Bivariate line fitting methods for allometry

David I. Warton

Department of Statistics, School of Mathematics,

*University of New South Wales NSW 2052, Australia**

Ian J. Wright, Daniel S. Falster and Mark Westoby

Department of Biological Sciences, Macquarie University, NSW 2109, Australia

March 17, 2005

*email: David.Warton@unsw.edu.au phone: (61)(2) 9385-7031 fax: (61)(2) 9385-7123

ABSTRACT

Fitting a line to a bivariate dataset can be a deceptively complex problem, and there has been much debate on this issue in the literature. In this review, we describe for the practitioner the essential features of line-fitting methods for estimating the relationship between two variables: what methods are commonly used, which method should be used when, and how to make inferences from these lines to answer common research questions.

A particularly important point for line-fitting in allometry is that usually, two sources of error are present (which we call measurement and equation error), and these have quite different implications for choice of line-fitting method. As a consequence, the point-of-view in this review and the methods presented have subtle but important differences from previous reviews in the biology literature.

Linear regression, major axis and standardised major axis are alternative methods that can be appropriate when there is no measurement error. When there is measurement error, this often needs to be estimated and used to adjust the variance terms in formulas for line-fitting. We also review line-fitting methods for phylogenetic analyses.

Methods of inference are described for the line-fitting techniques discussed in this paper. The types of inference considered here are testing if the slope or elevation equals a given value, constructing confidence intervals for the slope or elevation, comparing several slopes or elevations, and testing for shift along the axis amongst several groups. In some cases several methods have been proposed in the literature. These are discussed and compared. In other cases there is little or no previous guidance available in the literature.

Simulations were conducted to check whether the methods of inference proposed have the intended coverage probability or Type I error. We identified the methods of inference that perform well and recommend the techniques that should be adopted in future work.

Key words: model II regression, errors-in-variables models, standardised major axis, functional and structural relationships, measurement error, method-of-moments regression, test for common slopes, analysis of covariance.

Short title: Line-fitting for allometry

CONTENTS

I INTRODUCTION

Fitting a line to a bivariate cloud of data would seem a relatively simple and fundamental procedure in data analysis. However, there has been lively debate in the literature concerning which method is appropriate in what situation (Ricker, 1973; Jolicoeur, 1975; Sprent & Dolby, 1980; Ricker, 1982; Jolicoeur, 1990; Sokal & Rohlf, 1995; Carroll & Ruppert, 1996), and some of the issues discussed have never completely been resolved. Authors have offered distinctly different reasons for using one method instead of another (Sokal & Rohlf, 1995; Carroll & Ruppert, 1996, for example), and have advocated different methods (McArdle, 1988; Isobe *et al.*, 1990; Jolicoeur, 1990).

In this paper, line-fitting is discussed specifically in the context of allometry, the measurement of how one size variable scales against another (Reiss, 1989; Niklas, 1994). Allometry is a discipline in which alternatives to linear regression are routinely required, because lines are usually fitted to estimate how one variable scales against another, rather than to predict the value of one variable from another. Other disciplines in which such methods are commonly required are astronomy, physics and chemistry (Isobe *et al.*, 1990).

Describing the relationship between two variables typically involves making inferences in some more general context than was directly studied. Given measurements of brain and body mass for a sample of mammals, we would like to interpret results as being meaningful for *all* mammals. Statistical procedures that assist in generalising – making claims about a population, based on a sample – are known as methods of inference. In allometry, we would like to make inferences about the slope and sometimes the elevation of lines that are fitted to data.

This paper reviews the methods of line-fitting commonly used in allometry, their uses, and how to make inferences from lines fitted to a dataset. We identify fundamental points with a logical basis or a wide consensus in the literature, common misinterpretations, points of controversy and ways forward from these.

In describing line-fitting methods and their uses (Sections III and IV), we emphasise the distinction between two types of error, equation error and measurement error. This distinction leads us to a different viewpoint than that taken by most reviewers of this subject in the past, and it leads us to discuss a method of line-fitting that has not been used before (to our knowledge) for allometric problems when both forms of error are present and non-ignorable.

In reviewing methods of inference for common methods of line-fitting (Sections V and VI), we consider procedures for the major axis (MA), standardised major axis (SMA), and modifications of these methods for instances where the line is constrained to pass through the origin or when measurement error is accounted for in estimation. Methods of inference for linear regression are not considered here, being well-known (Draper & Smith, 1998, chapter 14) and available in standard statistics packages. We focus on the techniques appropriate for:

- Testing if slope and elevation equal a particular value, and estimating confidence intervals for slope and elevation (Fig. 1a)
- Testing if several lines have a common slope (Fig. 1b)
- Testing if several lines have a common elevation (Fig. 1c)
- Testing for no shift along lines of common slope (Fig. 1d)

Fig. 1 summarises schematically the hypothesis of interest in each of these situations, for a particular dataset that is explained in more detail in the following section.

We have found methods for comparing several independent lines to be particularly useful, and so devote considerable time in this paper to this topic. Such methods are useful for exploring how the relationship between two variables changes across functional groups, populations, environments, *etc.*

Some of the tests for comparing several bivariate allometric relationships (Figures 1b-c)

are analogous to analysis of covariance, but for the MA and SMA lines rather than for linear regression. Analysis of covariance is of limited usefulness in allometry, because linear regression is often inappropriate. Despite this, analysis of covariance has often been used in previous allometric work, because of an apparent unavailability of alternative methods of inference (for example, by Wilkinson & Douglas, 1998). However, there is no longer a need to resort to analysis of covariance in situations where it is considered inappropriate, given the methods described in this paper.

This review contains several novel contributions to the literature on line-fitting in allometry:

- Several points are made regarding usage and interpretation of methods that are new to the biology literature.
- We discuss a method of line-fitting that has not been used before (to our knowledge) for allometric problems when both equation and measurement error are present and non-ignorable.
- A geometric interpretation of methods of inference is presented, where possible.
- Some useful developments for comparing several lines are reviewed that are not well known in the biology literature (Flury, 1984; Warton & Weber, 2002).
- New methods are suggested in this paper, when no guidance is currently available.
- Simulations have been conducted (Appendix E) to explore the properties of the methods discussed in this paper. The simulation results lead us to some new conclusions.

Guidelines on measurement error calculation are given in Appendix C, calculation formulas for the methods of inference considered here are provided in Appendix D, simulations assessing the efficacy of these methods are presented in Appendix E, and algorithms for resampling are given in Appendix F.

II SOME ALLOMETRIC EXAMPLES

This section briefly introduces allometry and describes examples of where allometry is used.

In allometry, typically there are two size variables y and x which are believed to be related by the equation

$$y = \gamma x^{\beta}$$

This is often referred to as the "allometric relation" (Harvey & Pagel, 1991) or "allometric equation" (Reiss, 1989). The x and y variables are log-transformed, so that the above equation can be reexpressed as

$$Y = \log(\gamma) + \beta X$$

$$Y = \alpha + \beta X$$

where we have made the substitutions $Y = \log(y)$, $X = \log(x)$, and $\alpha = \log(\gamma)$. There is a linear relationship between Y and X. The log transformation is used for two different reasons. Firstly, it allows the relationship between the two size variables to be expressed as a linear relationship, which simplifies estimation. Secondly, it puts the size variables on a multiplicative or logarithmic scale. This is a sensible scale for interpreting most size variables, since growth is a multiplicative process.

It should always be checked whether or not log-transformed size variables are linearly related, because it may not be the case that two size variables are related by the allometric equation. Experience shows, however, that it is commonly a good approximation to the relationship between two size variables.

In some allometric work, it may not be considered desirable to log-transform variables. In the remainder of this article, we refer to the fitting of a linear relationship between Y and X, where these variables may or may not have been transformed. So, for example, Y might represent log(seed mass), log(brain mass) or height.

Throughout this paper we will refer to three examples, each of which is useful for highlighting different aspects of line-fitting for allometry.

Fig. 2 is a plot of brain mass against body mass for 62 mammal species, for data from Allison & Cicchetti (1976). On the logarithmic scale, these two variables appear to be linearly related, and the slope of the relationship has been hypothesised to be $\frac{2}{3}$ or $\frac{3}{4}$, based on arguments reviewed by Schoenemann (2004).

Fig. 3 is a plot of plant height vs basal diameter for *Rhus trichocarpa* saplings, from Osada (in preparation). Note that whereas the points on the plot in Fig. 2 were species, the points in this case represent individual saplings.

Fig. 1 represents a third example dataset that will be discussed in more detail below.

(1) Example with several independent lines

Fig. 1 refers to an example from our own experience (Wright & Westoby, 2002) which is particularly useful for discussing methods of inference about allometric lines. All four of the methods of inference described in this paper were of interest for this dataset.

The data in Fig. 1 are leaf longevity data (in years) against leaf mass per area (kg/m^2), for plant species sampled at four different sites (Wright & Westoby, 2002). Leaf mass per area can be interpreted as the plant's dry mass investment in the leaf, on a per unit area basis. Leaves with higher LMA are more expensive, from the plant's point-of-view, but they tend to live longer. A consistent positive relationship between these two variables has been documented in environments all over the globe (Wright *et al.*, 2004).

One question of interest is whether leaf longevity is directly proportional to leaf mass per area (Question a – depicted in Fig. 1a). If this is the case, then a doubling of mass investment in leaves doubles lifetime. In terms of light capture, this would mean that there is no lifetime advantage to putting more mass into leaves of a given area, because

the potential lifetime light capture (leaf area×longevity) would be directly proportional to the mass initially invested in the leaf. If leaf longevity and LMA are directly proportional to each other, then the log-transformed variables will be linearly related with a slope one. Hence we wish to test (for any particular site) whether or not the slope is one (Fig. 1a, the dotted line has slope one).

Another question of interest is whether there are differences in the nature of the relationship between leaf longevity and LMA, for different plant communities. In particular:

Question b Does the slope of the relationship between leaf longevity and LMA change across different sites? If so, then across communities, this suggests different leaf longevity gains from additional mass investment in leaves. This is depicted in Fig. 1b, for two high rainfall sites that differ in soil nutrient levels.

Question c Is there a shift in elevation across different sites? If so, then across communities, leaf longevity differs for plants with similar LMA. This suggests that leaves in different communities have different opportunities for total lifetime light capture, for a given mass investment in the leaf. Fig. 1c shows two low nutrient sites with different rainfall, which may differ in elevation (but have a common slope).

Question d Is there a shift along a common axis across different sites? If so, then LMA and leaf longevity tend to vary across sites, but the relationship between the two variables remains the same, *i.e.* species at neither site have an overall advantage in terms of lifetime light capture for leaves of a given structure. Fig. 1d shows two low rainfall sites with different nutrient levels, which may share a common axis but differ in location along this axis.

It should be noted that if the slope of the relationship does change across sites, then Questions c and d cannot be addressed. This is for the same reasons as in analysis of covariance – elevation and location along the line are not comparable for lines with different slopes.

III LINE FITTING METHODS AND THEIR USES

The three methods of bivariate line-fitting of primary interest in this paper are best known as linear regression, major axis (MA) estimation and standardised major axis (SMA) estimation. The MA and SMA methods are sometimes collectively called "model II regression", and SMA is currently better known as the "reduced major axis". We deliberately avoid using such terms in this manuscript, and offer reasons not to use these terms in Appendix A.

Linear regression, MA and SMA are all least squares methods – the line is estimated by minimising the sum of squares of residuals from the line, and the methods can be derived using likelihood theory assuming normally distributed residuals (Sprent, 1969, for example). The differences in methods of estimation of the lines are due to differences in the direction in which errors from the line are measured, which is illustrated graphically in Fig. 4.

Some definitions will need to be made, based on Fig. 4, which will be useful later in understanding methods of inference for these lines. An axis in the direction of the fitted line can be defined as the "fitted axis", and an axis parallel to the direction residuals are measured in could be defined as the "residual axis". We will refer to scores along fitted and residual axes as "(fitted) axis scores" and "residual scores", respectively. If the residual scores were shifted to have a mean of zero, they would become residuals in the conventional sense. The use of residual scores rather than residuals is important later in discussions about elevation – but in other cases, use of residual scores rather than residuals is not essential.

The fitted and residual axes are useful in understanding estimation and methods of inference for these lines. For example, the linear regression, MA and SMA slopes can all be derived as the value of the slope such that the residual and fitted axis scores are uncorrelated (Warton & Weber, 2002). Further, for the methods of inference for MA and SMA

considered in this paper, the only thing that differs between the MA and SMA cases is the way that the residual and fitted axes are defined.

In interpreting Fig. 4, it is important to make the distinction between two possible sources of error, described by Fuller (1987) as measurement error and equation error. Measurement error is generally well understood, it is where measured values do not represent the true values of the subjects being measured. Equation error is a term that is more often neglected – it is where the actual values of the subjects do not fall exactly along a straight line. For example, it is apparent that humans have an unusually large brain for their body mass (the highlighted point in Fig. 2). There are various possible explanations for this, none of which is error in measuring the average brain size or body mass of humans.

Note that what we describe as "measurement error" is not only error in measurement of a particular subject, but it may also include sampling error, if the subject of interest is a collection of individuals (a population or species). In fact Riska (1991) referred to measurement error as "sampling error", recognising variation introduced through sampling as the main source of measurement error in most allometric work. For example, the subjects in Fig. 2 are species, so measurement error for brain mass includes error measuring the brain and sampling error due to the fact that not all individuals of a species have the same brain mass. The subjects in Fig. 3 are individual plants, so there is no sampling error in estimating basal diameter and height. However, if the Y variable in Fig. 3 were leaf area, then there would be sampling error. Not all leaves on a sapling are the same size, so the measured values of leaf area would depend on what leaves were sampled.

The distinction between measurement and equation error has been made by other authors in the past. Equation error has been referred to as "natural variability" (Ricker, 1982), "natural variation" (Sokal & Rohlf, 1995), "biological error" (Riska, 1991) and "intrinsic scatter" (Akritas & Bershady, 1996), amongst other terms. The implications of measurement error for choice of line fitting method are different from the implications of equation error, but in much of the literature (even the statistical literature) this has not

been recognised (Carroll & Ruppert, 1996).

Whereas measurement error can be estimated from repeated measurement, equation error can not, and its nature depends on the purpose of line-fitting. Do humans have unusually large brains for their body size, or unusually small bodies for their brain size, or a bit of both? Any of these statements is correct, so it can be appropriate to attribute equation error to the Y variable, the X variable or both, depending on the purpose of line-fitting.

When equation error only is present, any of linear regression, MA and SMA might be appropriate methods of analysis. When a non-ignorable quantity of measurement error is also present, often this should be estimated and the line-fitting methods modified to incorporate this, as will be described below.

(1) Linear regression

Regression is a method of fitting lines for prediction of the Y-variable. Regression involves "conditioning on the X-variable" (Kendall & Stuart, 1973, chapter 26) – in other words, regression can be used for questions of the form "if we observed a subject whose value on the X-variable is x, what do we expect its value of Y to be?

Regression is useful whenever a line is desired for predicting one variable (which will be called Y) from another variable (which will be called X). The purpose of regression can be seen in the method of line estimation – the line is fitted to minimise the sum of squares of residuals measured in the Y direction, $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$. Such a line has fitted Y values as close as possible to the observed Y values, which is a sensible thing to do if you are interested in predicting Y values, once given a set of X values.

Regression is the appropriate method of line-fitting in most practical instances, because most problems can be expressed as problems of prediction. One of the more common research questions is "is Y associated with X?", which can be rewritten as "for subjects with different X values, are the expected Y values different?" This second question can

be answered by fitting a regression line and testing if the slope is significantly different from zero. "How strongly are Y and X associated?" is another question that can be answered using regression. A suitable statistic to answer this question is the square of the correlation coefficient, R^2, the proportion of variation in the Y variable that can be explained by linear regression on X.

Galton (1886) gave regression its name due to the property of "regression towards mediocrity" (or regression to the mean), where predicted values of observations tend to be closer to the mean than observed values, in general. Galton (1886) considered the height of parents ("mid-parents", with female heights transformed to a comparable scale as male height) compared to the height of their children, which is reproduced in Fig. 5. For these or any two variables that are expected to be scattered evenly around the one-to-one line, a fitted regression line will be flatter than a slope of one no matter how large the dataset is. In fact, the line will be close to a slope of r, the correlation coefficient. This situation is natural from the point-of-view of prediction – if a father is really tall, then his son would probably be tall too, but not as tall as him (in the same way that if a student got a really high score in a subject in a test – higher than they had ever got before – then they might expect to do well in the next test for this subject, but not quite as well as last time). While regression to the mean is useful in prediction, it is not appropriate when the value of the slope of the axis or line-of-best-fit is of primary interest.

The following are points concerning usage of linear regression that have occasionally been confused in the literature:

- Regression can be used irrespective of whether the X variable is fixed by the experimenter or a random variable (Draper & Smith, 1998, chapter 1). To estimate a regression line, the X-variable is conditioned on or "fixed". This fixing of X is a mathematical construction, but it has on occasion been confused with experimentally fixing a variable at a particular value. Some appear to have interpreted "regression requires fixing of the X variable" as meaning that X needs to be experi-

mentally fixed to use regression (Niklas, 1994; Quinn & Keough, 2002, for example), which we believe is a misunderstanding arising from different uses of the term "fixed" in statistics and in experimental sciences.

- Linear regression can be used when X is measured with error, as long as results are only interpreted in the context of predicting Y from X measured with error. If X has been measured with error (as it usually is), linear regression gives a biased estimator of the slope of the regression of Y against X (Fuller, 1987, page 3). This does not, however, mean that the use of linear regression is no longer appropriate when there is measurement error. On the contrary, a simple linear regression of Y can be used to answer some common questions – is Y related to X (Fuller, 1987, page 4), what is the predicted Y when X is observed (with error) to be $x \ldots$ Measurement error and regression will be considered in more detail later.

- Regression can be used to predict a causal variable, *i.e.* the causal variable can be treated as the Y variable and the outcome variable can be treated as the X variable. Further, there does not need to be causation for regression to be applied (Draper & Smith, 1998, for example). Regression only requires a desire to predict one variable from another, not causation. Confusion can arise because of two distinct conventions – the convention in graphing of always putting the causal variable (if there is one) on the X axis, and the convention in regression of always putting the predictor variable on the X axis. The variable being predicted needs to be a random variable, so it must not be fixed by sampling method, but what type of random variable is predicted (causal, outcome, *etc*) is entirely up to the researcher.

(2) Major axis and standardised major axis

When there are two variables, the major axis (MA) or standardised major axis (SMA) can be used to describe some axis or line-of-best-fit. The purpose of line-fitting is not to predict Y from X, it is simply to summarise the relationship between two variables. Such

a line is a summary in the sense that a single dimension is used to describe two-dimensional data. This is also known as data reduction or dimension reduction.

There are at least three contexts in which these methods are useful:

- Allometry – when the purpose of the study is to describe how size variables are related, typically as a linear relationship on logarithmic scales.
- "Law like relationships" (Sprent, 1969) are essentially the same application as allometry but in a more general setting – testing if a particular theoretical relationship holds for data from any discipline, *e.g.* is pressure inversely related to volume.
- Testing if two methods of measurement agree. When the methods of measurement are unbiased, this is a special case of "law-like relationships" where the true values of subjects are known to lie on the line $Y = X$ (the one-to-one line), *i.e.* there is no equation error.

The major axis is the line that minimises the sum of squares of the shortest distances from the data points to the line. The shortest distance from a point to a line is perpendicular to it, so in this method residuals are measured perpendicular to the line. The major axis is equivalent to the first principal components axis calculated from the covariance matrix, and fitted through the centroid of the data.

Implicit in minimising the distance from a point to the line is the assumption that departures from the line in the X and Y directions have equal importance. This is expressed in the errors-in-variables literature by assuming that the ratio of the variances of residuals in the X and Y directions is one (although as discussed later, it is not advisable to think of line-fitting in allometry as an errors-in-variables model).

