OLS rather than RMA regression), Ankney (1992) spectacularly concluded that for any given body height, men had ~100 g heavier brains (and by inference an expectation of greater intelligence: Rushton and Ankney, 1996) than women. This politically, biologically, and statistically questionable conclusion caused a rush of excitement throughout the scientific community as well as the public, as evidenced by a series of correspondences published in Nature. Although most critics accepted Ankney's statistical analysis and instead focused on the inappropriateness of setting up expectations about intelligence, it apparently was only Dolf Schluter (1992) who spotted the statistical problem. Probably due to the limited space, he did not focus on the critical distinction between OLS and RMA regression, but instead pointed out that the opposite of Ankney's conclusion follows when x and y axes are exchanged. Schluter also used OLS regression to show that, when regressing body heights over brain masses, one comes to the conclusion that, for a given brain mass, women were on average actually 10 cm shorter than men, meaning they had relatively larger brains. Apparently Ankney did not see or was not convinced by Schluter's comment, so he continued following his own conclusion, and went on asking why females did not score lower on intelligence tests than men despite them having the relatively smaller brains (Rushton and Ankney, 1996). Also, it appears to me that Ankney's but not Schluter's conclusion stuck to many minds of the public, so there is still a need to rectify this issue. As far as I am aware, the data of Ho et al. (1980) was never reanalyzed appropriately using RMA regression. Although the raw data of Ho et al. (1980) is not available, it is still possible to draw the respective OLS and RMA regression lines for both sexes (Fig. 1) by extracting all the necessary parameters from the paper. The bold red and dark blue OLS regression lines of brain mass over body height (taken from Ankney, 1992) are very shallow, reflecting the very low r^2 values (around 5%). This situation with much scatter exaggerates the apparent sexual dimorphism, leading to the conclusion of male brain mass being about 100 g greater than female brain mass for a given body height. Schluter's inversed OLS regression lines (taken from Schluter, 1992) for regressing body height on brain mass are also indicated (bold pink and light blue). When turning the page by 90° they beckon the opposite conclusion: for any given brain mass, men are about 10 cm taller than women, hence women have relatively larger brains. The according RMA regression lines can be computed as follows. First, they have to pass through the average body height (women 162.6 cm, men 174.8 cm) and average brain mass (women 1,252 g, men 1,392 g) values for the two sexes, respectively. Second, their slope is the geometric mean (square-root of the product) of the two OLS regression slopes (hence the alternative name "geometric mean regression," Sokal and Rohlf, 1995). For instance, according to the bold red line (OLS, women) brain mass increases by 3.1 g/cm of body height, and the bold pink line (Inverted OLS, women) represents an increase by 52.6 g/cm of body height, the geometric mean of which is 12.8 g/cm [note that 52.6 = (1/r)*12.8 and 3.1 = r*12.8, with r = 0.243]. The best way to think about a RMA equation is that it represents the intermediate between two OLS equations, one for regressing y over x and one for regressing x over y. It is thereby the solution on which the two OLS equations would converge if the correlation coefficient r approached unity (Smith, 2009). The RMA slope for men is marginally steeper (13.4 g/cm)

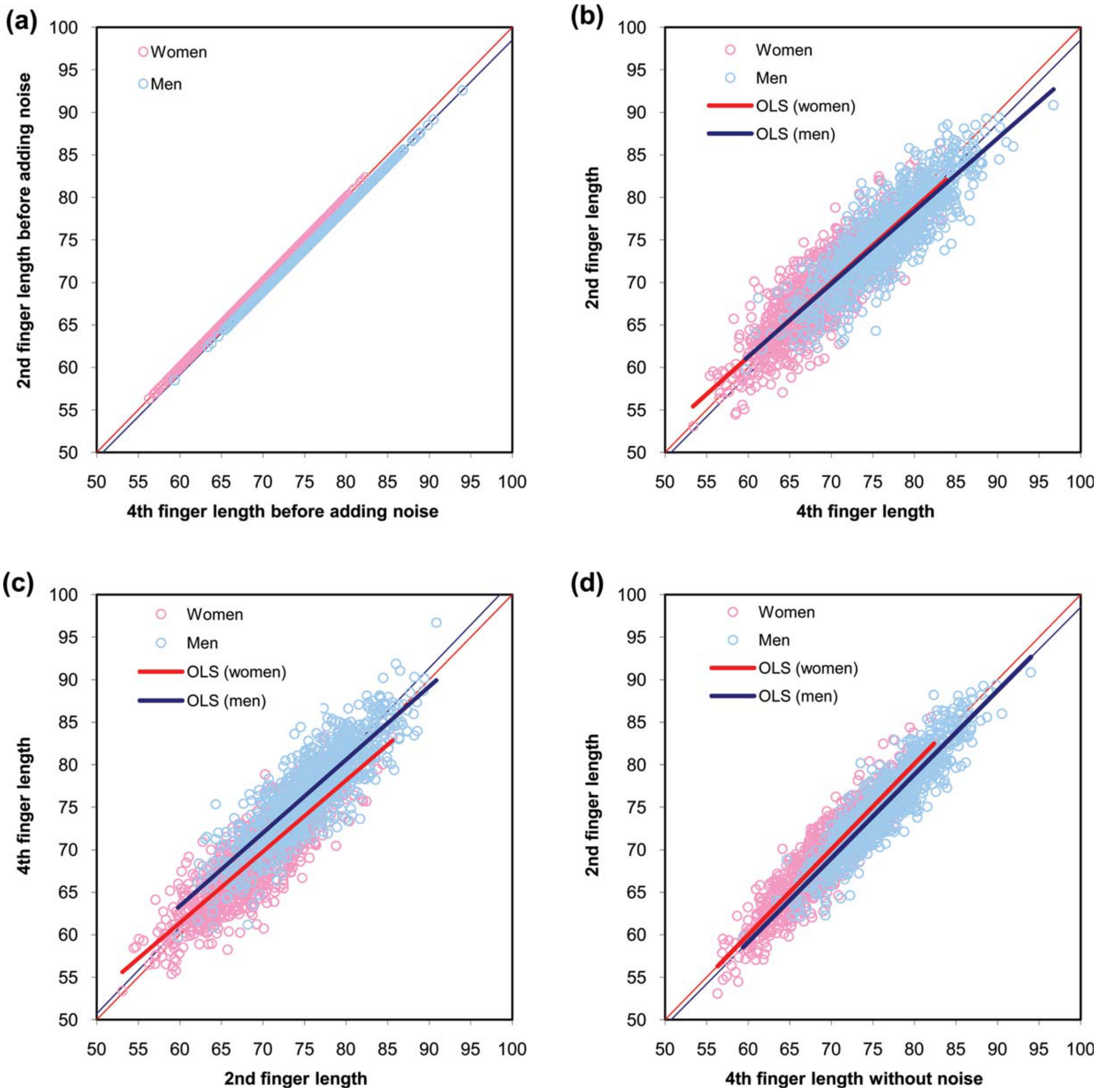

Fig. 2. Panel (**a**) shows how the random data were generated along the lines of isometry for females ($y = x$) and males ($y = 0.985x$), respectively. Panel (**b**) shows the relationship between 2nd and 4th finger lengths after adding noise along both the *x*- and the *y* axis to the data from (a). Separate OLS regression lines are shown for the two sexes (bold) as well as the initial lines of isometry (thin lines). Panel (**c**) shows the same data set as in (b) but with *x* and *y*-axes exchanged. Note that in this inverted data set, the male isometry line (thin blue line) is steeper ($y = x/0.985 = 1.01523x$) than the female isometry line (thin red line). Again, separate OLS regression lines are shown for the two sexes (bold). Panel (**d**) shows the 2nd finger length [*y* axis of (b)] over the initial 4th finger length before the noise was added [*x* axis of (a)]. Note that in this case, when the *x* axis contains no error, OLS regression lines coincide with the lines of isometry.

than that for women. At an average human body height of 168.7 cm, women's brains are on average about 20 g heavier (and not 100 g lighter) than that of men.