The standardised major axis is the major axis calculated on standardised data, then rescaled to the original axes. This is typically done when two variables are not measured on comparable scales, in which case it might not seem reasonable to give the X and

Y directions equal weight when measuring departures from the line. This technique is equivalent to finding the first principal component axis using the correlation matrix, then rescaling data.

The direction in which error is estimated for SMA is given in Fig. 4. See appendix B for further explanation.

There are many competitors to the major axis and standardised major axis methods, although these methods are relatively infrequently used. Examples include the "OLS bisector" (Isobe *et al.*, 1990, the average slope of the two linear regression lines: for predicting Y and for predicting X) and Bartlett's three group method (Nair & Shrivastava, 1942). The *ad hoc* nature of these approaches seems undesirable – the methods are not model-based, and lack the geometrical interpretation of MA or SMA (MA being the major axis of the bivariate ellipse, MA and SMA attributing errors from the line to a known direction...).

It is important to recognise that when finding a line-of-best-fit through data, there is no single correct method. The major axis, standardised major axis and alternatives all estimate a line-of-best-fit in different ways, and measure slightly different things about the data. The choice between the major axis and standardised major axis (and some alternatives) is based on assumptions about how equation error is partitioned between the X and Y directions (as in Fig. 4). However, because equation error is not a physical entity that can be directly measured, there is no single correct way to partition it into the X and Y directions. Statisticians describe the underlying line as "unidentifiable" (Moran, 1971; Kendall & Stuart, 1973) in this case.

The following are important points concerning the use of major axis or standardised major axis methods.

- When collecting data with a view to fitting MA or SMA lines, subjects should be randomly selected and not chosen conditionally on the values of the X or Y

variable. In regression, it is common for samples to be selected systematically to represent a large range of X values. In fact, this is a desirable sampling strategy in regression, because it allows the line to be estimated much more efficiently than if simple random sampling were used. However, when fitting MA or SMA lines, both X and Y variables are treated as random and so need to be sampled so that they are random. If the X variable were sampled so that the variance on this axis was high, this would bias the major axis or standardised major axis slope towards zero – the observed slope would usually be flatter than the true slope.

- MA/SMA methods should not be used simply because X is measured with error. It has on occasion been claimed that the major axis or standardised major axis needs to be used when the X variable is subject to measurement error (Niklas, 1994; Sokal & Rohlf, 1995). However, if the purpose of the line-fitting can be expressed in terms of prediction, a regression method should be used instead (Carroll & Ruppert, 1996; Draper & Smith, 1998). Confusion can arise about the reason for using MA or SMA because these methods attribute error from the line to the X variable as well as Y, whereas regression attributes error to just Y, as in Fig. 4.

- In allometry, you should not use information about measurement error to choose between MA, SMA and related methods. In allometry, equation error will invariably be present, and the direction in which equation error operates depends how you look at the data and not on anything that can be measured. Harvey & Pagel (1991) estimated the ratio of variances of measurement errors in X and Y, then used an errors-in-variables model assuming known error variance ratio. This model is known in the biology literature as the structural or functional relationship (Sprent & Dolby, 1980; Rayner, 1985; McArdle, 1988; Sokal & Rohlf, 1995). The difficulty that Harvey & Pagel (1991) encountered was that they were only able to estimate measurement error and not equation error, so in using this approach they essentially assumed that equation error was either zero or proportional to measurement error. While this approach has received much consideration in the literature, it should not

be used when equation error is present, a point made best by Carroll & Ruppert (1996). In allometry, equation error is often large compared to measurement error, in which case it would be more reasonable to assume there is no measurement error than to assume no equation error. Alternative methods that explicitly account for measurement error are described below.

(3) Line-fitting when accounting for measurement error

The presence of any measurement error will bias estimates of the slope of a line (Fuller, 1987), except in some special cases. In all studies, some amount of measurement error is present. In this section, we will consider when measurement error needs to be taken into account in analyses, and describe the most common method of modifying line-fitting methods to take measurement error into account.

To discuss measurement error, some terminology is needed. The error in X will be written as δ_X, and the error in Y as δ_Y. The variables that are observed are not X and Y but $(X + \delta_X)$ and $(Y + \delta_Y)$. Given a measurement $(X + \delta_X)$, it is not possible to tell exactly what the true value of X is and what the error (δ_X) is. (If, for example we observe the value 10.4, then we know that $X + \delta_X = 10.4$ but can not solve for the values of X and δ_X.) It is usually reasonable to assume that measurements are unbiased (so the true means of δ_X and δ_Y are zero), that δ_X and δ_Y are independent of each other, and that δ_X and δ_Y only depend on X and Y through their variances.

It many instances it is not necessary to account for measurement error in X or Y when fitting a line. Typically, this is either because measurement error is negligible, or because the questions of interest can be answered using a regression of $(Y + \delta_Y)$ vs $(X + \delta_X)$, and there is no need to estimate some relationship between Y and X measured without error. The following are situations in which it is appropriate to use regression without correcting for measurement error:

- To test if Y and X are related. Testing for an association between $(Y + \delta_Y)$ and $(X + \delta_X)$ is appropriate in this situation (Fuller, 1987). If there is no evidence of an association between $(X + \delta_X)$ and $(Y + \delta_Y)$, then there is no evidence of an association between X and Y.

- To predict values of Y from the observed values of X that have been measured with error. In this case, we want to predict Y given a value of $(X + \delta_X)$, and so a regression of Y against $(X + \delta_X)$ should be used, in the same way that a regression of Y vs X should be used to predict Y given a value of X.

- In regression situations when there is measurement error in Y only, and the magnitude of the measurement error is not a function of Y. In this situation the regression line of $Y + \delta_Y$ vs X is unbiased. Measurement error would only need to be considered if it was desirable to partition error variance into the components due to equation error vs measurement error.

Note that the first two of these cases are particularly common in regression applications. Consequently, a large proportion of instances where a regression line is fitted do not require adjustment for measurement error.

It is only necessary to account for measurement error if it is important that the fitted line describes a relationship between Y and X, rather than between the variables measured with error ($Y + \delta_Y$ and $X + \delta_X$). The following are examples of this:

- When slopes or correlation coefficients are to be compared to those from other studies which may have different magnitudes of measurement errors. Different amounts of measurement error bias results by different amounts (Fuller, 1987), which would need to be accounted for in comparisons.

- When theory predicts that the slope of the line relating Y and X should take a particular value – in such a case clearly the slope needs to be estimated without bias. For example, it is of interest to test if the slope of the relationship between log(brain

mass) and log(body mass) in Fig. 2 is consistent with the value $\frac{2}{3}$, or $\frac{3}{4}$ (Schoenemann, 2004), or to test if seed output of plant species is inversely proportional to seed mass (Henery & Westoby, 2001, and others).

Note that the typical situations in which MA or SMA are fitted correspond to one or both of these cases – so unless measurement error is negligible, it would need to be accounted for.

When can measurement error be considered negligible? Akritas & Bershady (1996) were reluctant to advise on this issue and instead recommended accounting for measurement error no matter how small it may be – after all, this approach will never lead to a biased estimator. McArdle (2003) suggested that when considering the influence of measurement error on a linear regression slope, a useful procedure is to estimate the proportion of the sample variance in the X variable that can be considered to be due to measurement error, p, and to calculate $\frac{p}{1-p}$. This is an estimate of the proportion of attenuation, *i.e.* it is an estimate of the proportional decrease in the estimated regression slope due to measurement error. If this decrease is of a scale that does not alter conclusions, it could be ignored. When considering the effect of measurement error on MA and SMA slopes, the simplest rule is to recalculate slopes accounting for measurement error and compare these to the original slope estimates. We have done this for several datasets and found relatively small effects of measurement error (slope changed by $< 8\%$). Nevertheless, we can not claim that measurement error is generally negligible in allometry because its magnitude will vary with the type of variable measured and the number of repeated measures taken on each subject.

To account for measurement error, the average measurement error variance of observations on X and Y needs to be estimated based on repeated measures. There does not need to be the same number of repeated measures for each subject, and the measurement error variance does not need to be the same for different subjects, as described in Riska (1991) and Akritas & Bershady (1996). More details and examples of how to estimate

measurement error are given in appendix C.

Before taking repeated measures to estimate a measurement error variance, careful thought is often required to identify what constitutes a repeated measurement. For example, if the subjects in analyses are species occurring in some region, the repeated measurements are observations of different individuals in the region. Note that if there are several populations in the region of interest, a representative sample should contain (randomly selected) individuals across all populations. Now consider a situation in which the subjects are individuals measured during a period of a week, but there may be systematic changes in subjects over the course of the week (due to growth, for example). Then repeated measurements of an individual would be measurements taken at random times over the week.

In the presence of measurement error whose variance is estimated from repeated measures, consistent estimators of slopes of lines can be obtained by replacing the sample covariance matrix by a methods-of-moments estimator, as follows. If measurement errors in the X and Y directions are independent of each other and of the true value of X or Y,

$$Var(X + \delta_X, Y + \delta_Y) = Var(X, Y) + Var(\delta_X, \delta_Y)$$

so

$$Var(X, Y) = Var(X + \delta_X, Y + \delta_Y) - Var(\delta_X, \delta_Y)$$

Writing out the sample estimates of these covariance matrices term-by-term:

$$\begin{pmatrix} s^2_X & s_{X,Y} \\ s_{X,Y} & s^2_Y \end{pmatrix} = \begin{pmatrix} s^2_{X+\delta_X} & s_{X+\delta_X,Y+\delta_Y} \\ s_{X+\delta_X,Y+\delta_Y} & s^2_{Y+\delta_Y} \end{pmatrix} - \begin{pmatrix} s^2_{\delta_X} & 0 \\ 0 & s^2_{\delta_Y} \end{pmatrix}$$

and so the covariance matrix of the true X and Y values can be estimated as

$$\begin{pmatrix} s^2_{X+\delta_X} - s^2_{\delta_X} & s_{X+\delta_X,Y+\delta_Y} \\ s_{X+\delta_X,Y+\delta_Y} & s^2_{Y+\delta_Y} - s^2_{\delta_Y} \end{pmatrix}$$

The terms $s^2_{\delta_X}$ and $s^2_{\delta_Y}$ would need to be estimated from repeated measures as in appendix C, the remaining terms in the above are the sample variances and covariances of the observed variables.

For example, consider the regression slope. The standard estimator of the regression slope when measurement error is not accounted for is

$$\hat{\beta}_{\text{reg}} = \frac{s_{X+\delta_X, Y+\delta_Y}}{s^2_{X+\delta_X}}$$

Replacing the relevant terms to account for measurement error, this becomes:

$$\hat{\beta}_{\text{MM,reg}} = \frac{s_{X+\delta_X, Y+\delta_Y}}{s^2_{X+\delta_X} - s^2_{\delta_X}} = \frac{s^2_{X+\delta_X}}{s^2_{X+\delta_X} - s^2_{\delta_X}} \hat{\beta}_{\text{reg}}$$

This is known as methods-of-moments regression (Carroll & Ruppert, 1996). Methods-of-moments regression can also be derived as the maximum likelihood solution when all variables are normally distributed (Fuller, 1987). An alternative and more complicated method is available for the case where data are species means (Kelly & Price, 2004), and it is unclear whether there are any advantages to the use of this method.

Adjusting for measurement error in estimating variance terms in a similar fashion leads to the following methods-of-moments standardised major axis slope estimate:

$$\text{sign}(s_{X+\delta_X, Y+\delta_Y}) \sqrt{\frac{s^2_{Y+\delta_Y} - s^2_{\delta_Y}}{s^2_{X+\delta_X} - s^2_{\delta_X}}}$$

and the following methods-of-moments major axis slope estimate:

$$\frac{1}{2s_{X+\delta_X, Y+\delta_Y}} \left\{ s^2_{Y+\delta_Y} - s^2_{\delta_Y} - s^2_{X+\delta_X} + s^2_{\delta_X} + \sqrt{(s^2_{Y+\delta_Y} - s^2_{\delta_Y} - s^2_{X+\delta_X} + s^2_{\delta_X})^2 + 4s^2_{X+\delta_X, Y+\delta_Y}} \right\}$$

Akritas & Bershady (1996) proposed estimators equivalent to the above for major axis and standardised major axis slopes that account for measurement error. Akritas & Bershady (1996) proposed obtaining methods-of-moments regression slope estimators and transforming these to find the MA and SMA slope using identities relating these slopes (Isobe *et al.*, 1990, Table 1). To our knowledge, no other authors have attempted to account for measurement error when estimating MA and SMA slopes. Instead, most authors have taken the view that MA and SMA inherently account for measurement error – and as previously discussed, this will lead to biased slope estimators, except in particular circumstances.

There are some difficulties with the use of methods-of-moments estimators of the covariance matrix in errors-in-variables models:

- While variance formulas are available for methods-of-moments regression (Fuller, 1987) and methods-of-moments MA and SMA (Akritas & Bershady, 1996), these do not always perform well in small samples (Appendix E). Resampling methods might need to be used to construct confidence intervals and test hypotheses.
- It is possible, although unlikely, for the variance estimates to be negative (if the estimated measurement error for a variable were larger than its sample variance), in which case the method should not be used until more accurate measurements can be obtained. Something is very wrong if most of the variation in a variable is due to inaccuracies in measurement.
- If measurement error variance is large compared to the sample variance, then the slope estimator can behave erratically – the line may fit the data poorly, and the slope estimator may be inefficient, having a large standard error.

The difficulties listed above can be addressed by using resampling for inference and ensuring that measurement error is relatively small. The size of the measurement error variances can be controlled by the number of repeated measurements. Whereas the precision of the measurements themselves may not be able to be improved on, averaging independent measurements dramatically reduces measurement error – the variance of measurement error is then the variance of a mean, which has the form $\frac{\sigma^2}{n}$, and n can be chosen by the experimenter. For example, the variance of measurement error is halved if the number of repeated measurements that are averaged is doubled.

(4) Line-fitting for phylogenetically independent contrasts

Often it is of interest to investigate the evolutionary divergence of traits, rather than simply to investigate cross-species patterns across traits at present time. In such a situation,

rather than asking "How are brain mass and body mass related?", it is of interest to ask "As mammals evolved, how were changes in brain mass related to changes in body mass?" as in Schoenemann (2004).

There has been much discussion of the issue of how to calculate divergences (Harvey & Pagel, 1991, for example), and of the general questions for which these types of analyses are useful (Westoby, Leishman & Lord, 1995). In the following, we will pass over these issues and discuss the method of fitting allometric lines given a set of divergences in variable X and divergences in variable Y, across some set of divergence events. These divergences should be independent and have equal variance for each variable, which may require adjustments similar to those described in Felsenstein (1985).

It should be noted that whereas methods of analysing divergences for linear regression are well established (Felsenstein, 1985; Grafen, 1989), no methods have previously been described for fitting MA and SMA, to our knowledge. This is despite the fact that fitting MA and SMA for divergence data is potentially of wide interest – much allometric work is comparative across different taxa, and it is common in comparative work to study traits in the context of evolutionary divergence.

All the line-fitting methods described in this paper can be modified for use with divergence data by replacing the terms $\bar{x}$ and $\bar{y}$ with zero. This ensures that the line passes through the origin, which is important for two reasons. Firstly, the origin represents the point where there is no evolutionary divergence in the two variables measured (both X and Y divergences are zero). Secondly, the sign attached to any divergence is arbitrary, which implies that the mean difference on X and Y should be zero (Felsenstein, 1985), *i.e.* the center of the data is the origin. Fitting lines through the origin for divergence data was discussed by Harvey & Pagel (1991), although it should be emphasised that the same logic applies equally well to MA or SMA.

Consider, for example, the standardised major axis slope in the absence of measurement

error

$$\hat{\beta}_{\mathrm{SMA}} = \frac{s_Y}{s_X} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

where there are N samples denoted $(x_1, y_1), \ldots, (x_N, y_N)$, and $\bar{x}$ is the sample mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$. For SMA fitted to divergences, $\bar{x}$ and $\bar{y}$ can be set to zero, so the slope estimator is

$$\hat{\beta}_{\mathrm{SMA}} = \sqrt{\frac{\sum_{i=1}^{N} y_i^2}{\sum_{i=1}^{N} x_i^2}}$$

IV REGRESSION, MA, OR WHAT?

The main contexts in which different methods of line-fitting are used have been summarised in Table 1. This can be a useful guide in deciding which line-fitting method is appropriate for a given situation. It can be helpful when going through this process to think about which statistic is of primary interest (labelled "key statistic" in Table 1). Usually, if the statistic of primary interest is the slope ($\hat{\beta}$), then MA or SMA is appropriate rather than linear regression. On the other hand, linear regression can always be used if primarily interested in the P-value for the test of no relationship between Y and X, or predicted values ($\hat{y}$), or the strength of the linear relationship (R^2).

Often there can be doubt about which method is appropriate, and one potentially confusing point is that the same dataset could be analysed using different methods of line-fitting depending on what the data are to be used for. This happened in the case of Moles & Westoby (2004), who synthesised data across various demographic stages to see if the various advantages of large seeded species (in predation, seedling development, *etc*) compensated for the lower number of seeds produced by a parent plant of a large-seeded species. In doing this, the question of interest was "do plants with big seeds have an overall life-history advantage over small seeded species?", which is a question of prediction where seed mass is the predictor. This at first seemed to be an unusual line-fitting method to adopt, however, because seed number vs seed mass is a classic example of allometry and so such data is usually fitted with a major axis or standardised major axis.

Feigelson & Babu (1992, page 64) similarly present an example dataset from astronomy where either regression or MA/SMA could be appropriate, depending on the purpose of fitting the line.

(1) Major axis or standardised major axis?

To this point, no guidance has been given concerning which is the more appropriate of the major axis and the standardised major axis. This is an issue that has seen debate in the biology literature for 30 years (Ricker, 1973; Jolicoeur, 1975; McArdle, 1988; Jolicoeur, 1990; Legendre & Legendre, 1998) a debate that was never really resolved. Interestingly, there has been little debate in the principal components literature, which discusses equivalent methods.

A key point to keep in mind is that MA and SMA slopes estimate different things about the data, and so MA and SMA lines are not directly comparable, as emphasised by Isobe *et al.* (1990) and Feigelson & Babu (1992).

In practice, these two methods give similar results if the variances of the two variables are similar (say, within a factor of 1.2) or if correlation is high, in which case it does not actually matter which method is used. In fact, the methods are identical for tests of whether a slope is equal to ± 1 or not, which is commonly the test of interest in allometry. In other cases, however, the major axis and standardised major axis slopes can lead to quite different results.

There have been several general recommendations regarding the use of MA vs SMA that are essentially free from controversy. These recommendations are summarised in Table 2, although some of the points require further elaboration:

Efficiency While it is not disputed that SMA slopes are estimated more efficiently than MA slopes, this result has been interpreted in different ways in the literature. Isobe *et al.* (1990) use efficiency as grounds for choosing a method for line-fitting, hence the

relatively small confidence bands for SMA slopes are interpreted as advantageous. On the other hand, Jolicoeur (1990) considered such narrow confidence bands as "unrealistic", given that they are so much narrower than the confidence bands for MA slopes. However, the latter interpretation can be rejected, because the confidence intervals for a SMA slope are known to be exact or close to exact in most practical instances (as demonstrated in Appendix E).

Scale dependence and log transformation It has previously been argued that if variables are log transformed, the variables are on a comparable scale, in which case the scale dependence of the major axis is irrelevant (Jolicoeur, 1975; Legendre & Legendre, 1998). However, scale dependence remains an issue for log transformed variables if the power of the X or Y variable is arbitrary. For example, it could be argued that you could equally well plot height vs basal area rather than height vs basal diameter in Fig. 3. But basal area is proportional to the square of diameter, so this constitutes an arbitrary scale change if variables have been log-transformed.