Although this looks like a female "victory" in terms of relative brain size in this particular data set, it has to be mentioned that this analysis still suffers from a number of shortcomings, that cannot be resolved without access to the original data and that are slightly beyond the scope of this commentary. Actually, fitting an appropriate regression line is far from trivial, and the interested reader is referred to Warton et al. (2006) for a more detailed treatment of the subject. The biggest flaw to be considered is the difference in dimensionality between body height (linear) and brain mass (volumetric). As a consequence of this, the relationship is expected to be curvilinear, which can be remedied by proper transformations (e.g., by using the cube-root of brain mass or by log-transformation of both x and y). The difference in dimensionality may also provoke doubts whether the assumption of RMA (equal biological noise and measurement error in x and y) are really met precisely. Note that regressing brain mass over body mass would solve the issue of dimensionality at the cost of

TABLE 1. GLMs to the data sets shown in Fig. 2b–d, respectively

Referring to	Dependent	Predictor	B	SE	*t*	*P*
Fig. 2b	2nd finger	Intercept	9.183	0.900	10.2	$<10^{-23}$
		4th finger	0.866	0.012	74.4	$<10^{-99}$
		Sex (female)	0.169	0.153	1.1	0.27
Fig. 2c	4th finger	Intercept	12.628	0.870	14.5	$<10^{-45}$
		2nd finger	0.849	0.011	74.4	$<10^{-99}$
		Sex (female)	−2.244	0.143	−15.6	$<10^{-52}$
Fig. 2d	2nd finger	Intercept	−0.803	0.718	−1.1	0.26
		4th finger without noise	0.996	0.009	107.1	$<10^{-99}$
		Sex (female)	1.107	0.117	9.5	$<10^{-20}$

Note that the sex-by-covariate interaction was always non-significant, and hence it was always excluded for simplicity. In contrast to this, different slopes for males and females were fitted in Fig. 2.

another problem that was highlighted by Arnold and Green (2007). Since brain mass is a part of total body mass, *y* is not being regressed over *x* but rather over *x*+*y*, which leads to a correlated error structure and up-wards biased slope estimates.

EXAMPLE 2: MAKING NEGATIVE SIZE DIMORPHISM DISAPPEAR

Over the past decade there has been a great deal of interest (>300 publications: Voracek and Loibl, 2009) in the study of digit ratio (i.e., the relative length of the index finger over that of the ring finger), because this peculiar morphological trait has been found to be sexually dimorphic, to correlate with all sorts of sex-hormone related behavioral and physiological traits, and, following various lines of evidence, has been suggested to reflect sex-hormone exposure during embryo development (Manning, 2002). Recently, Kratochvíl and Flegr (2009) claimed that all or most of this might be caused by a statistical artifact of using a ratio that is not size-independent. Using an OLS regression of the length of the second over the length of the fourth finger, they find that this regression line does not pass through the origin, but instead has a positive intercept ($a \pm$ SE = 11.2 $\pm$ 1.7 mm, $P < 0.00001$). Hence they conclude that the relationship is allometric rather than isometric and therefore that ratios will change with absolute size. The ratio of 2nd to 4th digit lengths will decrease with increasing size, seemingly explaining why men on average have lower digit ratios than women. Accordingly, when re-examining their data in a GLM (here ANCOVA) with 2nd digit as the dependent variable, 4th digit as a covariate and sex as a fixed effect, there is no longer a sex effect ($P = 0.23$), even though digit ratio itself was highly significantly dimorphic ($P < 0.0001$). The GLM suggests that there is no sexual dimorphism in 2nd digits once controlling for size differences in the 4th digit. It is easy to imagine that most researchers would agree with their conclusion. Moreover, three follow-up studies have discussed and examined the propositions of the Kratochvíl and Flegr (2009) study in great detail (Hönekopp and Watson, 2010; Manning, 2010; Tobler et al., 2011). Although all of them, by various means, reach the conclusion that digit ratio remains sexually dimorphic even when accounting for allometry, none of the studies questions the validity of using OLS regression. To the contrary, they all comply with the inadequate OLS regression, leading to shallow slopes and positive intercepts, rather than the here appropriate RMA regression. In the following I would like to identify the specific problem inherent in the OLS regression analysis of digit ratio, which is best achieved with a simulated data set rather than with real data, as will become apparent. Hence, I generated a data set more or less closely following the parameters from the original Kratochvíl and Flegr (2009) publication.

As shown in Fig. 2a, I started out by drawing 1,000 values from a normal distribution with a mean of 69 mm and a standard deviation (SD) of 5 mm to represent the 4th digits of 1,000 women. Likewise, 1,000 values for 4th digits of men were created with a mean of 77 mm and SD of 5 mm. For the 2nd digit of women, I assumed perfect isometry with a slope of 1 (setting 2nd = 4th), which corresponds to a digit ratio of 1. Given a sexual dimorphism in digit ratio of about 1.5%, I used an isometric slope of 0.985 (2nd = 0.985*4th), which corresponds to a male digit ratio of 0.985. In this case, when there is no scatter, OLS and RMA regression lines coincide, both going through the origin (intercept $a = 0$) with slopes of 1 and 0.985 for women and men respectively (Fig. 2a). In a second step, I added noise to each data point (randomly drawn from a normal distribution with mean = 0, SD = 2 mm), and this was done independently for all of the 2nd and all of the 4th digits (Fig. 2b). An overall (sexes pooled) OLS regression line yields $r^2 = 0.81$ and $a \pm$ SE = 9.8 $\pm$ 0.7 mm, which, for the present purpose, is sufficiently close to the data from Kratochvíl and Flegr (2009; $r^2 = 0.81$, $a \pm$ SE = 11.2 $\pm$ 1.7 mm). In fact, the similarity of intercepts suggests that human digits are probably very close to being isometric. Hence, contrary to Kratochvíl and Flegr (2009), human digit ratio appears to be fairly independent of absolute size and therefore no allometric adjustments are required. In Fig. 2b one can clearly see how the OLS regression lines (sexes now separate; in bold) are shallower than the lines of isometry (thin lines) around which the data were created. The fact that the male and female OLS regression lines end up being so close together that a GLM (Table 1, top) judges them as statistically indistinguishable ($P = 0.27$), is merely coincidental. To clearly illustrate the flaw in this analysis, I have applied Dolph Schluter's trick in exchanging the *x* and *y* axes using the exactly same data set (Fig. 2c). Note that the male isometry line now has a slope of 1/0.985 = 1.01523. This means that there is "positive" sexual dimorphism in inverted digit ratios (4th/2nd), meaning the larger sex has the relatively larger trait values. Again, OLS regression lines are shallower than the isometry lines, but this time, male and female lines do not coincide. Accordingly, a GLM (Table 1, middle) with the 4th digit as the dependent variable, the 2nd digit as a covariate and sex as a fixed factor yields a highly significant ($P < 10^{-52}$) sex difference. This shows that "negative" size dimorphisms (the smaller sex having relatively larger trait values; e.g., corpus callosum size; Smith, 2005) tend to be obscured by the shallower OLS regression lines, while "positive" size dimorphisms tend to be exaggerated.

There are two reasons why the presented GLMs are inappropriate. First, there is collinearity between the predictors, namely sex and size. When using Type 1 sums of squares rather than Type 3 sums of squares it becomes visible that the conclusion of whether the effect of sex is significant depends on the order of entering the predictors into the model, a clear sign of collinearity problems (Freckleton, 2011). However, collinearity is not the primary problem here. OLS regression produces slopes that are downward biased by the statistical noise in x, and these slopes result in biased estimates for the predictor sex (and hence wrong inference) as soon as there is any collinearity between sex and size. Hence, only when there is absolutely no sexual dimorphism in size to correct for, which is practically never reached in a real data set, will the GLM output be unbiased with regard to the sex effect. The joint problem caused by collinearity and statistical noise in x has recently been examined and discussed by Freckleton (2011).