Inference in complex problems It was explained previously that the essential difference between MA and SMA is that data are implicitly standardised before line-fitting for SMA. This standardisation of data complicates inference (Anderson, 1984; Jolliffe, 2002, for example). For the more commonly encountered situations where a method of inference might be required, methods have been developed for both MA and SMA, as reviewed in sections V and VI. However, in some situations a procedure may be available for MA but not for SMA. For example, if comparing the slopes of several axes that have been constructed through three or more dimensions, the methodology of Flury (1984) could be used in the major axis case, but no equivalent approach is currently available for the standardised major axis.

Some more controversial claims have also been made concerning whether MA or SMA should be preferred. We have summarised these in Table 3, and included some arguments why these claims can be disputed, to emphasise that these claims do not provide a strong basis for preferring SMA to MA, or vice versa.

The authors tend to prefer using SMA, while also considering the use of MA or the OLS bisector approach (Isobe *et al.*, 1990) as reasonable alternatives in most situations. Despite fitting both MA and SMA in many contexts, we have not yet encountered a situation where use of MA instead of SMA led to a qualitatively different interpretation of results, and we believe that such an instance would be exceptional. However, we emphasise that it is good practice to quote with a slope estimate the method by which it was obtained, as emphasised by Feigelson & Babu (1992). Different line-fitting methods estimate (slightly) different things about the data, so a slope estimate needs to be interpreted in the context of the method used to estimate it.

V INFERENCE FOR A SINGLE MA OR SMA LINE

In this section, we will discuss methods of inference about slope and elevation for a major axis or standardised major axis. Table 4 summarises the calculation formulas for the recommended methods.

(1) One-sample test of the slope

To test if the true slope is equal to some value b, a simple approach to use is to test if the residual and axis scores are uncorrelated, when these are calculated using b as the slope. For example, to test if the standardised major axis slope is equal to $\frac{2}{3}$, calculate the variables $Y+\frac{2}{3}X$ and $Y-\frac{2}{3}X$ and test the hypothesis that these variables are uncorrelated (as in Fig. 6). Using this approach leads to the standard F tests for the linear regression, major axis and standardised major axis cases (Draper & Smith, 1998; Creasy, 1957; Pitman, 1939). These test statistics are equivalent to the likelihood ratio tests derived assuming bivariate normality (Warton & Weber, 2002), and are exact* if errors from the

*An exact test has a test statistic that exceeds the critical value for the significance level p with probability exactly p.

line are normally distributed, although there is an additional assumption for exactness in the MA and SMA cases. For MA and SMA, the test does not make the distinction between whether it is the fitted axis or the residual axis that has a slope close to b, so it must be known *a priori* which of the sample axes is estimating the true MA/SMA axis. This is not a restrictive assumption in allometry, where it is usually known *a priori* whether a positive or negative relationship is to be expected, and there is usually an axis along which the vast majority of the variation is explained, as in Figures 1-3.

For the test that $b = 1$ (testing for isometry), the MA and SMA tests are mathematically identical. The test in this case is whether $Y - X$ is uncorrelated to $Y + X$, or in other words, if the data were rotated by 45^0, would the subsequent values be uncorrelated? This approach is related to Tukey's mean-difference plots (Chambers *et al.*, 1983), or in the medical literature, Bland-Altman plots (Bland & Altman, 1986).

(2) One-sample test for elevation

For all types of lines considered in this paper, the sample elevation is calculated so that the fitted line goes through the centroid of the data $(\bar{x}, \bar{y})$. This leads to the formula $\bar{y} - \hat{\beta}\bar{x}$ for all line-fitting methods. Another way to think about the sample elevation is as the sample mean of residual scores $Y - \hat{\beta}X$.

The sample elevation is approximately normally distributed, and so a one-sample t-test can be used to test if the true elevation is equal to some value a. Irrespective of whether linear regression, MA or SMA is used, the variance of $\hat{\alpha}$ is approximately

$$\frac{\sigma^2}{N} + \mu_x^2 \mathrm{var}(\hat{\beta})$$

where σ^2 is the variance of residual scores $Y - \hat{\beta}X$ when $\hat{\beta}$ is treated as fixed, and μ_x is the mean of the X variable (Robertson, 1974). This expression consists of two components – the first is due to uncertainty estimating the centroid $(\bar{x}, \bar{y})$, and the second is due to uncertainty estimating the slope $(\hat{\beta})$, as in Fig. 7. In practice, σ^2 and μ_x need to be

replaced by their sample estimates, which leads to the standard formula for the variance of elevation for linear regression (Draper & Smith, 1998, for example). The estimated variance of elevation is the same for SMA as for linear regression, because the variance of the slope is the same.

(3) Confidence intervals for slope and elevation

A confidence interval for a parameter can always be constructed based on a one-sample test for the parameter, by finding the range of values for which the one-sample test is non-significant at the chosen level of confidence. So, for example, a 95% confidence interval for a major axis slope could be constructed as the interval containing all values of b such that the correlation coefficient between the variables $bY + X$ and $Y - bX$ is not significantly different to zero at the 0.05 level. This is the method by which the expressions for confidence intervals in Table 4 were calculated, which are the recommended expressions for calculating confidence intervals (as in Jolicoeur & Mosimann, 1968; Jolicoeur, 1990, and elsewhere). We will refer to this as the exact method of calculating confidence intervals.

Several alternative methods exist for making inferences about MA or SMA slopes. Although some of the alternative methods usually work well, the exact method is preferred on theoretical grounds and on the basis of simulation work presented in appendix E. In brief:

- Given the variance of the MA or SMA slope, the t_{N-2} distribution can be used to find approximate confidence intervals (Ricker, 1973; Sokal & Rohlf, 1995; Quinn & Keough, 2002). This requires the assumption that the sampling distribution is normal, which is usually reasonable, although not exactly true for MA and SMA slopes. Hence confidence intervals are not exact, although they are a good approximation (Appendix E).

- Clarke (1980) derived confidence limits for $\log(\hat{\beta}_{\mathrm{SMA}})$, given that the sampling distri-

bution of SMA (and indeed MA) slope is closer to a log-normal distribution than to a normal distribution. This enables good approximate inference about SMA slopes in small samples (Clarke, 1980; McArdle, 1988).

- Isobe *et al.* (1990) proposed a method that is asymptotic (*i.e.* valid for large sample sizes) of making inferences about slope that makes less restrictive assumptions than the exact method. We found the asymptotic method performed poorly in small sample simulations for both normal and non-normal data, while the exact method was robust to non-normality, suggesting that it is not necessary to seek more robust methods of inference (Appendix E).

- Legendre & Legendre (1998) described a method of constructing confidence intervals for elevation, which uses the relation $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ to estimate confidence limits for $\hat{\alpha}$ by substituting in the upper and lower confidence limits for $\hat{\beta}$. This method treats $\bar{y}$ and $\bar{x}$ as fixed, *i.e.* it accounts for sample variation in estimating the slope (Fig. 7b) but not the centroid (Fig. 7a). This method performs very poorly (Appendix E) and should not be used, particularly when $\bar{x}$ is close to zero because in this case most sample variation in $\hat{\alpha}$ is due to uncertainty in estimating the centroid of the data.

VI INFERENCE FOR COMPARING SEVERAL MA OR SMA LINES

It is often of interest to compare several MA or SMA lines, as in the leaf allometry example (section II, Fig. 1b-d). The types of hypotheses of interest are analogous to analysis of covariance, although for lines calculated using MA or SMA estimation rather than using linear regression. Most of the testing procedures described here were proposed relatively recently (Warton & Weber, 2002) or are proposed in this paper.

Calculation formulas for the multi-sample tests are described in Appendix D. These

formulas tend to have a more complicated form than for one-sample tests.

(1) Testing for common slope

Testing for common slope amongst several lines, as in Fig. 1b, is a first step in making inferences about several lines. This test is a necessary preliminary to testing for equal elevation or no shift along the axis, given that such tests make little sense unless the lines being compared share a common slope. However, testing for common slope is of interest in its own right, because in allometry inferences about the slope of the line are usually of primary interest.

In the case of linear regression, an F-statistic is constructed to compare the sums of squares when a common slope is fitted and the sums of squares when each group is fitted with a regression line of different slope (Sokal & Rohlf, 1995, for example). One might think that a similar approach could be used in the MA and SMA cases, but with sums of squares defined differently, as suggested by Harvey & Mace (1982) and others. However, this statistic does not follow the F distribution needed for comparing several MA or SMA slopes (Appendix E, Table 11), presumably because the numerator and denominator sums of squares are not independent. Further, this method assumes equal residual variances across groups, which is not always reasonable, and does not need to be assumed in alternative test statistics (Clarke, 1980; Flury, 1984; Warton & Weber, 2002).

We recommend using a likelihood ratio test for common MA or SMA slope (Flury, 1984; Warton & Weber, 2002). Calculation details can be found in Appendix D. The test was derived assuming bivariate normality, although it is known to be robust to non-normality (Warton, in review). This same test was proposed for MA slopes in the appendix of Harvey & Mace (1982) and attributed to J. Felsenstein, although an algorithm for estimating the common slope was not described. Flury (1984) developed common principal components analysis, and described a method of common slope estimation that can be used for major axes (and can also be used for more than two variables). Warton & Weber (2002) modified

the method due to Flury (1984) for the bivariate standardised major axis case. Warton & Weber (2002) also demonstrated that when the test statistic is used with Bartlett corrections, it is well approximated by the chi-squared distribution even when sample sizes in each group average 10 and when data are not bivariate normal, but errors from the line are normally distributed.

When using the likelihood ratio tests of Flury (1984) or Warton & Weber (2002), the common MA/SMA slope estimator does not have a closed form solution, and so is calculated by iteration (and available software does this in negligible time). Alternative slope estimators could be used, for example, the pooled sums of squares could be calculated across groups and the standard slope estimator used (Krzanowski, 1984), which is analogous to the estimator of the linear regression common slope. This has the advantage of being simpler to calculate, although it has the disadvantage of making more restrictive assumptions – pooling sums of squares implicitly assumes that the variance matrix is the same for all groups, and the procedure performs poorly when groups have the same slope but different variances or correlations (as in simulations, Appendix E Table 11). In practice, variances and correlations can be quite different for different groups (as in the example in Warton & Weber, 2002), so pooling of sums of squares can not be recommended in general.

An alternative method of comparing two SMA slopes was proposed by Clarke (1980), and reviewed in McArdle (1988). The test maintains close to exact significance levels in small samples, as does the likelihood ratio test (Warton & Weber, 2002), although the test due to Clarke (1980) does not compare more than two SMA slopes. The method due to Clarke (1980) could be modified to compare several SMA slopes using a Wald test, along the lines of the test for equal elevation described later. While this method is a reasonable alternative, we lean towards using the likelihood ratio test, given that it is a single procedure that can be used for both MA and SMA, and it is known to have good properties.

Warton & Weber (2002) describes a procedure for testing for common slope in a more general context than MA and SMA – when the error variance ratio is unknown. The error variance ratio determines the direction of the residual axis, which can be anywhere from vertical to horizontal. Different choices of error variance ratio lead to different slope estimates and (usually slightly) different results in multi-sample inference. Issues relating to choice of error variance have been discussed previously. If one is uncomfortable with the choice of MA vs SMA or an alternative slope estimator, then testing for common slope with an unknown error variance ratio offers a conservative test, giving the smallest possible test statistic across all possible choices of error variance ratio.

A confidence interval can be constructed for the common slope, using likelihood ratio techniques. The method has not previously been described in detail in the literature, although it is relatively straightforward in principle. A likelihood ratio statistic for testing if the common slope equals some value b ($H_0 : \beta_i = b \quad H_a : \beta_i = \beta \neq b$ for each i), can be constructed, as in Appendix D. A confidence interval is then the interval containing all possible values of b for which the test statistic is not significant (at the chosen level of confidence). Although the confidence limits are not easily written in closed form, they can be computed using an optimisation algorithm.

(2) Testing for common elevation

As was described in relation to Fig. 1c, it is often of interest to test for equal elevation amongst several lines that have been fitted with MA or SMA lines of common slope. We propose using a Wald statistic for inference, as described in Appendix D, in preference to the F-statistic that has received some attention in the literature. We will briefly describe the F statistic first, explain its problems, then describe the Wald statistic.

Harvey & Mace (1982) and others suggested using analysis of variance of residual scores as a test for common elevation, which we refer to as using an F statistic. This method was used by Wright, Reich & Westoby (2001). The reasoning behind this approach is

that the sample elevation can be calculated as the sample mean of the residual scores, $Y - \hat{\beta}X$, so testing for equal means of the residual scores is equivalent to testing for equal elevation. Note that in calculating the residual scores, $\hat{\beta}$ is the estimated common slope, because the lines being compared must have common slope for the test of elevation to be meaningful. As far as we are aware, no other procedure for comparing elevations of MA/SMA has previously been proposed.

The problem with the F-test is that while it accounts for uncertainty in estimating the centroid of each group, it does not consider uncertainty estimating the common slope – in Fig. 7, (a) is accounted for but (b) is ignored. This is not a problem if the mean of the X variable is the same for all groups. However, the F statistic can depart considerably from an F-distribution when there is a common elevation, but the means of the X variable differ systematically among groups (Appendix E). In this situation the error estimating $\hat{\beta}$ has different consequences for the sample elevation of groups with different X means, such that the sample elevations have unequal and correlated error. In contrast, the F statistic assumes that the sample elevations have equal and uncorrelated error. Resampling in an appropriate fashion (as described in Appendix F) can ensure valid inference despite this, however a simpler alternative is to use a different test statistic.

A Wald statistic, as described in Appendix D, can be recommended for comparing several elevations. This is a simple type of statistic that is traditionally used for inference (Rao, 1973, for example) in situations such as this one, when it is difficult to calculate the null likelihood function hence the likelihood ratio statistic (a statistician's first choice test statistic). Wald statistics are commonly encountered – they appear in the standard output from most statistics packages for multiple linear or logistic regression (often labelled "t" and "Z" statistics, respectively). A Wald statistic simply tests if parameter estimates are significantly far from their hypothesised values, by comparing the distance from hypothesised values to its standard error. In testing for common elevation, there are several parameters of interest, so the Wald statistic involves a vector of parameters and their variance matrix. By using the correct formula for the variance matrix, which is a

multivariate version of $s^2_{\hat{\alpha}}$ from Table 4, the Wald statistic incorporates both the sources of uncertainty illustrated in Fig. 7. This statistic was demonstrated in simulations to maintain close to exact Type I error (Appendix E) irrespective of possibly unequal sample sizes, correlations, and means of the X variable.

(3) Testing for no shift along a common axis

Fig. 1d describes a situation in which it was believed that data from two sites were scattered around a common axis, with no difference in elevation, but it was hypothesised that there might be a shift along the axis. It was believed that both sites would contain species sharing a common trade-off between the two leaf traits, however in the higher nutrient site, species would generally have shorter-lived leaves with higher leaf mass per area.

We propose a Wald test for equal mean fitted axis scores. The fitted axis scores measure the location of a point along the fitted axis, so it is natural to test for shift along the axis using the mean fitted axis score of each group. As explained in Appendix D, this is equivalent to a test for equal elevation of the (standardised) minor axis of each group, *i.e.* of the line fitted through the centroid which is in the direction of the residual axis. Consequently, the method of testing is very similar to testing for common elevation (Appendix D).

Wright *et al.* (2002) tested for no shift along a common axis using an analysis of variance of the axis scores. This was done given the lack of an alternative procedure, and is not recommended. This procedure does not account for sampling error in estimating the common slope, as for the F statistic for common elevation, and so this procedure is sensitive to differences in means of the X variable that are not attributable to shifts along the fitted axis.

VII INFERENCE FOR RELATED LINE-FITTING METHODS

This section describes how the methods of inference described in this paper can be modified for use with related line-fitting methods: MA or SMA without an intercept, and in the presence of measurement error.

(1) MA or SMA with no intercept

If MA or SMA lines are forced to go through the origin (as for divergence data), the methods of inference described in this paper can still be used, after two simple changes:

- The averages of the X and Y variables are set to zero in all calculation formulas. This affects calculation formulas via changes to the sample variances and covariance of X and Y, which become sums of squares and products of x_i and y_i rather than of $(x_i - \bar{x})$ and $(y_i - \bar{y})$.
- All terms in calculation formulas of the form $(N-2)$ or $(n_i - 2)$ should be replaced by $(N-1)$ and $(n_i - 1)$, respectively. These terms represent the residual degrees of freedom. Because the elevation is fixed at zero, there is one less parameter to be estimated, so the residual degrees of freedom are increased by one. Note that this only affects one-sample inference about the slope, because the recommended methods of multi-sample inference about slopes do not involve using a t or F distribution.

For example, when given N pairs of phylogenetically independent contrasts $(x_1, y_1), \ldots, (x_N, y_N)$, the estimated SMA slope with the line forced through the origin is

$$\hat{\beta}_{\mathrm{SMA}} = \sqrt{\frac{\sum_{i=1}^{N} y_i^2}{\sum_{i=1}^{N} x_i^2}}$$

A $100(1-p)\%$ confidence interval for β has the same form as previously:

$$\hat{\beta}(\sqrt{B+1} \pm \sqrt{B})$$

except now $B = \frac{1-r_{xy}^2}{N-1} f_{1-p,1,N-1}$, where r_{xy} is the sample correlation coefficient calculated without centering the data:

$$r_{xy}^2 = \frac{\left(\sum_{i=1}^{N} x_i y_i\right)^2}{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2}$$

The only changes that have been made here to the relevant formulas of Table 4 are that sample means have been replaced with zero, and $N - 2$ has been replaced by $N - 1$.

Note that if the lines are forced through zero, it is no longer of interest to make inferences about elevation, because the elevation of all lines is exactly zero.

(2) MA or SMA adjusting for measurement error

Methods-of-moments MA was proposed by Akritas & Bershady (1996, section 3) and discussed together with methods-of-moments SMA in section III. Akritas & Bershady (1996) presented a variance formula for methods-of-moments MA, which can be extended to methods-of-moments SMA using the relevant formula in Table 1 of Isobe *et al.* (1990).

A simple alternative to the method of Akritas & Bershady (1996) is to use the recommended methods of inference for when there is no measurement error, after adjusting the variance estimates to account for measurement error. This is in fact the method that Akritas & Bershady (1996) used, although they modified the formulas from Isobe *et al.* (1990) rather than the formulas for exact methods.

Simulations in Appendix E suggest that it is better to use the asymptotic method due to Akritas & Bershady (1996) than to modify exact methods of inference, but only in moderate to large samples ($N > 30$). In small samples where there is non-negligible measurement error, it may be necessary to use resampling, given that no suitable alternative has been found. Carroll *et al.* (1995, Appendix A.6) provide some guidelines for resampling measurement error models.

Further studies investigating methods of inference for measurement error models would

be useful. In particular, we only conducted simulations to consider confidence interval estimation of methods-of-moments MA or SMA slopes, and we did not consider other types of inference problems.

VIII ROBUSTNESS OF INFERENTIAL PROCEDURES TO FAILURE OF ASSUMPTIONS

Several assumptions are made in the above inferential procedures, and it is important to know how sensitive inferences are to these assumptions, and what can be done to check that the assumptions are satisfied. These issues are summarised in Table 5, drawing on general design principles (Cox, 1958), robustness of linear models in general (Miller, 1986), and recent work on the robustness of inferences about MA and SMA (Warton, in review).

Some of the points in Table 5 require some further elaboration:

- Normality is the least important assumption, because the central limit theorem ensures robustness to failure of this assumption (see Appendix E for examples of this robustness). However, an important consideration with non-normality is loss of power – least squares methods have low power for data from long-tailed distributions (Staudte & Sheather, 1990, for example).

- Resampling will rarely help ensure robustness to failure of assumptions for linear models. Resampling algorithms are described in Appendix F, but these implicitly assume independence, linearity and equal variance. Even when residuals are non-normal, resampling only ensures valid inference, it does not generally ensure higher power, if the same test statistic is used. Resampling might be useful for small samples, when residuals are moderately non-normal, but its main use is for inference when alternative methods are unavailable, not as a method of robust inference.