To really understand the main reason why the first two GLMs in Table 1 lead to erroneous conclusions, one has to go back to the assumption of OLS regression that was violated (noise in x). Therefore, I repeated the first analysis (Table 1 top, Fig. 2b), but this time noise was added to the data points only along the y axis (the same amount as before), but not to the x axis. In other words, I regressed the y axis from Fig. 2b over the x axis from Fig. 2a. Importantly, the OLS regression lines now coincide with the lines of isometry around which the data were set up (Fig. 2d; very minor deviations being due to noise in data creation). Hence, the reason for the shallower slope in Fig. 2b compared to 2d, lies in the addition of noise to the values of x. This can easily be seen from another formula describing the slope of an OLS regression line: $b = \mathrm{cov}(xy)/\mathrm{var}(x)$, where $\mathrm{cov}(xy)$ is the covariance between x and y, and $\mathrm{var}(x)$ is the variance in x (Sokal and Rohlf, 1995). Adding statistical noise to the values of x increases the variance in x and lowers the regression slope (because of the larger denominator).

The GLM to the data in Fig. 2d shows a non-significant sex-by-4th finger length interaction ($P = 0.26$), so it cannot tell apart the slight difference in male versus female slopes (0.985 vs. 1). However, a reduced GLM (Table 1, bottom) with a common slope for the two sexes (estimated to be about 0.996) shows a non-significant intercept (in line with isometry), and a parameter estimate for the main effect of sex that is approximately intermediate between the two erroneous earlier estimates. Hence, apart from the failure to distinguish between male and female slopes (for which there is not enough power) the last GLM in Table 1 gives about a fair representation of the data created. However, this was only possible by excluding the noise from the x axis, which cannot be done in real data sets.

A correct way of analyzing the data in Fig. 2b would be fitting MA regression lines. MA rather than RMA lines are appropriate for this artificial data set, since I added absolutely equal amounts of noise to the data (rather than proportionally equal amounts). Using the "smatr" package in R (Warton and Ormerod, 2007), I did this for 10,000 such artificial data sets as in Fig. 2b, which only differ from each other in sampling noise. Reassuringly, the average of all slopes and intercepts coincide quite precisely with the starting parameters for the isometry lines (intercepts = 0, slopes = 1 and 0.985). However, across the 10,000 replicates the 95% confidence intervals for the male and female slopes are very wide. Because means can be estimated more precisely than slopes, it would be a bad idea to estimate the sexual dimorphism in digit ratio from the male and female MA slopes. The 95% confidence interval for sexual dimorphism in digit ratio (male digit ratio/female digit ratio) estimated by this method ranges from 0.936 to 1.038 (mean = 0.985), despite the considerable sample size of 1,000 individuals of each sex. In contrast, the classical method of first calculating the digit ratio of every individual and then assessing the sexual dimorphism in this trait produces a much narrower confidence interval across the 10,000 replicates (0.981–0.988). Hence, because slope estimates are much more sensitive to sampling noise than estimates of means, one should not be surprised by the output of a MA or RMA analysis. In the present example one would always find a non-significant difference in slopes, but a highly significant difference in intercepts (i.e., means when assuming equal slopes). So, the purpose of this line fitting can only be to roughly confirm the assumption of isometry, such that we can continue to use the classical ratio of digit lengths without having to worry about size being a confounding factor. It is clearly not desirable to replace this traditional ratio with residuals from RMA lines, the slopes of which would vary widely between empirical studies already due to sampling noise alone. This would be against all efforts of making research procedures more standardized. Hence, despite of all criticism of ratios, digit ratio researchers should stick to their initial trait of study. Yet for large data sets, it would be worth reporting the equations of RMA lines.

WHEN TO CHOOSE OLS OR RMA?

The central question about when to implement OLS versus RMA regression was recently discussed by Smith (2009). The most critical distinction lies in the distribution of statistical noise between the x and the y axis. OLS regression of y over x assumes all noise to be in y and none in x, and yields the characteristic shallow slope of $b_{yx} = r(\mathrm{SD}y/\mathrm{SD}x)$. The inverted OLS regression of x over y assumes the very opposite distribution of noise and yields the steep inverted slope of $b_{xy} = (1/r)(\mathrm{SD}y/\mathrm{SD}x)$. RMA regression assumes that the noise is distributed over the y and x axes in proportion to their overall variances ($\mathrm{SD}y^2/\mathrm{SD}x^2$), yielding the intermediate geometric mean slope of $b = (\mathrm{SD}y/\mathrm{SD}x)$. As mentioned earlier, when the noise in x is much smaller than that in y, and only due to quantifiable measurement error, then specific modeling of that error is advisable (Warton et al., 2006; Freckleton, 2011). Hence, RMA regression is not a universal cure for cases with some error in x. Some caution regarding RMA may already be necessary for the brain mass example (Fig. 1), since the amounts of biological noise in *x* and *y* cannot be known. However, the intermediate RMA solution is clearly more sensible than either of the two OLS solutions (the regular and the inverted), which delineate the range of all possible solutions in dependence on the relative distribution of errors. The most pragmatic approach is to decide upon whether one seeks a single regression line that best describes the symmetrical relationship between two traits (RMA) or whether one considers the relationship between *x* and *y* to be asymmetrical (*y* depending on *x*, OLS). For two morphological traits, *y* does not depend on *x*, but rather the two are correlated because they share genetic and environmental effects that affect both of them in a correlated way. In contrast, OLS is appropriate for cases where *x* is causal to *y*.

THE BEHAVIOR OF RESIDUALS AND RATIOS

Another important issue to mention regards the behavior of residuals derived from OLS versus RMA regressions. For simplicity, I will focus only on residuals measured parallel to the *y* axis (otherwise see Warton et al., 2006; Smith, 2009). To explore this, I look at residuals from the female data set in Fig. 2b. Although residuals from the OLS regression line—by definition—show no remaining correlation with the *x* axis (r = 0), RMA residuals show a remaining negative correlation with *x* (r = −0.23), just like the classical measure of digit ratio shows such a negative correlation with *x*, that is, the ratio's denominator (r = −0.23). It is this correlation that lead to a lot of criticism against the use of ratios.

What seems to be the strength of OLS methodology turns into a weakness when considering the remaining correlation with the "true" size differences. In real life, "true size" (here meaning the underlying cause of covariance between *x* and *y*) is a latent variable that cannot be measured, but in our simulation it is represented by the *x* axis of Fig. 2a, that is, size differences that are shared between *x* and *y* (making them correlated) before adding uncorrelated noise to *x* and *y*. Problematically, OLS residuals still show a considerable correlation with the noise-free measure of size of the 4th finger, that is, the *x* axis of Fig. 2a (r = 0.26), whereas RMA residuals and digit ratio practically do not (both r = 0.04). It is the noise in the *x*-variable and the subsequently shallower OLS regression line that lead to a systematic under-correction for real size differences.

These differences in behavior between OLS and RMA residuals become even more extreme, as the correlation between *y* and *x* becomes weaker. I give a final example from the study of relative tarsus length in birds to illustrate this. In my captive population of zebra finches, tarsus length is only moderately positively correlated with the cube root of body mass (r = 0.40, N = 988). Hence, RMA residuals of tarsus length over cube root of body mass (or ratios of tarsus divided by cube root of mass) show a disturbingly strong negative correlation with this measure of size (r = −0.55, and r = −0.54, respectively), whereas OLS residuals by definition do not (r = 0). However when tested against another aspect of size, such as wing length, the previously obtained OLS residuals of tarsus length prove to be not quite size independent (correlation with wing length r = 0.19, $P < 10^{-8}$) whereas RMA residuals and ratios are (r = 0.005, and r = 0.009, respectively). Hence, while OLS regression looks cosmetically fine (no residual correlation with cube root of body mass) it systematically underestimates general size effects on tarsus length, yielding tarsus residuals that are inadequately corrected for size, which can lead to serious misinterpretations (Green, 2001; Peig and Green, 2010).

CONCLUSION

Following a series of articles criticizing the use of ratios (e.g., Packard and Boardman, 1988; Ranta et al., 1994) it may have appeared to many researchers that fitting all relevant predictors into a single GLM would provide a solution of universal validity. However, GLMs can be misleading in the study of allometry and probably quite a few other situations. Caution is required whenever there is considerable statistical noise in the predictor variable. This is frequently the case, but the conditions under which conclusions become truly erroneous are a bit more specific. In the present example, GLMs were misleading because the two competing explanatory variables (sex and the size covariate) were correlated (collinearity). Hence, only when the groups to be compared differ in the size covariate, which however will typically be the case, will the group effect be estimated wrongly. More generally, it seems advisable to always visually examine the respective scatter plots, and to bear in mind the shallow slopes of OLS regressions. RMA regressions do not provide a universal cure to all cases with some error variation in *x*, but they are the most sensible solution for all cases where *x* and *y* probably contain similar proportions of biological noise, as in studies of allometry.