- Robust alternatives to MA and SMA could be drawn from the principal components

analysis literature (reviewed by Jolliffe, 2002, section 10.4). There are two general approaches for robust principal components lines – using robust estimators of the variance matrix, or using different criteria for line estimation. A simple example of using robust estimates of variances is to calculate the robust SMA slope as the ratio of median absolute deviations, and calculate the elevation so that rather than passing through the centroid, it passes through the point $(\tilde{x}, \tilde{y})$ where $\tilde{x}$ and $\tilde{y}$ are sample medians rather than means. An example of an alternative line-fitting criterion is to estimate a robust MA slope by finding the rotation of the data for which some robust measure of correlation is zero (Isler *et al.*, 2002). In using such techniques, resampling-based inference may be required, because of a lack of development of alternative methods of inference.

IX SOFTWARE

Most of the inferential procedures described in this paper are not available in standard software packages. With the exception of one-sample tests of slope, specialised computer software is required for all methods of inference. Such software is available in several formats:

- As a stand-alone package known as (S)MATR (Falster *et al.*, 2003), with accompanying documentation `http://www.bio.mq.edu.au/ecology/SMATR/`.
- As spreadsheet formulas in Microsoft Excel.
- As an R package.
- As a Matlab toolbox.

All software can be found from the first author's website `http://web.maths.unsw.edu.au/~dwarton/programs.html`. The Excel spreadsheets do not include software for testing hypotheses about elevation or shift along the axis.

X CONCLUSIONS

1. In selecting a method of line-fitting, it is essential to consider what the line is to be used for and whether measurement error is negligible or not (Table 1).

2. The distinction between equation and measurement error has often been made in previous reviews, however we believe that the different consequences of the different sources of error has not been sufficiently appreciated.

3. When equation error is present, the type of line-fitting method to use is determined by the research question of interest, as in Table 1, and not by estimates of error magnitude.

4. When measurement error is present, its magnitude should be estimated. This allows the impact of measurement error on results (in particular, on the estimated slope) to be considered and corrected for, if required.

5. For the first time, users of the major axis and standardised major axis have available a set of tools for the most commonly encountered tasks in allometric analysis – for inference about the slope or elevation of a single line, and for comparing several lines in a framework analogous to analysis of covariance.

6. We have described how to modify methods of inference for use when data are phylogenetically independent contrasts or are measured with error.

7. There are several areas where further methodological research would be useful:

    - Improved inference about method-of-moments lines – currently, methods of inference about such lines are only approximate, and in small samples the approximation can perform poorly (Appendix E).
    - Robust alternatives to MA and SMA – the MA and SMA methods are special cases of least squares approaches. Such methods of estimation are known to lack robustness to outliers, hence they are inefficient for long-tailed distributions.

- Inference for SMA when there are more than two variables – whereas the methods of inference for MA have natural extensions to more than two dimensions, the data standardisation for SMA complicates inference.

XI ACKNOWLEDGEMENTS

Thanks to Neville Weber and Brian McArdle for helpful discussions during early stages of this work.

A Terminology

This appendix briefly reviews some alternative terminology used for line-fitting methods. Methods of line-fitting have been known under a variety of names, as reviewed in Table 6. A few of these terms are misleading, hence not recommended for general use:

- "Model II regression" (Sokal & Rohlf, 1995). This commonly used term is misleading on two counts – the method does not involve regression, and is not model II, in the senses in which these terms were first used. The term "regression" was first used because of the property of "regression to the mean", a property that the user is trying to avoid with MA and SMA methods. The term "model II" was originally suggested under the belief that what distinguishes MA and SMA methods from linear regression is that the X variable is random not fixed, which is the difference between model I and model II ANOVA as defined by Eisenhart (1947). However, the distinction is not random vs fixed X, it is that a line-of-best-fit is required rather than a line for predicting Y. It is true that the X variable must be random to use MA or SMA, but it is not the case that X has to be fixed for linear regression (for more details, see section reviewing regression). A closer analogy to MA or SMA is ANOVA where between group differences are in part attributed to misclassification

of subjects (this is rare in practice).

- "Errors-in-variables models" is a term used widely in the statistical literature for line-fitting when there is measurement error in both variables. However, MA and SMA are not used because there is measurement error (as described previously).

- "Functional relationship" and "structural relationship" are terms that were used to describe major axis and standardised major axis methods in the statistics literature throughout most of the second half of the 20th century. However, these terms have different meanings in other literatures. For example, to many biologists, a functional relationship is a causal relationship.

There are two further terms that are not misleading, however more appropriate terms are available:

- "Geometric mean functional relationship" is an alternative to "standardised major axis" that gets its name from the fact that the slope estimator is the geometric mean of the two linear regression slope estimators. While this is an interesting property, the method is more naturally described as a standardised version of the major axis.

- "Reduced major axis" is not as specific a term as "standardised major axis", which makes it clear that the modification that has been made to the major axis method is standardisation of data. According to Jolicoeur (1975), the term "reduced" was introduced by Kermack & Haldane (1950) as an inaccurate translation of the French work "reduit", which is better translated as "standard" or "standardised" than as "reduced".

B Equation error estimation for the standardised major axis

This appendix briefly describes how equation error is treated in the standardised major axis model.

The standardised major axis can be considered to measure residuals in a direction that is the fitted standardised major axis reflected about the Y axis. This interpretation has not often been used in the past, however it can be derived as follows. For standardised bivariate data, the major axis always has slope either 1 or -1, or more specifically, $sign(r)$ (the sign of the correlation coefficient multiplied by 1). The residuals are measured perpendicular to this line, and so have slope $-sign(r)$. Rescaling to the original axes, the slope of the fitted line is $sign(r)\frac{s_y}{s_x}$, the ratio of standard deviations with the same sign as the correlation coefficient, and the residuals are measured in the direction $-sign(r)\frac{s_y}{s_x}$, which is the fitted line reflected about the Y axis. This result was implicitly used in deriving methods of inference that are used for the standardised major axis (Pitman, 1939; Warton & Weber, 2002).

A less useful derivation of the standardised major axis finds the line that minimises the sum of triangular areas between the line and each data point (Teissier, 1948). This derivation is best considered a geometric peculiarity – statistical theory offers no reason why you would wish to minimise such a quantity.

C Estimating measurement error variance

This appendix outlines how to estimate the variance of measurement errors, or the average measurement error variance if this differs across subjects. If the variable being measured is X, and it is measured with error as $X + \delta$, then we would like to estimate a quantity

s^2_δ such that an unbiased estimator of the variance of X is

$$s^2_{X+\delta} - s^2_\delta$$

where $s^2_{X+\delta}$ is the sample variance of the X subjects that were measured with error. Note that δ is used in this appendix rather than δ_X, which was used in the remainder of the manuscript. To estimate s^2_δ, the only requirement is that repeated measurements of $X+\delta$ must be available.

To define some notation – let us say that N values of X were sampled, and random variables representing each of these N subjects are $X_1+\delta_1$, $X_2+\delta_2$, ..., $X_N+\delta_N$. There are n_i repeated measurements of the ith subject $(X_i+\delta_i)$, where the number of repeated measurements is not necessarily equal for all subjects (and so n_i is a function of i). The repeated measurements of $X_i+\delta_i$ measured with error are $x_i+\delta_{i,1}$, $x_i+\delta_{i,2}$, ..., $x_i+\delta_{i,n_i}$.

Note that measurement error variance should be estimated on the scale on which the lines are to be fitted. For example, in Fig. 2, error variance is estimated for log(brain mass) not for brain mass.

Note also that it can require careful thought to determine what units represent replicate measures of a subject – individuals of a species from throughout their known distribution (for Fig. 2) or repeated measurements on an individual (for Fig. 3), *etc.*

The repeated measurements are most often averaged to estimate $X_1+\delta_1$, $X_2+\delta_2$, ..., but this does not need to be the case. In particular, the log of a variable might be analysed (as in Figures 2 and 3), but the species means might be estimated before log transformation, to preserve the interpretation of the species value as an average. In this appendix, methods of estimating the average measurement error variance s^2_δ are described for the following cases:

- When all $Var(\delta_i)$ are equal.
- When not all $Var(\delta_i)$ are equal.

- In a leaf mass per area example, where repeated measurements are not averaged on the log scale.

(1) All measurement errors have equal variance

The first case to be considered is the simplest one – where the variance of measurement error on each subject is equal.

In this case, the variance of measurement error can be estimated using one-factor analysis of variance. The factor used in analysis is the subject (X_1, X_2, ...), which is species in Fig. 2 or individual in Fig. 3. The observations within groups are the measurements taken of the subject, which are measurements of log(brain or body mass) for different individuals of a species in Fig. 2, or remeasurements of log(height or basal diameter) for the same individual in Fig. 3. An estimate of average measurement error variance can be obtained using the mean squared error of the analysis of variance (MSE) as

$$s_\delta^2 = \frac{MSE}{N} \sum_{i=1}^{N} \frac{1}{n_i}$$

(2) Measurement error variances not equal

It is often the case that measurement errors do not have equal variance for different values of X. For example, in measuring average seed weight for an individual plant, measurement error will largely be due to the measurement process itself (rather than due to sampling error). The standard error of repeated measures on a set of scales might be 0.05 grams, for example, irrespective of the size of the seed. While this is not a function of seed weight, it will be a function of log(seed weight), as 0.05 grams is relatively larger for small seeds than for large seeds.

To estimate average measurement error variance, the sample variances of repeated mea-

sures of each subject are first calculated, $s^2_{\delta_1}, s^2_{\delta_2}, \ldots, s^2_{\delta_N}$. These are then combined as:

$$s^2_\delta = \frac{1}{N}\sum_{i=1}^{N}\frac{s^2_{\delta_i}}{n_i} \tag{1}$$

If the same number of repeated measures are taken for all subjects, this estimate will be the same as the one proposed in the previous section. Riska (1991) and Akritas & Bershady (1996) independently arrived at this estimator of measurement error for linear regression.

(3) When the data are not averages of repeated measures

On some occasions, a variable used in line-fitting might be a function of repeated measures, but it might not simply be an average of these measures on the scale on which data are analysed. In these cases a simple formula such as equation (1) does not apply. Alternatives are either to derive an alternative formula, or to use a resampling approach to estimate measurement error variance.

One approach that can be used to arrive at an estimate of measurement error variance is to use the following result:

$$Var(g(X)) \approx g'(\mu)^2 Var(X)$$

where $g'(\mu)$ is the derivative of the function $g(\cdot)$ evaluated at μ (Kendall & Stuart, 1969, sections 10.6 and 10.7). This result is only an approximation, made by assuming that over the range of values of X that are observed, $g(\cdot)$ is approximately linear. The approximation will work better when X is less variable and when $g(\cdot)$ is closer to linear.

Using this result, if repeated measurements are averaged on the untransformed scale, and then averages are used in analysis on the log scale,

$$s^2_{\delta_i} \approx \frac{1}{\bar{x}_i^2} s^2_{x,i} \tag{2}$$

where $\bar{x}_i$ is the sample mean and $s^2_{x,i}$ is the sample variance of the repeated measures for

the ith subject. If data are transformed using $\log_{10}(\cdot)$ for analysis, then $s^2_{\delta_i}$ is approximated as $(\log_{10} e)^2$ multiplied by the expression in equation (2).

(4) Example – log(LMA) using species averages

Consider the calculation of measurement error when estimating leaf mass per area (LMA) in kg/m^2 for the data used in Wright, Reich & Westoby (2001). Measurements were taken of LMA and other leaf traits, for about 20 species in each of four sites. Several individuals were sampled of each species, and the average of LMA calculated for each species. Standardised major axes were estimated to find the slope of the line of best fit relating LMA and a measure of photosynthetic rate (A_{mass}), when both variables were log-transformed.

Although the log-transformation was used for analysis, the species means and measurement error were not estimated from repeated measurements on the log-transformed data. Instead, species means were estimated on the untransformed scale (kg/m^2 for LMA), for two reasons:

- So that the summary statistic for species is average LMA value – LMA being the variable of interest, rather than log(LMA), and the average being a statistic that is simpler to interpret than the back-transformed average or "geometric mean".
- The distribution of repeated measurements of LMA within a species is not usually long-tailed or strongly skewed, and so the sample mean is usually a good summary statistic.

The averaging on the untransformed scale slightly complicates estimation of the measurement error variance $s^2_{\delta_i}$ for log(LMA) – equation (2) needs to used.

Table 7 gives a summary of the repeated measures of LMA for each species, for one of the sites sampled by Wright *et al.* (2001). The term s_{δ_i} is estimated using equation (2), with

a correction factor of $(\log_{10} e)^2$, because $\log_{10}$(LMA) is used in further analyses in Wright *et al.* (2001) rather than $\log_e$(LMA).

The average measurement error variance s^2_δ, can be calculated using equation (1), from the values of n_i and $s^2_{\delta_i}$ in Table 7. The estimated value is 0.00014. This is small compared to the sample variance of the log(LMA) values for the 18 species means, $s^2 = 0.0034$.

For all species, the sample mean and measurement error variance are similar to the values they would otherwise be if the repeated measurements were log-transformed before averaging. However, this may not always be the case.

An alternative method of measurement error estimation would be to estimate the terms $s^2_{\delta_i}$ by resampling. For each species, the repeated measurements could be resampled with replacement, the sample mean reestimated for each resampled dataset, and the variance of the log-transformed means across the resampled datasets could be used to estimate $s^2_{\delta_i}$.

For the other three sites and other variables measured in Wright *et al.* (2001), we similarly found that measurement error variance was small compared to sample variance (usually $<$ 3%, but 10% in one case). Measurement error was larger for log(A_{mass}) than for log(LMA), and so SMA slopes of log(A_{mass}) vs log(LMA) were slightly flatter when accounting for measurement error (the most substantial reduction being from a slope of 1.24 to 1.15).

D Calculations for multi-sample tests

In this appendix, we describe calculation formulas that can be used to conduct tests for comparing the lines estimated from several independent samples (section VI). First we will define some terms that are used in the calculation formulas below.

We will assume there are g groups, and that the ith group consists of n_i pairs of observations $(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \ldots, (x_{in_i}, y_{in_i})$, and that the total sample size is $N = \sum_{i=1}^{g} n_i$. We will use standard definitions of the sample mean, variance, covariance, and correlation

coefficient of the X and Y variables in each group:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

$$s^2_{x,i} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \qquad s^2_{y,i} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

$$s_{xy,i} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i) \qquad r_i = \frac{s_{xy,i}}{s_{x,i}s_{y,i}}$$

We will also define $s^2_{\mathrm{f},i}(b)$, $s_{\mathrm{rf},i}(b)$ and $s^2_{\mathrm{r},i}(b)$, the sample variances and covariances of fitted axis and residual scores from a line of slope b, for the ith group. The residual axis from a line of slope b can be defined as $Y - bX$, so the sample variance of residual scores is

$$s^2_{\mathrm{r},i}(b) = s^2_{y,i} - 2bs_{xy,i} + b^2 s^2_{x,i}$$

This should be multiplied by the factor $\frac{n_i-1}{n_i-2}$ when b is estimated from the sample data (Sprent, 1969), although this will make negligible difference in large samples. For MA, the fitted axis is $bY + X$, and for SMA the fitted axis is $Y + bX$, so

$$s^2_{\mathrm{f},i}(b) = \begin{cases} b^2 s^2_{y,i} + 2bs_{xy,i} + s^2_{x,i} & \text{for MA} \\ s^2_{y,i} + 2bs_{xy,i} + b^2 s^2_{x,i} & \text{for SMA} \end{cases}$$

$$s_{\mathrm{rf},i}(b) = \begin{cases} s_{xy,i} + bs^2_{y,i} - bs^2_{x,i} - b^2 s_{xy,i} & \text{for MA} \\ s^2_{y,i} - b^2 s^2_{x,i} & \text{for SMA} \end{cases}$$

and as previously, these terms are multiplied by the factor $\frac{n_i-1}{n_i-2}$ if b is estimated from the data. (In passing, note that the MA or SMA slope for the ith group is the value $\hat{\beta}_i$ that satisfies $s_{xy,i}(\hat{\beta}_i) = 0$, *i.e.* the value that ensures residual and axis scores are uncorrelated.)

(1) Common slope test

To conduct a test for common slope, first the common slope must be estimated. A maximum likelihood estimate for the common slope is the value $\hat{\beta}$ satisfying the equation

(Warton & Weber, 2002)

$$0 = \begin{cases} \sum_{i=1}^{g} n_i \left(\frac{1}{s^2_{\mathrm{r},i}(\hat{\beta})} - \frac{1}{s^2_{\mathrm{f},i}(\hat{\beta})} \right) s^2_{\mathrm{rf},i}(\hat{\beta}) & \text{for MA} \\ \sum_{i=1}^{g} n_i \left(\frac{1}{s^2_{\mathrm{r},i}(\hat{\beta})} + \frac{1}{s^2_{\mathrm{f},i}(\hat{\beta})} \right) s^2_{\mathrm{rf},i}(\hat{\beta}) & \text{for SMA} \end{cases}$$

This equation can be solved iteratively, given an initial estimate (such as that suggested by Krzanowski, 1984, below). In iteration, the current estimate of $\hat{\beta}$ is used to calculate $s^2_{\mathrm{r},i}(\hat{\beta})$ and $s^2_{\mathrm{f},i}(\hat{\beta})$, then these values are plugged into the above equation, which is solved for the new estimate of $\hat{\beta}$ (Warton & Weber, 2002).

To test if there is a common slope, a (Bartlett-corrected) likelihood ratio statistic (Warton & Weber, 2002) is

$$-\sum_{i=1}^{g}(n_i - 2.5) \log \left(1 - r^2_{\mathrm{rf},i}(\hat{\beta})\right) \sim \chi^2_{g-1}$$

In the context of estimating common principal components, Krzanowski (1984) suggested that instead of estimating the common slope iteratively by maximum likelihood, it could be found in a single step by pooling sums of squares:

$$\tilde{\beta} = \begin{cases} \frac{1}{2s_{xy,\mathrm{p}}} \left(s^2_{y,\mathrm{p}} - s^2_{x,\mathrm{p}} + \sqrt{(s^2_{y,\mathrm{p}} - s^2_{x,\mathrm{p}})^2 + 4s^2_{xy,\mathrm{p}}} \right) & \text{for MA} \\ \mathrm{sign}(s_{xy,\mathrm{p}}) \frac{s_{y,\mathrm{p}}}{s_{x,\mathrm{p}}} & \text{for SMA} \end{cases}$$

where

$$s^2_{x,\mathrm{p}} = \sum_{i=1}^{g}(n_i - 1)s^2_{x,i} \qquad s_{xy,\mathrm{p}} = \sum_{i=1}^{g}(n_i - 1)s_{xy,i} \qquad s^2_{y,\mathrm{p}} = \sum_{i=1}^{g}(n_i - 1)s^2_{y,i}$$

Note, however, that a test statistic using this pooled estimator will not have a chi-square distribution when residual variances are different in different groups, as demonstrated in simulations (Table 11).

An alternative test proposed by Harvey & Mace (1982) and others is to use an F test analogous to that in the linear regression case. A measure analogous to sums of squares for a line of slope b fitted to the ith group is

$$SS(b, i) = (n_i - 2)k(b)s^2_{\mathrm{r},i}(b)$$

where $k(b)$ is a correction factor so that residual scores are measured on an appropriate scale (mathematically, so that the Jacobian matrix of the transformation to residual and axis scores has determinant 1):

$$k(b) = \begin{cases} 1 + b^2 & \text{for MA} \\ 2b & \text{for SMA} \end{cases}$$

In the major axis case, $SS(b, i)$ is the minimum possible sum of squares of distances from the points to a line of slope b, for the ith group. In the standardised major axis case, $SS(b, i)$ is the analogous term calculated on standardised data then back-transformed, or equivalently, it is the sum of triangular areas discussed by Teissier (1948).

An F test is then constructed to compare fitting a common slope ($\tilde{\beta}$, estimated by pooling sums of squares) to fitting each group with its own slope ($\hat{\beta}_i$):

$$\frac{(N - 2g) \sum_{i=1}^{g} \left(SS(\tilde{\beta}, i) - SS(\hat{\beta}_i, i) \right)}{(g - 1) \sum_{i=1}^{g} SS(\hat{\beta}_i, i)}$$

While in the linear regression case a test statistic of this form is distributed as $F_{g-1,N-2g}$, simulations have shown that this distribution is generally a poor approximation when comparing the slopes of MA or SMA lines (Table 11).

(2) CI for common slope

If it is believed that there is a common slope, to test if the common slope is equal to b, a (Bartlett corrected) likelihood ratio test statistic is

$$\sum_{i=1}^{g} (n_i - 2.5) \log \left(\frac{s^2_{\mathrm{f},i}(b)\, s^2_{\mathrm{r},i}(b)}{s^2_{\mathrm{f},i}(\hat{\beta})\, s^2_{\mathrm{r},i}(\hat{\beta})} \right) \sim \chi^2_1$$

A $100(1 - p)\%$ confidence interval for the common slope can be estimated by finding the range of values for b for which this test statistic is non-significant at level p. Use of this approach to constructing confidence intervals was suggested by Warton & Weber (2002).

Solving for the confidence interval is best done using an optimisation routine.

(3) Test for common elevation

Two statistics will be described here for testing for common elevation – an F statistic and a Wald statistic. The Wald statistic makes less restrictive assumptions and was shown in simulations (Appendix E) to maintain nominal significance levels in a range of scenarios we consider to be realistic. Hence the Wald statistic is recommended for general use.

The F statistic described here is the test statistic for an analysis of variance of residual scores $Y - \hat{\beta}X$, but with denominator degrees of freedom of $N - g - 1$ not $N - g$. As previously, the term $\hat{\beta}$ is an estimator of the common slope. An F statistic of this form can be used because the elevation is the sample mean of the residual scores, $\hat{\alpha}_i = \bar{y}_i - \hat{\beta}\bar{x}_i$. If $\bar{\alpha} = \frac{1}{N}\sum_{i=1}^{g} n_i\hat{\alpha}_i$, then the F statistic can be written as

$$\frac{(N-g-1)\sum_{i=1}^{g} n_i(\hat{\alpha}_i - \bar{\alpha})^2}{(g-1)\sum_{i=1}^{g}(n_i-2)s^2_{r,i}(\hat{\beta})}$$

This test statistic does not account for the possibility that the X means might not be equal, or that the residual variances (estimated by $s^2_{\mathrm{r},i}$) might not all be equal. The estimated variance of $\hat{\alpha}_i$ is

$$\frac{s^2_{\mathrm{r},i}}{n_i} + \bar{x}_i^2 s^2_{\hat{\beta}}$$

but an analysis of variance of residual scores ignores the second part of this expression, and replaces $s^2_{\mathrm{r},i}$ with a pooled estimate. This is only reasonable if the true mean of the X variable ($\mu_{x,i}$) is the same for all groups, and if all residual variances are the same.

If the $\mu_{x,i}$ can not be assumed to be equal, then a Wald statistic is more appropriate, as described in the following. First we will define the vector containing the g sample elevations as $\hat{\mathbf{A}} = [\hat{\alpha}_1 \ldots, \hat{\alpha}_g]'$, the g-vector of sample means of the X variable as $\bar{\mathbf{X}} = [\bar{x}_1 \ldots \bar{x}_g]'$, and the g-vector of residual mean standard errors as $s^2(\bar{\mathbf{R}}) = [s^2_{r,1}/n_1 \ldots s^2_{r,g}/n_g]'$. Then the variance matrix of the g sample elevations is approximately

$$s^2(\hat{\mathbf{A}}) = \mathrm{diag}(s^2(\bar{\mathbf{R}})) + \bar{\mathbf{X}}\bar{\mathbf{X}}' s^2_{\hat{\beta}}$$

where $\mathrm{diag}(\mathbf{v})$ indicates a diagonal matrix with the vector $\mathbf{v}$ along the diagonal, and where the variance of the estimator of the common slope ($s^2_{\hat{\beta}}$) is a function of the variances of

the one-sample estimates of slope (which are calculated using the formula in Table 4):

$$s_{\hat{\beta}}^{-2} = \sum_{i=1}^{g} s_{\hat{\beta}_i}^{-2}$$

This expression for the variance of the common slope estimator can be derived by calculating the Fisher information (Kendall & Stuart, 1973, section 18.16) of the common slope estimator.

If the null hypothesis is written in the form $H_0 : \mathbf{L}\hat{\mathbf{A}} = \mathbf{0}$ for some matrix $\mathbf{L}$, then the Wald statistic for testing H_0 is

$$(\mathbf{L}\hat{\mathbf{A}})' \left(\mathbf{L}s^2(\hat{\mathbf{A}})\mathbf{L}'\right)^{-1} (\mathbf{L}\hat{\mathbf{A}}) \overset{\text{approx}}{\sim} \chi^2_{g-1}$$

The chi-squared approximation applies if $\hat{\mathbf{A}}$ is normally distributed (Kendall & Stuart, 1969, section 15.10), which is true in large samples (Robertson, 1974, for example) and a reasonable approximation in small samples, as suggested by simulations (Table 13).

When testing for common elevations, a suitable choice of $\mathbf{L}$ is $\mathbf{L} = \left[\mathbf{1}_{(g-1)\times 1} \mid -\mathbf{I}_{(g-1)\times(g-1)}\right]$, where $\mathbf{1}_{a\times b}$ is the $a \times b$ matrix in which every element is one, and $\mathbf{I}_{a\times a}$ is the $a \times a$ identity matrix.

If an estimate of common elevation is desired, this can be calculated as

$$\hat{\alpha} = \frac{\mathbf{1}_{1,g} \left(s^2(\hat{\mathbf{A}})\right)^{-1} \hat{\mathbf{A}}}{\mathbf{1}_{1,g} \left(s^2(\hat{\mathbf{A}})\right)^{-1} \mathbf{1}_{g,1}}$$

This is the mean elevation, accounting for unequal and correlated error of the different sample errors.

Another statistic that could be used to test for common elevation is a likelihood ratio statistic, although estimation would not be straightforward in general, because there would need to be iteration between three different steps: estimating common slope, common elevation, and variance terms.

(4) Test for no shift along the fitted axis

To test for no shift along a common fitted axis, the F test analogous to that used for testing for common elevation is

$$\frac{(N-g-1)\sum_{i=1}^{g} n_i(\hat{\mu}_{\mathrm{f},i}-\bar{\mu}_{\mathrm{f}})^2}{(g-1)\sum_{i=1}^{g}(n_i-2)s^2_{\mathrm{f},i}(\hat{\beta})}$$

where $\hat{\mu}_{\mathrm{f},i}$ is the mean fitted axis score,

$$\hat{\mu}_{\mathrm{f},i} = \begin{cases} \hat{\beta}\bar{y}_i + \bar{x}_i & \text{for MA} \\ \bar{y}_i + \hat{\beta}\bar{x}_i & \text{for SMA} \end{cases}$$

and $\bar{\mu}_{\mathrm{f}} = \frac{1}{N}\sum_{i=1}^{g} n_i\hat{\mu}_{\mathrm{f},i}$.

A Wald statistic for testing for no shift along a common fitted axis has a similar form as when testing for common elevation. The covariance matrix of $\hat{\mathbf{M}}_{\mathrm{f}} = [\hat{\mu}_{\mathrm{f},1}\ \ldots, \hat{\mu}_{\mathrm{f},g}]'$ is

$$s^2(\hat{\mathbf{M}}_{\mathrm{f}}) = \begin{cases} \mathrm{diag}(s^2(\bar{\mathbf{F}})) + \bar{\mathbf{Y}}\bar{\mathbf{Y}}' s^2_{\hat{\beta}} & \text{for MA} \\ \mathrm{diag}(s^2(\bar{\mathbf{F}})) + \bar{\mathbf{X}}\bar{\mathbf{X}}' s^2_{\hat{\beta}} & \text{for SMA} \end{cases}$$

where $s^2(\bar{\mathbf{F}}) = [s^2_{\mathrm{f},1}/n_1\ \ldots\ s^2_{\mathrm{f},g}/n_g]'$, and the Wald statistic for testing $H_0 : \mathbf{L}\hat{\mathbf{M}}_{\mathrm{f}} = \mathbf{0}$ is

$$(\mathbf{L}\hat{\mathbf{M}}_{\mathrm{f}})' \left(\mathbf{L}s^2(\hat{\mathbf{M}}_{\mathrm{f}})\mathbf{L}'\right)^{-1} (\mathbf{L}\hat{\mathbf{M}}_{\mathrm{f}}) \overset{\text{approx}}{\sim} \chi^2_{g-1}$$

As previously, the Wald statistic is recommended in favour of the F statistic, because its distribution is not sensitive to unequal variances along different fitted axes, nor to shifts along the residual axis (*i.e.* differences in location along the X or Y axes that are not due to shifts along the fitted axis).

A test for equal average fitted axis mean is equivalent to a test for equal elevation of the (standardised) minor axis. For the SMA case, this is obvious, since the fitted axis mean is in fact the elevation of the standardised minor axis. In the MA case, the fitted axis mean is proportional to the Y-intercept of the minor axis, and hence the equivalence holds in this case also.

E Simulations

This appendix presents results of simulations that check the coverage probability of confidence intervals that have been proposed, and checks the Type I error of test statistics that have been proposed. In all cases, 95% confidence and the 5% significance level will be used.

As a general rule, the values of parameters used in simulations were in the range we consider to commonly arise in practice, although for sample size, values were chosen to be on the smaller side of this range. In all simulations, the variances of both variables were chosen to be equal, without loss of generality. Methods of inference are scale independent for SMA, and it can reasonably be assumed that properties of inferential methods do not change with scale for MA.

In all simulations, coverage probabilities and Type I error were estimated from 10000 datasets. This means that an exact confidence interval will usually (95% of the time) have an estimated coverage probability in the range (94.6%, 95.4%), and an exact test will usually (95% of the time) have an observed Type I error rate in the range $(4.6\%, 5.4\%)$.

The sign of the slope was assumed known to be positive *a priori* in all simulations, given that the sign of the slope is almost always known in practice. Hence coverage probability was estimated for the portions of (primary or secondary) confidence interval that were positive, and the (standardised) minor axis was used in calculations in the rare instances when it had positive slope, rather than using a (standardised) major axis with negative slope.

Three distributions were considered in simulations – bivariate normal with variances 1, a 9:1 mixture of bivariate normal distributions with variances 1 and 4, and a 9:1 mixture with variances 1 and 9. Note that the last of these distributions is particularly long-tailed – its kurtosis coefficient is larger than that for the double exponential, for example. For conciseness, results are often presented for the bivariate normal distribution only, and it

can be noted that the effect of using a different distribution was negligible.

(1) Confidence intervals for slope

Simulations were conducted to measure the coverage probability of confidence intervals for slopes of the major axis (β_{MA}) and the standardised major axis (β_{SMA}), using the exact method described in this paper (Pitman, 1939; Creasy, 1957; Jolicoeur & Mosimann, 1968), using the t_{N-2} distribution with variances taken from Table 4, and using a t_{N-2} distribution but with variance estimates taken from Isobe *et al.* (1990). The exact method gives exact confidence intervals if sufficient information is known *a priori* to distinguish the (standardised) major and minor axes, and if residuals are normally distributed. Other methods are asymptotic, *i.e.* they approach exactness as sample size increases.

Results demonstrate that although the exact method is only exact for normally distributed residuals, it remained very close to nominal levels for the non-normal distributions considered here, even for small samples (Table 8). This justifies recommendation of this method in practice. It is also apparent from Table 8 that although the t_{N-2} method is not exact, it still works very well. In these simulations, coverage probabilities for the t_{N-2} method were practically indistinguishable from those for the exact method, except for slightly liberal confidence intervals for β_{MA} in small samples. The method due to Isobe *et al.* (1990) performed reasonably except when $N = 10$, and coverage probabilities were not as close to nominal levels as for alternative methods, so this approach is not recommended.

(2) Confidence intervals for the slope when the line if fitted through the origin

Some simulations were conducted to demonstrate that when the line is fitted through the origin, the major axis (β_{MA}) and the standardised major axis (β_{SMA}) slopes and confidence intervals can be calculated by modifying existing formulas – setting all sample

means to 0, and replacing $N-1$ with $N-2$, in the formulas of Table 4. As previously, the exact method allows exact inference for normally distributed residuals, and close to exact inference for non-normal residuals (Table 9). This remained true even when the data were not actually centered at zero, as long as the true line passed through the origin.

(3) Confidence intervals for elevation

Simulations were conducted to measure the coverage probability of confidence intervals for elevation (α) in small samples, when calculated using the method described in this paper or the method of Legendre & Legendre (1998). Confidence intervals were considered for the major axis elevation and standardised major axis elevation.

Data were generated from the bivariate normal distribution with sample size 20, correlation 0.5, variances 1. The location of the centroid was varied over four points (which make up a square): $(0,0)$, $(0,10)$, $(10,10)$, $(10,0)$. This allowed consideration of the effects of shifts along the X-axis and Y-axis separately. The respective true elevation in the four simulations was $0, 10, 0, -10$.

Results in Table 10 illustrate that the confidence intervals for elevation proposed in this paper can work well. In contrast, the method proposed by Legendre & Legendre (1998) worked well for $\mu_x = 10$ but was particularly poor when $\mu_x = 0$. See the main body of the manuscript for an explanation of the behavior of the method due to Legendre & Legendre (1998).

(4) Type I error of tests for common slope

Some simulations were conducted to measure the Type I error of different tests for common slope. Simulations were conducted for two groups of data generated as bivariate normal with variances 1, total sample size 40, sampling either balanced or unbalanced, and correlation either the same for the two groups or different. These simulations compared three

test statistics:

- An F statistic, analogous to what is done in linear regression to compare several slopes. Results demonstrated that this statistic does not have an F distribution – the critical value at the 0.05 significance level was exceeded as much as 30% of the time (Table 11).

- The likelihood ratio test described in Warton & Weber (2002). Results here (Table 11) and elsewhere (Warton & Weber, 2002; Warton, in review) demonstrate that this statistic maintains close to exact significance levels at the 0.05 level under a broad range of conditions, for samples of size 20 and indeed smaller.

- The likelihood ratio test, but with common slope estimated by pooled sums of squares (as suggested by Krzanowski, 1984) instead of by maximum likelihood estimation as in Flury (1984) or Warton & Weber (2002). Type I error remained close to nominal levels for this method only when correlation was the same across groups. In other instances, Type I error was inflated (Table 11). The inflation arose because the common slope was not estimated using the maximum likelihood estimator, so the null likelihood function was underestimated and the test statistic was overestimated.

(5) Confidence intervals for common slope

Simulations have been conducted to measure the coverage probability of confidence intervals for the common slope of the major axis (β_{MA}) and the standardised major axis (β_{SMA}) in small samples. Details on how these confidence intervals are calculated can be found in Appendix D. The same set of simulation conditions were used as for Table 11.

Confidence intervals for a common SMA slope are close to exact for all simulations, and a reasonable approximation for a common MA slope (Table 12).

(6) Type I error of tests for common elevation

Simulations have been conducted to measure the Type I error of F and Wald statistics to test for common elevation. These tests are described in more detail in Appendix D.

Simulations were conducted for data generated as bivariate normal with variances 1, for 2 groups with a total sample size of 40. Simulations considered difference relative locations of the centroids of the two samples, different sampling designs, and different correlations. In all cases, the null hypothesis is true (the true elevations of both groups are equal). All the test statistics are invariant under any location change applied to all groups, so the centroid of the first group is fixed at the origin without loss of generality.

From simulation results it is clear that the Wald statistic should be used in practice (Table 13). This statistic can maintain close to nominal levels for the chi-squared distribution irrespective of differences between groups in sample size, residual variances, or means of the X variables. In contrast, the F statistic based on ANCOVA is sensitive to unequal X means, and to unequal residual variances in unbalanced designs.

Simulations including resampling of these test statistics have also been conducted (results not shown). If any of the above test statistics were resampled by bootstrapping residuals within each group (as described in Appendix F), then the resampled statistic maintains close to nominal significance levels. On the other hand, if residuals are permuted between groups, the F statistic can depart substantially from nominal levels if residual variances are not equal across groups, analogous to the situation described for analysis of variance by Boik (1987).

(7) Confidence intervals for methods-of-moments slope

Simulations were conducted to measure the coverage probability of confidence intervals for slopes of the methods-of-moments major axis ($\beta_{MM,MA}$) and the methods-of-moments

standardised major axis ($\beta_{MM,SMA}$). The methods of confidence interval construction considered were: ignoring measurement error and using exact methods (Pitman, 1939; Creasy, 1957; Jolicoeur & Mosimann, 1968), correcting variance terms for measurement error and using exact methods, using the t_{N-2} distribution but with variance estimates taken from Akritas & Bershady (1996). For the simulation results of Table 14, bivariate normal data were used, and measurement errors were also normally distributed. Sample sizes were 10, 30 or 90, with measurement error variances estimated from 2, 4 or 8 repeated measures of each observation. Measurement error variances were either 0.4 or 0.2.

Note that the repeated measures were averaged before line-fitting, which reduces the size of the measurement error variance as a factor of the number of repeated measures. For example, with $\sigma^2_{\delta_y} = 0.4$ and $n_{\text{rep}} = 4$, the variance of averaged repeated measures is $\frac{0.4}{4} = 0.1$. Because the variances of Y is one, this means that the variance of averaged repeated measures is 10% of the size of the variance of Y. However, the variance of average repeated measures can be large compared to the variance of residual scores (σ^2_{r}) – $\sigma^2_{\text{r}} = 1 - \rho$ when variances are one, which is 0.5 and 0.25 in simulations.

Measurement error was introduced into Y only, because if measurement error is not corrected for, the most extreme situation in which measurement error biases slope is when measurement error is in one variable only.

Results can be summarised as follows:

- Exact methods ignoring measurement error were poor at larger sample sizes and larger measurement errors, because of bias. However, this method often led to more accurate confidence intervals than the competing methods in small samples, and had reasonably accurate coverage probability when measurement error was not large.

- There was substantial undercoverage for methods that correct for measurement error whenever sample size and number of repeated measures were small ($n = 10$ and $n_{\text{rep}} = 2$). This suggests that to use these methods, measurement error variance

needs to be estimated reasonably well (*i.e.* from more than 30 observations in total, including repeated measures).

- All methods performed better when the measurement error was smaller compared to error from the line (*i.e.* compared to residual variance) – hence they performed better when $\sigma^2_{\delta_y}$ was smaller, n_{rep} larger, or ρ smaller.

- The exact method adjusted for measurement error usually had good coverage probabilities when variance of measurement error was less than 20% of the variance of residual scores. However, in other situations it performed poorly, and it was quite slow to converge to a coverage probability of 95% with increasing sample size.

- The only method that lead to consistently good coverage probabilities in moderate-large samples was the method due to Akritas & Bershady (1996). However, this method was usually too liberal for $N = 10$.

As a tentative rule, measurement error could be ignored if the variance of measurement error were less than 20% of the variance of residual scores. Under this rule, the "exact" confidence intervals ignoring measurement error performed reasonably well in simulations. However, the rule should be used with caution: the estimated slope without correcting for measurement error is biased (in our simulations, the bias was up to 10%), and confidence intervals ignoring measurement error are not consistent (*i.e.* in very large samples, the confidence intervals will not contain the true slope, due to bias).