ACKNOWLEDGEMENTS

The author thanks Hans Winkler, Andy Green, Mihai Valcu, Holger Schielzeth, Shinichi Nakagawa, Mathias Kölliker, John Manning, and Irmgard Teschke for important suggestions on an earlier version of the manuscript. Mihai Valcu helped with running simulations and contributed with critical discussion.

LITERATURE CITED

Ankney CD. 1992. Sex-differences in relative brain size—the mismeasure of woman, too. Intelligence 16:329–336.
Arnold TW, Green AJ. 2007. On the allometric relationship between size and composition of avian eggs: a reassessment. Condor 109:705–714.
Freckleton RP. 2011. Dealing with collinearity in behavioral and ecological data: model averaging and the problems of measurement error. Behav Ecol Sociobiol 65:91–101.
Galton F. 1886. Regression towards mediocrity in hereditary stature. J Anthropol Inst 15:246–263.
Green AJ. 2001. Mass/length residuals: measures of body condition or generators of spurious results? Ecology 82:1473–1483.
Ho KC, Roessmann U, Straumfjord JV, Monroe G. 1980. Analysis of brain-weight. 2. Adult brain-weight in relation to body height, weight, and surface-area. Arch Pathol Lab Med 104:640–645.
Hönekopp J, Watson S. 2010. Meta-analysis of digit ratio 2D:4D shows greater sex difference in the right hand. Am J Hum Biol 22:619–630.
Kratochvíl L, Flegr J. 2009. Differences in the 2nd to 4th digit length ratio in humans reflect shifts along the common allometric line. Biol Lett 5:643–646.

Legendre P, Legendre L. 1998. Numerical ecology. 2nd ed. Amsterdam: Elsevier.
Manning JT. 2002. Digit ratio: a pointer to fertility, behavior, and health. New Jersey: Rutgers University Press.
Manning JT. 2010. Digit ratio (2D:4D), sex differences, allometry and finger length of 12–30-year olds: evidence from the British Broadcasting Corporation (BBC) internet study. Am J Hum Biol 22:604–608.
McArdle BH. 1988. The structural relationship: regression in biology. Can J Zool 66:2329–2339.
Packard GC, Boardman TJ. 1988. The misuse of ratios, indexes, and percentages in ecophysiological research. Physiol Zool 61:1–9.
Peig J, Green AJ. 2010. The paradigm of body condition: a critical reappraisal of current methods based on mass and length. Funct Ecol 24:1323–1332.
Ranta E, Laurila A, Elmberg J. 1994. Reinventing the wheel—analysis of sexual dimorphism in body-size. Oikos 70:313–321.
Rushton JP, Ankney CD. 1996. Brain size and cognitive ability: correlations with age, sex, social class, and race. Psychon Bull Rev 3:21–36.
Schluter D. 1992. Brain size differences. Nature 359:181.
Smith RJ. 2005. Relative size versus controlling for size—interpretation of ratios in research on sexual dimorphism in the human corpus callosum. Curr Anthropol 46:249–273.
Smith RJ. 2009. Use and misuse of the reduced major axis for line-fitting. Am J Phys Anthropol 140:476–486.
Sokal RR, Rohlf FJ. 1995. Biometry. New York: Freeman.
Tobler M, Healey M, Olsson M. 2011. Digit ratio, color polymorphism and egg testosterone in the Australian painted dragon. PLoS ONE 6:e16225.
Voracek M, Loibl LM. 2009. Scientometric analysis and bibliography of digit ratio (2d:4d) research, 1998–2008. Psychol Reports 104:922–956.
Warton D, Ormerod J. 2007. Smatr: (standardised) major axis estimation and testing routines. R package version 2.1. Available at: http://web.maths.unsw.edu.au/~dwarton.
Warton DI, Wright IJ, Falster DS, Westoby M. 2006. Bivariate line-fitting methods for allometry. Biol Rev 81:259–291.