If measurement error needs to be accounted for, the method due to Akritas & Bershady (1996) can be used if sample size is not small. Unfortunately, none of the methods considered here is reliable if sample size is small ($N \approx 10$), so the use of resampling methods is suggested in these instances.

F Resampling-based procedures

This appendix outlines the algorithms for resampling-based inference for the procedures described in the main text. The two techniques of resampling that will be considered are permutation testing under the reduced model (Freedman & Lane, 1983) and bootstrapping data to reflect H_0 (Hall & Wilson, 1991). The approach due to Freedman & Lane (1983) was originally proposed for the regression context, but it can be applied to residual and axis scores, since linear transformation to these variables essentially reduces MA or SMA to a regression problem. The approach due to Freedman & Lane (1983) has been found to maintain nominal significance levels more closely than other available permutation algorithms when residual variances are constant (Anderson & Robinson, 2001).

Bootstrapping and permutation testing arise from quite different models and philosophies (Westfall & Young, 1993, pages 169-177). Permutation tests arise from studies in which randomisation has been used in assigning treatments to subjects (*e.g.* in randomly choosing the mice to receive an injection of a hormone treatment), whereas bootstrapping arises from approximating the sampling distribution(s) of the population(s) being studied. Of these two, the latter is closer to the underlying model in allometry – there is usually no treatment applied by randomisation, and instead samples from different populations are being compared.

If resampling in allometry, we recommend bootstrapping in preference to permutation tests, although acknowledging that in practice the performance of the two methods will be almost identical. The only situation in which one might expect qualitatively different results is in multi-sample tests when the residual variances are unequal – in this situation, the bootstrapping algorithms are still applicable, but a different permutation testing algorithm is necessary for valid inferences.

(1) One-sample test of slope

The one-sample test that the slope equals some value b reduces to a test for correlation of the appropriate residual plot, $(bY + X, Y - bX)$ for the major axis, and $(Y + bX, Y - bX)$ for the standardised major axis. A resampling-based test would involve constructing this plot and resampling as appropriate. The following algorithm is proposed:

1. Construct the residual plot
2. Calculate the test statistic ($r_{\mathrm{rf}}(b)^2$, in the notation of Table 4)
3. Set *count* to 0 (or 1 for permutation testing).
4. For bootstrap testing, calculate residual and axis scores using the estimated slope $\hat{\beta}$: residual scores are $Y - \hat{\beta}X$, and axis scores are $\hat{\beta}Y + X$ for MA, $Y + \hat{\beta}X$ for SMA. For permutation testing, consider the residual and axis scores calculated using b as the slope instead of using $\hat{\beta}$. These will be used in step 5.
5. For *iter* steps (or $iter - 1$ for permutation testing), repeat the following:
   (a) Resample the residual scores (referred to as r*). For permutation testing under the reduced model, the residual scores are randomly reassigned to axis scores, and for the bootstrap, residual scores are resampled with replacement and assigned to axis scores.
   (b) Recalculate the test statistic: $r_{\mathrm{r}^*\mathrm{f}}(b)^2$ for permutation testing or $r_{\mathrm{r}^*\mathrm{f}}(\hat{\beta})^2$ for bootstrapping.
   (c) If $r^2_{\mathrm{r}^*\mathrm{f}} > r_{\mathrm{rf}}(b)^2$, add 1 to *count*.
6. Calculate the P-value as $\frac{count}{iter}$.

Note that for the bootstrapped datasets, the residual and axis scores are calculated using $\hat{\beta}$ rather than b. This is done to preserve the properties of the sample data in the resampled data – in particular, the variances and covariances of the residual and axis scores will resemble those of the sample data.

(2) Test for common slope

In the case of testing for a common slope, some alterations to the above algorithm are required. Obviously, the test statistic to be used will be different – the likelihood ratio test statistic given in Appendix D will be used in steps 2 and steps 5b-c. Alterations to the method of resampling are also required.

In this section, the notation of Appendix D will be used – so $\hat{\beta}$ now refers to the common slope, and $\hat{\alpha}_i$ refers to the sample elevation for the ith group, assuming common slope (unless otherwise specified).

For resampling (steps 4 and 5a), residuals are resampled rather than residual scores (although this is not essential if bootstrapping), and a back-transformation to the original axes is required. The fitted axis scores (F) and residuals (R) are:

$$(F, R) = \begin{cases} (\hat{\beta}Y + X, Y - \hat{\alpha}_i - \hat{\beta}X) & \text{for MA} \\ (Y + \hat{\beta}X, Y - \hat{\alpha}_i - \hat{\beta}X) & \text{for SMA} \end{cases}$$

which means that the back-transformation to the original variables, X and Y, is

$$(X, Y) = \begin{cases} \left(\frac{1}{1+\hat{\beta}^2}(F - \hat{\beta}R + \hat{\alpha}_i\hat{\beta}), \frac{1}{1+\hat{\beta}^2}(\hat{\beta}F + R - \hat{\alpha}_i)\right) & \text{for MA} \\ \left(\frac{1}{2\hat{\beta}}(F - R - \hat{\alpha}_i), \frac{1}{2}(F + R + \hat{\alpha}_i)\right) & \text{for SMA} \end{cases}$$

Because the test statistics are location and scale invariant, the back-transformations can be replaced by

$$(X', Y') = \begin{cases} (F - \hat{\beta}R, \hat{\beta}F + R) & \text{for MA} \\ (F - R, F + R) & \text{for SMA} \end{cases}$$

In the case of common slope testing by permuting residuals under the reduced model, steps 4 and 5a are:

4. Using the lines fitted with a common slope, construct the residuals, and axis scores.

5(a). Randomly reassign the residuals to axis scores. Back-transform to the original axes.

When testing for a common slope by bootstrapping, these steps become:

4. For the lines fitted with different slopes (*i.e.* using $\hat{\beta}_i$ and α_i calculated using data from the ith group only), construct the residuals and axis scores.

5(a). Resample residuals with replacement within each group, assign each to an axis score. Back-transform using the transformation based on the common slope estimator $\hat{\beta}$ (not using the $\hat{\beta}_i$).

When bootstrapping, the common slope $\hat{\beta}$ is used in back-transformation rather than $\hat{\beta}_i$ so that the resampled data reflect H_0 (Hall & Wilson, 1991), *i.e.* so that the resampled data are generated from distributions with a common slope. Also, resampling of residuals is done within each group rather than across groups to ensure that any differences between groups in variances of residuals are preserved.

(3) Test for common elevation

In the case of tests for common elevation, a resampling algorithm can be used that is similar to the one used for common slope testing, with only two differences:

- The test statistic is changed to the Wald statistic for equal elevation.
- In step 4, residuals need to be calculated using the estimated common elevation ($\hat{\alpha}$) rather than using the sample elevations for the different groups ($\hat{\alpha}_i$). This ensures that all groups of resampled data have common elevation.

(4) Test for no shift along the fitted axis

In the case of tests for no shift along the fitted axis, there are several changes to the resampling algorithm for common slope testing:

- The test statistic is changed to the Wald statistic for no shift along the fitted axis.

- In step 4, the axis scores must be subtracted by the common elevation of the (standardised) minor axis, *i.e.* fitted axis scores are replaced by $\hat{\beta}Y + X - \hat{\alpha}_{\mathrm{r}}$ for MA, and $Y + \hat{\beta}_i X - \hat{\alpha}_{\mathrm{r}}$ for SMA. This ensures that all groups of resampled data have a common location along the fitted axis.

- The fitted axis scores are resampled rather than the residuals.

- Residual scores may be used rather than residuals, for permutation testing and bootstrapping.

REFERENCES

Akritas, M. G. & Bershady, M. A. (1996). Linear regression for astronomical data with measurement errors and intrinsic scatter. *The Astrophysical Journal* 470, 706–714.

Allison, T. & Cicchetti, D. V. (1976). Sleep in mammals: ecological and constitutional correlates. *Science* 194, 732–734.

Anderson, M. J. & Robinson, J. (2001). Permutation tests for linear models. *Australian and New Zealand Journal of Statistics* 43, 75–88.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*, 2nd Edn. New York: John Wiley & Sons.

Bland, J. M. & Altman, D. G. (1986). Statistical method for assessing agreement between two methods of clinical measurement. *The Lancet* i, 307–310.

Boik, R. J. (1987). The Fisher-Pitman permutation test: a non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology* 40, 26–42.

Carroll, R. J. & Ruppert, D. (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *American Statistician* 50, 1–6.

Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995). *Measurement error in nonlinear models*. London: Chapman and Hall.

Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Belmont: Wadsworth.

Clarke, M. R. B. (1980). The reduced major axis of a bivariate sample. *Biometrika* 67, 441–446.

Cox, D. R. (1958). *Planning of experiments*. John Wiley and Sons.

Creasy, M. A. (1957). Confidence limits for the gradient in the linear functional relationship. *Journal of the Royal Statistical Society B* 18, 65–69.

Draper, N. R. & Smith, H. (1998). *Applied regression analysis*, 3rd Edn. New York: John Wiley & Sons.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* 3, 1–21.

Falster, D. S., Wright, I. J. & Warton, D. I. (2003). (S)MATR: Standardised major axis tests and routines. version 1.0. http://www.bio.mq.edu.au/ecology/SMATR/.

Feigelson, E. D. & Babu, G. J. (1992). Linear regression in astronomy. II. *The Astrophysical Journal* 397, 55–67.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist* 125, 1–15.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association* 79, 892–898.

Freedman, D. & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics* 1, 292–298.

Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley & Sons.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute* 15, 246–263.

Grafen, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London B* 326, 119–157.

Hall, P. & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* 47, 757–762.

Harvey, P. H. & Mace, G. M. (1982). Comparisons between taxa and adaptive trends: problems of methodology. In *Current problems in sociobiology* (ed. K. C. S. Group), pp. 343–361. Cambridge: Cambridge University Press.

Harvey, P. H. & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford: Oxford University Press.

Henery, M. L. & Westoby, M. (2001). Seed mass and seed nutrient content as predictors of seed output variation between species. *Oikos* 92, 479–490.

Isler, K., Barbour, A. & Martin, R. D. (2002). Line-fitting by rotation: A non-parametric method for bivariate allometric analysis. *Biometrical Journal* 44, 289–304.

Isobe, T., Feigelson, E. D., Akritas, M. G. & Babu, G. J. (1990). Linear regression in astronomy. I. *The Astrophysical Journal* 364, 104–113.

Jolicoeur, P. (1975). Linear regression in fishery research: some comments. *Journal of the Fisheries Research Board of Canada* 32, 1491–1494.

Jolicoeur, P. (1990). Bivariate allometry: interval estimation of the slope of the ordinary and standardized normal major axes and structural relationship. *Journal of Theoretical Biology* 144, 275–285.

Jolicoeur, P. & Mosimann, J. E. (1968). Intervalles de confiance pour la pente de l'axe majeur d'une distribution normale bidimensionnelle. *Biometrie-Praximetrie* 9, 121–140.

Jolliffe, I. T. (2002). *Principal components analysis*, 2nd Edn. New York: Springer-Verlag.

Kelly, C. & Price, T. D. (2004). Comparative methods based on species mean values. *Mathematical Biosciences* 187, 135–154.

Kendall, M. G. & Stuart, A. (1969). *The Advanced Theory of Statistics, volume 1.* London: Charles Griffin.

Kendall, M. G. & Stuart, A. (1973). *The Advanced Theory of Statistics, volume 2.* London: Charles Griffin.

Kermack, K. A. & Haldane, J. B. S. (1950). Organic correlation and allometry. *Biometrika* 37, 30–41.

Krzanowski, W. J. (1984). Principal component analysis in the presence of group structure. *Applied Statistics* 33, 164–168.

Lakshminarayanan, M. Y. & Gunst, R. F. (1984). Estimation of parameters in linear structural relationships: sensitivity to the choice of the ratio of error variances. *Biometrika* 71, 569–573.

Legendre, P. & Legendre, L. (1998). *Numerical ecology*, 2nd Edn. Amsterdam: Elsevier Science.

Lindley, D. V. (1947). Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society* 9, 218–244.

McArdle, B. H. (1988). The structural relationship – regression in biology. *Canadian Journal of Zoology* 66, 2329–2339.

McArdle, B. H. (2003). Lines, models, and errors: Regression in the field. *Limnology and Oceanography* 48, 1363–1366.

Miller, R. G., Jr (1986). *Beyond ANOVA, basics of applied statistics.* New York: John Wiley & Sons.

Moles, A. T. & Westoby, M. (2004). Seedling survival and seed size: a synthesis of the literature. *Journal of Ecology* 92, 372–383.

Moran, P. A. P. (1971). Estimating structural and functional relationships. *Journal of multivariate analysis* 1, 232–255.

Nair, K. R. & Shrivastava, M. P. (1942). On a simple method of curve fitting. *Sankhya* 6, 121–132.

Niklas, K. J. (1994). *Plant Allometry: The Scaling of Form and Process*. Chicago: University of Chicago Press.

Osada, N. (in preparation). Branching, biomass allocation, and crown development in a pioneer tree, *Rhus trichocarpa* .

Pitman, E. T. G. (1939). A note on normal correlation. *Biometrika* 31, 9–12.

Quinn, G. P. & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge: Cambridge University Press.

Rao, C. R. (1973). *Linear statistical inference and its applications*, 2nd Edn. New York: John Wiley & Sons.

Rayner, J. M. V. (1985). Linear relations in biomechanics: the statistics of scaling functions. *Journal of Zoology, Ser. A* 206, 415–439.

Reiss, M. J. (1989). *The allometry of growth and reproduction*. Cambridge: Cambridge University Press.

Ricker, W. E. (1973). Linear regressions in fishery research. *Journal of the Fisheries Research Board of Canada* 30, 409–434.

Ricker, W. E. (1982). Linear regression for naturally variable data. *Biometrics* 38, 859.

Riska, B. (1991). Regression models in evolutionary allometry. *American Naturalist* 138, 283–299.

Robertson, C. A. (1974). Large-sample theory for the linear structural relation. *Biometrika* 61, 353–359.

Schoenemann, P. T. (2004). Brain size scaling and body composition in mammals. *Brain, Behavior and Evolution* 63, 47–60.

Sokal, R. R. & Rohlf, F. J. (1995). *Biometry – the principles and practice of statistics in biological research*, 3rd Edn. New York: W. H. Freeman.

Sprent, P. (1969). *Models in regression and related topics*. London: Methuen.

Sprent, P. & Dolby, G. R. (1980). The geometric mean functional relationship. *Biometrics* 36, 547–550.

Staudte, R. G. & Sheather, S. J. (1990). *Robust estimation and testing*. New York: John Wiley & Sons.

Teissier, G. (1948). La relation d'allométrie: sa signification statistique et biologique. *Biometrics* 4, 14–53.

Warton, D. I. (in review). The robustness to failure of assumptions of tests for common slope of several allometric lines .

Warton, D. I. & Weber, N. C. (2002). Common slope tests for errors-in-variables models. *Biometrical Journal* 44, 161–174.

Westfall, P. H. & Young, S. S. (1993). *Resampling-based multiple testing – examples and methods for* P*-value adjustment*. New York: John Wiley & Sons.

Westoby, M., Leishman, M. R. & Lord, J. M. (1995). On misinterpreting the 'phylogenetic correction'. *Journal of Ecology* 83, 531–534.

Wilkinson, T. L. & Douglas, A. E. (1998). Host cell allometry and regulation of the symbiosis between pea aphids, *Acyrthosiphon pisum*, and bacteria, *Buchnera*. *Journal of Insect Physiology* 44, 629635.

Wright, I. J., Reich, P. B. & Westoby, M. (2001). Strategy shifts in leaf physiology, structure and nutrient content between species of high- and low-rainfall and high- and low-nutrient habitats. *Functional Ecology* 15, 423–434.

Wright, I. J., Reich, P. B. & Westoby, M. (2002). Convergence towards higher leaf mass per area in dry and nutrient-poor habitats has different consequences for leaf life span. *Journal of Ecology* 90, 534543.

Wright, I. J., Reich, P. B., Westoby, M., Ackerly, D. D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornelissen, J. H. C., Diemer, M., Flexas, J., Garnier, E., Groom, P. K., Gulias, J., Hikosaka, K., Lamont, B. B., Lee, T., Lee, W., Lusk, C., Midgley, J. J., Navas, M. L., Niinemets, U., Oleksyn, J., Osada, N., Poorter, H., Poot, P., Prior, L., Pyankov, V. I., Roumet, C., Thomas, S. C., Tjoelker, M. G., Veneklaas, E. J. & Villar, R. (2004). The worldwide leaf economics spectrum. *Nature* 428, 821–827.

Wright, I. J. & Westoby, M. (2002). Leaves at low versus high rainfall: coordination of structure, lifespan and physiology. *New Phytologist* 155, 403–416.

Table 1. Which method of line-fitting should be used when

Purpose	Key statistic	Appropriate method
Predict Y from X (X may even be random or may include measurement error)	$\hat{y}$	Linear regression
Test for an association between Y and X	P	Linear regression
Estimate the line best describing the bivariate scatter of Y and X	$\hat{\beta}$	MA or SMA
Test if the slope equals a specific value (1, or $\frac{3}{4}$, *etc*) for the line best describing the relationship between Y and X	$\hat{\beta}$	MA or SMA
Estimate the strength of the linear relationship between Y and X	R^2	Correlation
Predict Y from some underlying X that has been measured with error, so that only $(X + \delta)$ is observed	$\hat{y}$	Method-of-moments regression
Estimate the line best describing the bivariate scatter of Y and X, when only $(X + \delta_X)$ and $(Y + \delta_Y)$ are observed	$\hat{\beta}$	Method-of-moments MA or SMA

Table 2. Properties of MA and SMA that favour one or the other for line-fitting. The properties and recommendations listed here have a wide consensus or a strong logical basis. Some of the references given here relate to discussion of the equivalent question in the principal components literature – use of the covariance or correlation matrix for principal components analysis.

Property	Favours	Favoured in what situations	Explanation	Reference
Efficiency	SMA	All cases	SMA lines are estimated with greater precision (standard error of the slope is smaller).	Isobe *et al.* (1990); Jolicoeur (1990)
Scale dependence	SMA	When scale is arbitrary*	The major axis is scale dependent – if all Y values are doubled, the MA slope will not double.	Harvey & Pagel (1991); Sokal & Rohlf (1995); Jolliffe (2002)
Inference in complex problems	MA	When a method of inference for SMA is unavailable	For some complex problems, procedures for analysis are currently available for MA but not for SMA.	Anderson (1984); Jolliffe (2002)
Assumed error variances	MA	When there is no equation error and the measurement error variance is equal for X and Y	The major axis assumes the error variance is equal for X and Y, which is often a reasonable assumption when checking if two methods of measurement agree. Note that this argument does not hold if there is equation error (such as in allometry).	Sprent & Dolby (1980); Rayner (1985)

*Scale is arbitrary if the two variables are measured in qualitatively different units (*e.g.* kilograms and meters). Note that if both variables are log-transformed, units are no longer important and this consideration no longer applies, unless the power of X or Y is arbitrary (Is there a reason for plotting Y vs X rather than Y^2 vs X?)

Table 3. Controversial properties of line-fitting using MA and SMA that have been claimed to favour one or the other of MA and SMA line-fitting. We outline arguments against these recommendations, in the column labelled "But..."

Property	Favours	Claim	Reference	But...
Bias when error variance is misspecified	SMA	β_{SMA} is more robust to misspecification of error variances than β_{MA}	Lakshminarayanan & Gunst (1984); McArdle (1988)	There is no single correct specification of equation error in allometry, so there is no "true" error variance.
Assumed error variances	MA	For SMA, the assumptions made of error variances are unrealistic.	Jolicoeur (1975); Sprent & Dolby (1980)	Error variances can only be claimed to be "unrealistic" if they are due to measurement error. We do not recommend choice between MA or SMA on the grounds of measurement error.
Testing if X and Y are related	MA	β_{MA} can be used to test for a relationship between X and Y, but β_{SMA} cannot be.	Jolicoeur (1990); Legendre & Legendre (1998)	It is not essential that a single procedure be used both in testing for a relationship and in estimating the best-fitting relationship.
Permutation testing of the slope	MA	Permutation tests are not possible for β_{SMA}, because it is invariant under permutation of X or Y values.	Legendre & Legendre (1998)	Permutating X or Y is only appropriate for testing if X and Y are related. Permutation testing algorithms exist for both β_{MA} and β_{SMA}, as in Appendix F.
Exactness of primary CI	MA	If the secondary confidence interval is ignored, the CI for β_{SMA} is far from exact when correlation and sample size are small ($r^2 < 0.25$, $n = 10$).	Jolicoeur (1990)	This situation is of little practical interest. Typically, the sign of β_{SMA} is known *a priori*, in which case the CI is exact. If not, and if $n = 10$ and $r^2 < 0.25$ (which must be rare), the secondary CI should not be ignored.

Abbreviations – SMA=standardised major axis, MA=major axis, β_{SMA}= standardised major axis slope, β_{MA}=major axis slope, CI=confidence interval. Secondary CI: For MA and SMA slopes, there are two confidence intervals: one (usually) in the positive domain, and one (usually) in the negative domain. The secondary confidence interval is the one that does not contain the estimated slope.

Table 4. Calculation formulas for estimation of bivariate lines for linear regression, the major axis and standardised major axis, and for inference about the slope (β) or elevation (α) from one sample.

	Linear regression	Major axis	Standardised major axis
$\hat{\beta}$	$\frac{s_{xy}}{s_x^2}$	$\frac{1}{2s_{xy}}\left(s_y^2 - s_x^2 + \sqrt{(s_y^2 - s_x^2)^2 + 4s_{xy}}\right)$	$\text{sign}(s_{xy})\frac{s_y}{s_x}$
$\hat{\alpha}$	$\bar{y} - \hat{\beta}\bar{x}$	As for regression	As for regression
Residual axis (r)	$Y - \hat{\beta}X$	As for regression	As for regression
Fitted axis (f)	X	$\hat{\beta}Y + X$	$Y + \hat{\beta}X$
Test $H_0 : \beta = b$	$(N-2)\frac{r_{\text{rf}}^2(b)}{1-r_{\text{rf}}^2(b)} \sim F_{1,N-2}$	As for regression	As for regression
$s_{\hat{\beta}}^2$	$\frac{1}{N-2}\frac{s_y^2}{s_x^2}(1-r_{xy}^2)$	$\frac{(1+\hat{\beta}^2)^2}{N-2}\left(\frac{s_{\text{f}}^2}{s_{\text{r}}^2} + \frac{s_{\text{r}}^2}{s_{\text{f}}^2} - 2\right)^{-1}$	$\frac{1}{N-2}\frac{s_y^2}{s_x^2}(1-r_{xy}^2)$
$100(1-p)\%$ CI for β	$\hat{\beta} \pm s_{\hat{\beta}} t_{1-\frac{p}{2},N-2}$	$\frac{1}{2(s_{xy} \pm \sqrt{Q})}\left(s_y^2 - s_x^2 + \sqrt{(s_y^2 - s_x^2)^2 + 4s_{xy}^2 - 4Q}\right)$	$\hat{\beta}(\sqrt{B+1} \pm \sqrt{B})$, where
(primary)		where $Q = \frac{1}{N-2}(s_x^2 s_y^2 - s_{xy}^2) f_{1-p,1,N-2}$	$B = \frac{1-r_{xy}^2}{N-2} f_{1-p,1,N-2}$
Secondary CI for β	Not applicable	$\frac{1}{2(s_{xy} \pm \sqrt{Q})}\left(s_y^2 - s_x^2 - \sqrt{(s_y^2 - s_x^2)^2 + 4s_{xy}^2 - 4Q}\right)$	$-\hat{\beta}(\sqrt{B+1} \pm \sqrt{B})$
$s_{\hat{\alpha}}^2$	$\frac{s_{\text{r}}^2}{N} + \bar{x}^2 s_{\hat{\beta}}^2$	As for regression	As for regression
Test $H_0 : \alpha = a$	$\frac{\hat{\alpha}-a}{s_{\hat{\alpha}}} \overset{\text{approx}}{\sim} t_{N-2}$	As for regression	As for regression
$100(1-p)\%$ CI for α	$\hat{\alpha} \pm s_{\hat{\alpha}} t_{1-\frac{p}{2},N-2}$	As for regression	As for regression

Notation: we wish to estimate a line $Y = \alpha + \beta X$ from N pairs of observations of X and Y, $\bar{x}$ is the sample mean of the observations of X, s_x^2 is the sample variance, s_{xy} the sample covariance, r_{xy} the sample correlation coefficient. The variables "r" and "f" represent residual and fitted axis scores, respectively, and $r_{\text{rf}}(b)$ is the correlation between residual and axis scores, when these variables are calculated using a slope of b (not $\hat{\beta}$). The terms $t_{1-p,N-2}$ and $f_{1-p,1,N-2}$ represent the $100p\%$ critical values from the t_{N-2} and $F_{1,N-2}$ distributions, respectively.

Table 5. Assumptions of methods of inference for MA and SMA, how to check them, and when these assumptions matter.

Assumption	Is it satisfied?	Does this matter?	Reference
Residuals are independent	Sample randomly to guarantee this is satisfied	Yes! If there is dependence, standard errors/CI's are usually too small	Cox (1958, chapter 5)
Residuals are normally distributed	Check quantile plot of residuals	Not usually, although low power for long-tailed data	Miller (1986)
Y and X are linearly related	Check carefully a plot of residual vs axis scores for no pattern	Yes! Common slope tests are sensitive to non-linearity.	Miller (1986); Warton (in review)
Residuals have the same variance at all points along the fitted line	Check plot of residual vs axis scores for no pattern	Yes, but most methods are robust to moderate departures from equal variance	Warton (in review)

Note that "residual scores" and "axis scores" in this table refer to the variables given in Table 4, with a change of scale if desired. For testing for no mean shift along a fitted axis, all assumptions apply to fitted axis scores rather than to residuals.

Table 6. Terminology for methods of line-fitting. The first column contains terminology that is used in this manuscript, and is recommended for general use, and the second column contains terms that are equivalent to the proposed term, and have been used in the past

Preferred term	Other equivalent terms	Example reference
Linear regression	Model I regression	Sokal & Rohlf (1995)
Principal components	Principal components analysis	Jolliffe (2002)
	Model II regression	Sokal & Rohlf (1995)
	Errors-in-variables models	Fuller (1987, page 30)
	Structural or functional relationship	Lindley (1947)
Major axis	First principal component axis of the covariance matrix	Jolliffe (2002)
	Orthogonal regression	Isobe *et al.* (1990)
Standardised major axis	Reduced major axis	Kermack & Haldane (1950)
	Geometric mean functional relationship	Ricker (1973)

Table 7. Data on repeated measurements of LMA (in kg/m^2) for measurement error variance calculation. Data are for the 18 species sampled at West Head by Wright *et al.* (2001)

Species	n_i	$\overline{\text{LMA}}_i$	$s^2_{\text{LMA}_i}$	$s^2_{\delta_i}$
Acacia floribunda	8	0.10	0.00057	0.00125
Astrotricha floccosa	8	0.08	0.00015	0.00054
Allocasuarina sp	7	0.15	0.00014	0.00017
Correa reflexa	7	0.06	0.00009	0.00057
Dodonaea triquetra	7	0.10	0.00012	0.00033
Eucalyptus paniculata	7	0.13	0.00133	0.00225
Eucalyptus umbra	7	0.22	0.00417	0.00238
Lasiopetalum ferrugineum	6	0.11	0.00015	0.00039
Leptospermum polygalifolium	8	0.08	0.00018	0.00068
Lomatia silaifolia	6	0.13	0.00159	0.00302
Macrozamia communis	7	0.28	0.00025	0.00009
Persoonia linearis	6	0.16	0.00123	0.00145
Pomaderris ferruginea	8	0.08	0.00025	0.00085
Pultenaea daphnoides	9	0.10	0.00014	0.00029
Pultenaea flexilis	7	0.09	0.00022	0.00077
Synoum glandulosum	6	0.09	0.00037	0.00145
Syncarpia glomulifera	6	0.16	0.00022	0.00027
Xylomelum pyriforme	8	0.17	0.00066	0.00054

For repeated measurements of LMA for the ith species, n_i is the sample size, $\overline{\text{LMA}}_i$ is the sample mean, $s^2_{\text{LMA}_i}$ is the sample variance, $s^2_{\delta_i}$ is the estimated measurement error variance, calculated as described in the text. Note that "repeated measurements" in this context are measurements on different individuals.

Table 8. Simulations estimating the coverage probabilities (%) of 95% confidence intervals for the slope of the major axis (β_{MA}) and standardised major axis (β_{SMA}), when using the exact limits, a t_{N-2} approximation, or the variance estimates of Isobe *et al.* (1990), for data with different sample size (N), correlation (ρ), and from different distributions. Coverage probability was estimated from 10000 datasets.

		β_{MA}			β_{SMA}		
ρ	N	exact	t_{N-2}	Isobe	exact	t_{N-2}	Isobe
(a) Bivariate normal							
0.5	10	94.9	93.0	86.7	94.9	94.7	88.0
	30	95.3	94.3	92.1	95.3	95.4	92.6
	90	94.9	94.8	93.9	94.9	95.0	93.9
0.75	10	94.8	93.8	86.9	94.8	94.6	87.3
	30	95.2	95.1	92.2	95.2	95.2	92.1
	90	94.9	95.1	94.0	94.9	95.1	93.9
(b) 9:1 mixture, variances 1 and 4							
0.5	10	95.1	93.0	86.5	95.1	94.9	87.2
	30	95.1	93.9	91.7	95.1	95.1	91.5
	90	94.7	94.9	94.0	94.7	95.0	93.2
0.75	10	94.8	93.9	86.6	94.8	94.5	86.6
	30	94.7	94.6	92.0	94.7	94.7	91.5
	90	95.0	95.0	93.7	95.0	95.1	93.3
(c) 9:1 mixture, variances 1 and 9							
0.5	10	94.8	92.1	85.0	94.8	94.7	84.9
	30	94.8	93.7	91.9	94.8	94.8	90.4
	90	94.9	94.5	95.0	94.9	95.0	93.1
0.75	10	94.3	93.4	85.9	94.3	94.4	85.3
	30	94.8	94.4	91.9	94.8	94.7	90.7
	90	95.2	95.1	94.2	95.2	95.1	93.4

Table 9. Simulations estimating the coverage probabilities (%) of 95% confidence intervals for the slope of the major axis (β_{MA}) and standardised major axis (β_{SMA}) when fitted through the origin. Exact limits were used. Data were generated with different sample size (N), correlation (ρ), and centroid (μ_x, μ_y). Coverage probability was estimated from 10000 bivariate normal datasets.

			β_{MA}		β_{SMA}	
ρ	N	(μ_x, μ_y):	$(0,0)$	$(10,10)$	$(10,10)$	$(10,10)$
0.5	10		95.1	95.0	95.2	95.2
	30		95.0	94.7	95.0	94.7
	90		94.9	95.3	94.9	95.3
0.75	10		95.1	95.5	95.1	95.5
	30		95.2	95.5	95.2	95.5
	90		94.7	94.9	94.7	94.9

Table 10. Simulations estimating the coverage probabilities (%) of 95% confidence intervals for the elevation of the major axis and standardised major axis, using the t_{N-2} distribution and the method due to Legendre & Legendre (1998, pages 512-513, denoted "LL" below). Simulations varied the location of the centroid (μ_x, μ_y). Coverage probability was estimated from 10000 bivariate normal datasets.

	α_{MA}		α_{SMA}	
(μ_x, μ_y)	t_{N-2}	LL	t_{N-2}	LL
$(0,0)$	95.3	21.1	95.1	14.4
$(0,10)$	95.4	21.7	95.4	15.1
$(10,10)$	94.8	94.8	94.9	94.8
$(10,0)$	95.0	94.8	95.2	94.8

Table 11. Simulations estimating the Type I error (%) at the 0.05 significance level, of tests for common major axis slope (β_{MA}) and common standardised major axis slope (β_{SMA}). The test statistics are an F-test analogous to the linear regression case (F), and a maximum likelihood test where the common slope is either estimated using maximum likelihood ($\hat{\beta}$) or pooled sums of squares ($\tilde{\beta}$). Simulations varied the sample sizes of the two groups (n_1, n_2) and the correlation between Y and X for the two groups (ρ_1, ρ_2). Type I error was estimated from 10000 bivariate normal datasets.

		β_{MA}			β_{SMA}		
(n_1, n_2)	(ρ_1, ρ_2)	F	$\hat{\beta}$	$\tilde{\beta}$	F	$\hat{\beta}$	$\tilde{\beta}$
$(20, 20)$	$(0.75, 0.75)$	6.9	4.5	5.3	3.5	4.9	5.5
	$(0.6, 0.9)$	10.8	4.7	13.7	3.5	4.7	9.5
$(10, 30)$	$(0.75, 0.75)$	17.2	4.3	5.1	10.9	4.7	5.1
	$(0.6, 0.9)$	32.6	5.0	10.8	18.4	5.3	9.2
	$(0.9, 0.6)$	11.6	4.5	11.7	5.1	4.9	7.5

Table 12. Simulations estimating the coverage probability (%) of 95% confidence intervals for the common major axis slope (β_{MA}) and common standardised major axis slope (β_{SMA}). In simulations, bivariate normal data were generated, varying the sample sizes of the two groups (n_1, n_2) and the correlation between Y and X for the two groups (ρ_1, ρ_2). Coverage probability was estimated from 10000 bivariate normal datasets.

(n_1, n_2)	(ρ_1, ρ_2)	$\hat{\beta}_{\text{MA}}$	$\hat{\beta}_{\text{SMA}}$
$(20, 20)$	$(0.75, 0.75)$	94.8	95.2
	$(0.6, 0.9)$	94.7	95.0
$(10, 30)$	$(0.75, 0.75)$	94.0	94.5
	$(0.6, 0.9)$	95.0	95.2
	$(0.9, 0.6)$	93.3	94.6

Table 13. Simulations estimating the Type I error (at the 0.05 level) of F-tests and Wald tests (F and W, respectively) for common elevation. Simulations varied the sampling design (n_1, n_2), correlation (ρ_1, ρ_2), and the location of the centroid of the second sample $(\mu_{x,2}, \mu_{y,2})$. The location of the first centroid was always $(0, 0)$. Type I error was estimated 10000 bivariate normal datasets.

		α_{SMA}		α_{MA}	
ρ_1, ρ_2	(μ_{x2}, μ_{y2})	F	W	F	W
(a) $(n_1, n_2) = (20, 20)$					
$0.75, 0.75$	$(0, 0)$	5.5	5.3	5.4	5.5
	$(0.5, 0.5)$	6.5	5.3	5.9	5.3
	$(1, 1)$	10.7	5.6	8.0	5.2
$0.6, 0.9$	$(0, 0)$	5.3	5.3	5.3	5.5
	$(0.5, 0.5)$	5.7	5.2	5.5	5.3
	$(1, 1)$	8.5	5.8	7.3	5.8
(b) $(n_1, n_2) = (10, 30)$					
$0.75, 0.75$	$(0, 0)$	6.0	6.2	5.8	6.1
	$(0.5, 0.5)$	6.2	6.2	5.5	6.3
	$(1, 1)$	9.8	6.7	8.0	6.5
$0.6, 0.9$	$(0, 0)$	1.1	5.6	1.0	5.7
	$(0.5, 0.5)$	1.4	5.7	1.1	5.4
	$(1, 1)$	3.3	6.3	2.1	5.8
$0.9, 0.6$	$(0, 0)$	15.3	6.2	15.3	6.3
	$(0.5, 0.5)$	15.5	6.6	15.3	6.7
	$(1, 1)$	17.8	6.3	17.1	6.4

Table 14. Simulations estimating the coverage probabilities (%) of 95% confidence intervals for the slope of the methods-of-moments major axis ($\beta_{MM,MA}$) and methods-of-moments standardised major axis ($\beta_{MM,SMA}$), when there are repeated measurements of observations with measurement error (in Y only). Confidence intervals used the exact method ignoring measurement error (exact) or modified to account for measurement error ($\text{exact}_{\text{adj}}$), or used the method of Akritas & Bershady (1996, labelled AB). Simulations varied sample size (N), correlation (ρ), the variance of normally distributed measurement errors ($\sigma^2_{\delta_y}$), and the number of repeated measures (n_{rep}) of each subject. Repeated measures were averaged for each subject, and used to estimate measurement error variance. Coverage probability was estimated from 10000 bivariate normal datasets.

			$\beta_{MM,MA}$			$\beta_{MM,SMA}$		
ρ	N	n_{rep}	exact	$\text{exact}_{\text{adj}}$	AB	exact	$\text{exact}_{\text{adj}}$	AB
(a) $\sigma^2_{\delta_y} = 0.4$								
0.5	10	2	94.2	87.3	86.2	89.9	87.3	88.6
		4	94.8	91.6	85.9	89.1	91.6	87.5
		8	94.9	93.5	86.5	88.6	93.5	87.8
0.5	30	2	91.8	90.2	90.7	92.5	90.2	92.0
		4	93.8	92.8	91.2	92.7	92.8	91.5
		8	94.9	94.1	91.8	92.9	94.1	92.3
0.5	90	2	84.5	91.2	93.6	87.0	91.2	93.8
		4	92.2	92.9	93.5	92.8	92.9	93.4
		8	94.2	94.4	93.8	94.3	94.4	94.0
0.75	10	2	93.7	80.8	86.2	89.0	80.8	88.3
		4	94.5	88.8	87.0	88.7	88.8	87.7
		8	95.2	92.5	87.3	88.3	92.5	87.4
0.75	30	2	89.8	86.2	91.1	89.7	86.2	91.5
		4	93.4	91.8	92.0	92.3	91.8	92.1
		8	94.7	93.6	92.4	92.7	93.6	92.3

			$\beta_{MM,MA}$			$\beta_{MM,SMA}$		
ρ	N	n_{rep}	exact	$\text{exact}_{\text{adj}}$	AB	exact	$\text{exact}_{\text{adj}}$	AB
(a) $\sigma^2_{\delta_y} = 0.4$								
0.75	90	2	78.2	88.0	93.7	80.9	88.0	93.6
		4	90.3	91.8	94.0	90.8	91.8	94.0
		8	93.7	93.7	94.0	93.5	93.7	93.9
(b) $\sigma^2_{\delta_y} = 0.2$								
0.5	10	2	94.6	88.6	87.0	88.2	88.6	87.3
		4	95.0	92.3	87.3	88.3	92.3	87.5
		8	94.9	94.0	87.1	87.4	94.0	87.0
0.5	30	2	94.3	92.1	92.2	93.0	92.1	92.4
		4	94.6	93.3	92.1	92.5	93.3	92.2
		8	94.9	94.1	92.1	92.4	94.1	92.0
0.5	90	2	90.4	91.9	93.9	91.2	91.9	93.9
		4	93.5	93.2	93.9	93.4	93.2	94.0
		8	94.7	94.5	94.2	94.3	94.5	94.3
0.75	10	2	94.5	88.3	86.7	88.4	88.3	87.0
		4	94.8	92.3	87.4	88.7	92.3	87.4
		8	94.7	93.5	86.5	86.9	93.5	86.6
0.75	30	2	93.2	91.1	92.1	92.3	91.1	92.0
		4	94.5	93.1	91.6	92.5	93.1	91.8
		8	94.7	94.0	91.5	92.0	94.0	91.6
0.75	90	2	89.9	92.2	94.0	91.0	92.2	94.1
		4	93.7	93.4	93.8	93.3	93.4	93.8
		8	94.5	94.1	93.9	93.8	94.1	93.8

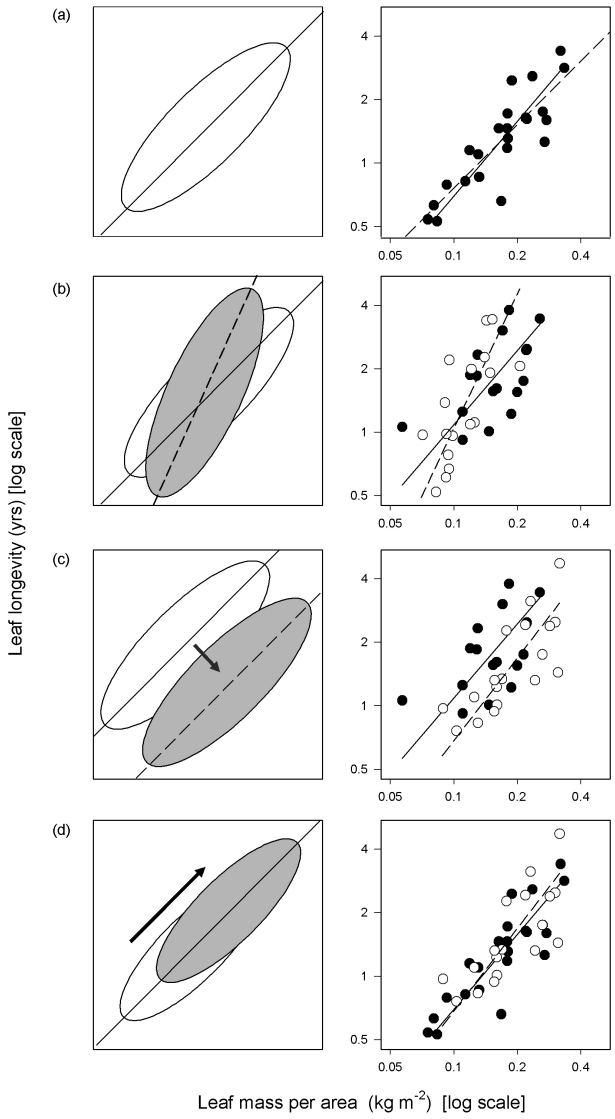

Fig. 1. An illustration of the four different types of tests considered in this paper: (a) testing if the slope equals a particular value (1 in this case, dotted line), (b) testing if slopes are different, (c) testing if elevations are different, (d) testing for shift along the axis. Data from Wright & Westoby (2002): leaf longevity (in years, log scale) vs leaf mass per area (kg/m^2, log scale), where each datapoint is for a different plant species. Species come in four natural groups, corresponding to higher vs lower rainfall and higher vs lower soil nutrient levels. Different pairs of groups have been plotted in (b-d), representing different rainfall or soil nutrient contrasts.

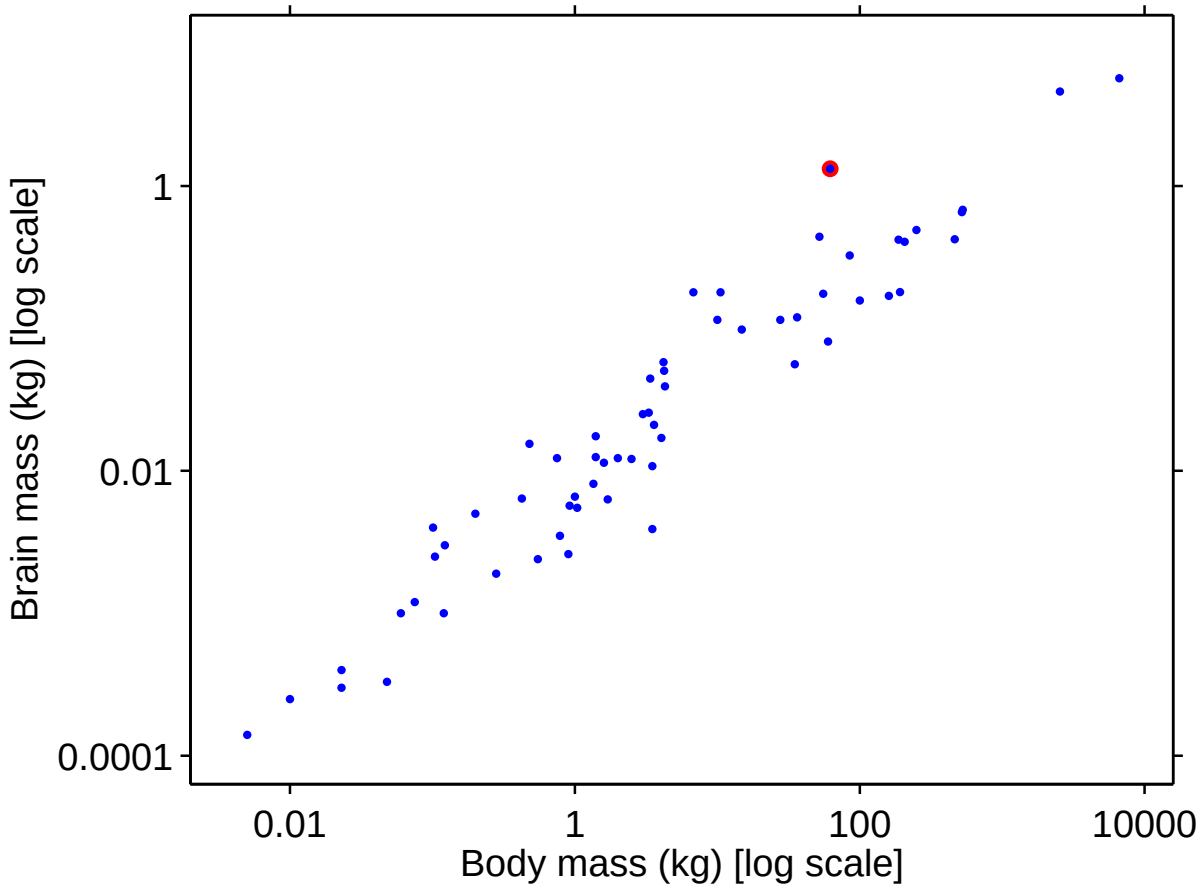

Fig. 2. A plot of brain mass against body mass for 62 mammal species. Humans are plotted using a larger symbol, and they have an unusually large brain considering their size (or is it a small body size considering their brain size...). Data from Allison & Cicchetti (1976).

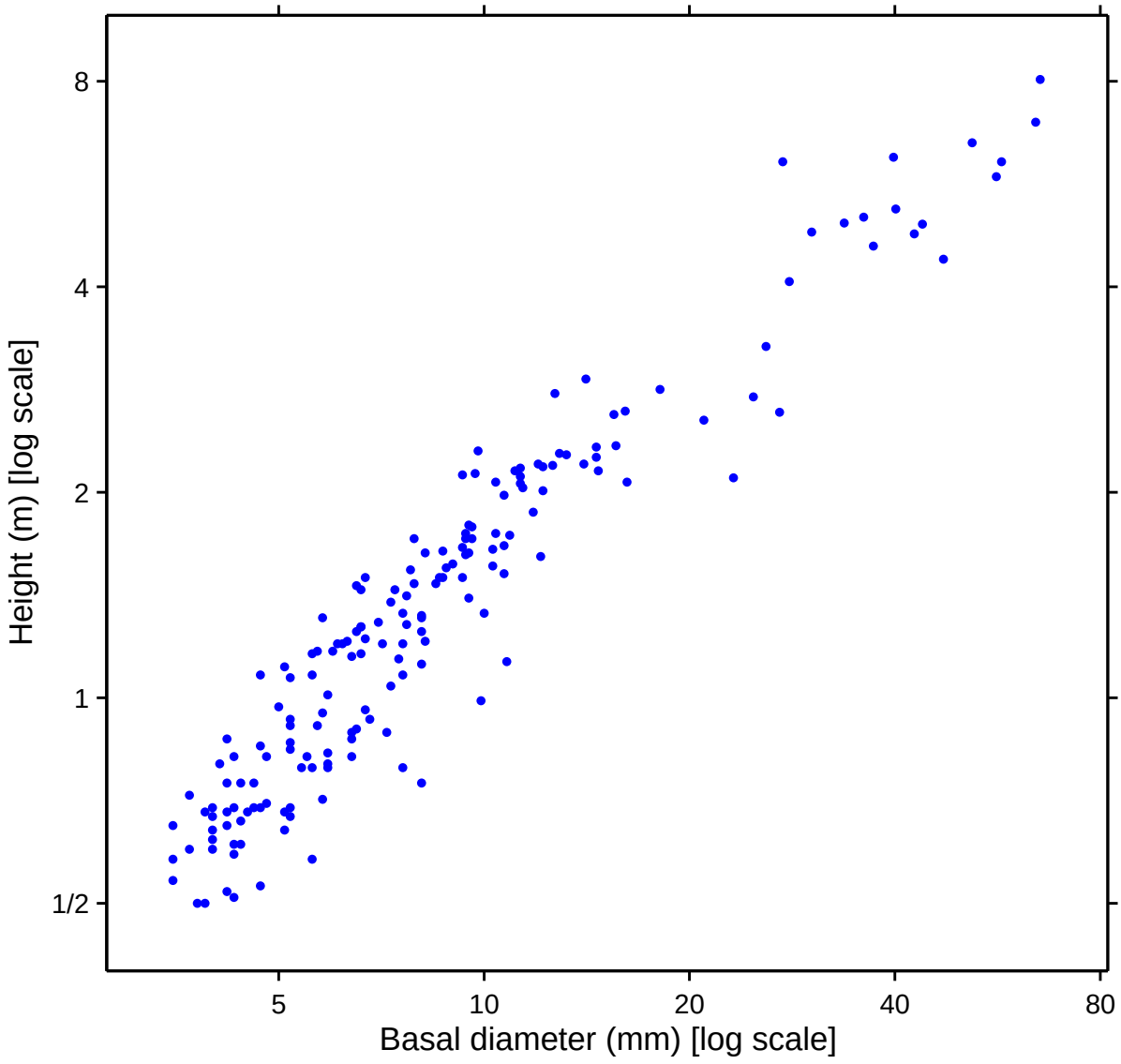

Fig. 3. A plot of height against basal diameter (measured at 10% height) for individual *Rhus trichocarpa* saplings. Data from Osada (in preparation).

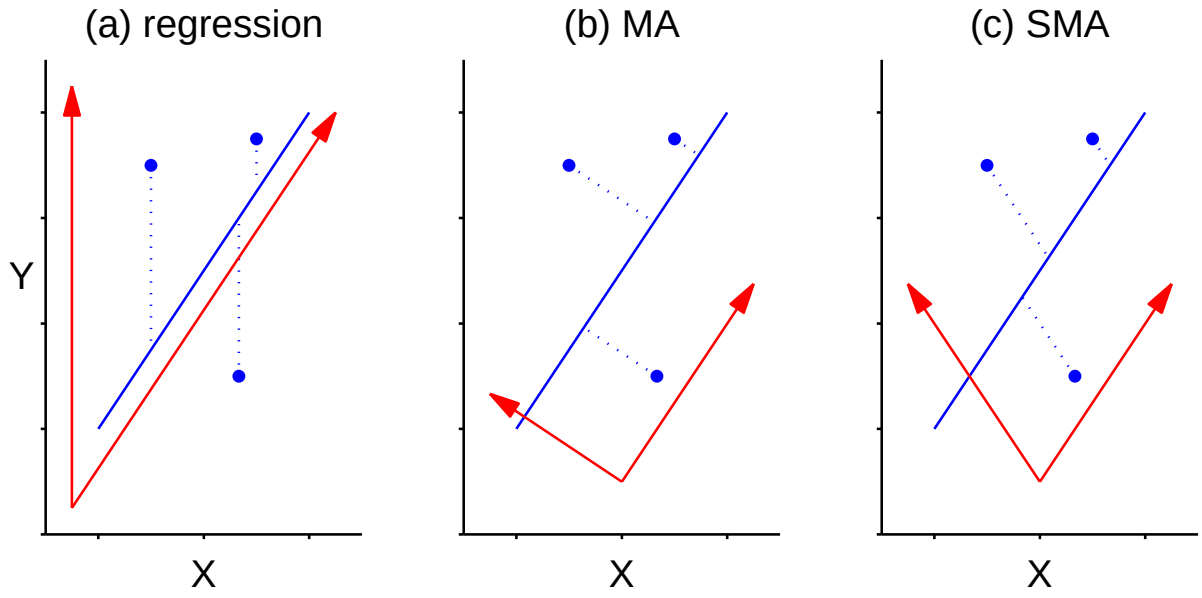

Fig. 4. The direction in which residuals are measured is (a) vertical for linear regression (b) perpendicular to the line for major axis estimation (c) the fitted line reflected about the Y axis for standardised major axis estimation. Axes are plotted on the same scale. The dotted lines indicate residuals, and the arrows represent the fitted and residual axes, which are useful for understanding methods of estimation and inference about these lines.

(a)

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

(b)

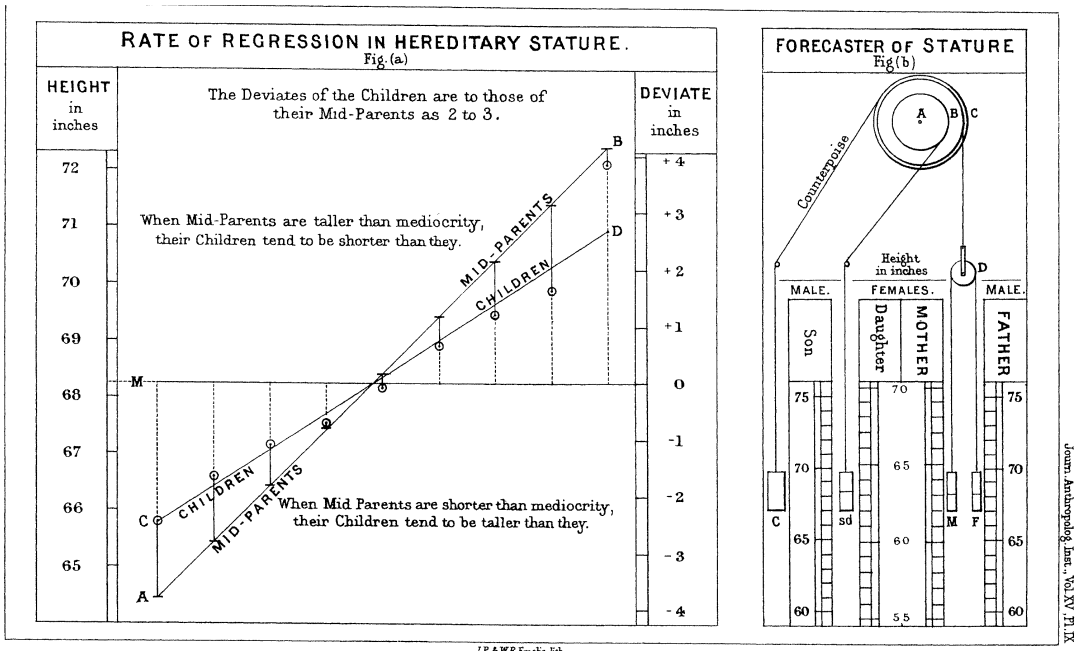

Fig. 5. Galton's height data for children and their parents, reproduced from Galton (1886). (a) Frequency table containing the data. (b) Schematic demonstrating regression to the mean, by comparing mid-parent height (line labelled "mid-parents") with predicted child's height (line labelled "children") for a regression based on the data in the table from (a). A child's predicted height is much closer to the mean than the mid-parent's height (child's height is about two thirds as far), hence the slope of the fitted regression is much flatter than expected from looking at the distributions of mid-parents and children separately.

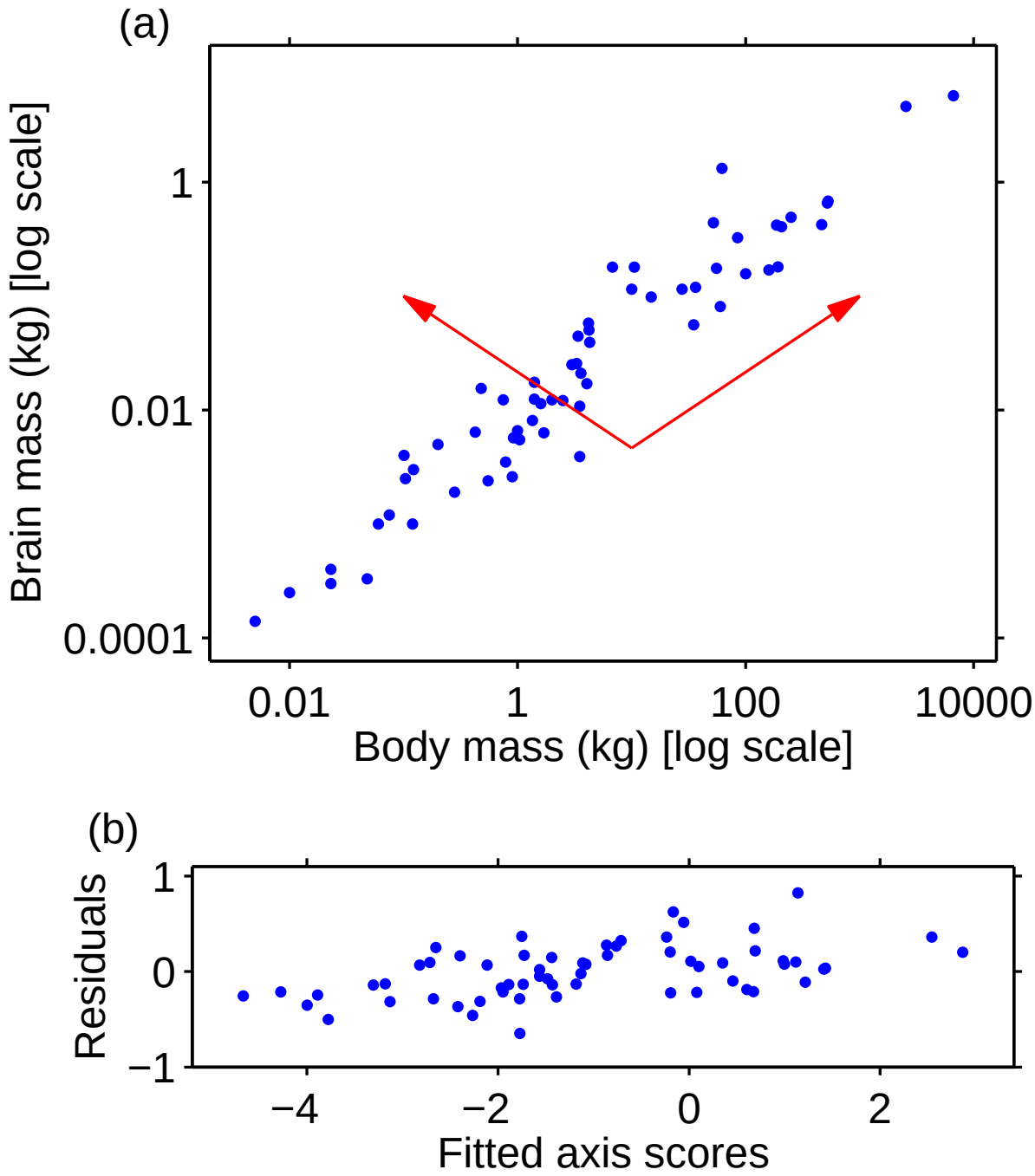

Fig. 6. A one-sample test of the SMA slope is a test for correlation between residual and fitted axis scores. (a) An example dataset (Allison & Cicchetti, 1976), with residual and fitted axes for SMA included, under the hypothesis that the SMA slope is 2/3. (b) Residuals plotted against fitted axis scores under this hypothesis. Note there is a trend in the residual plot for increasing residuals as axis scores increase – this is evidence against a SMA slope of 2/3.

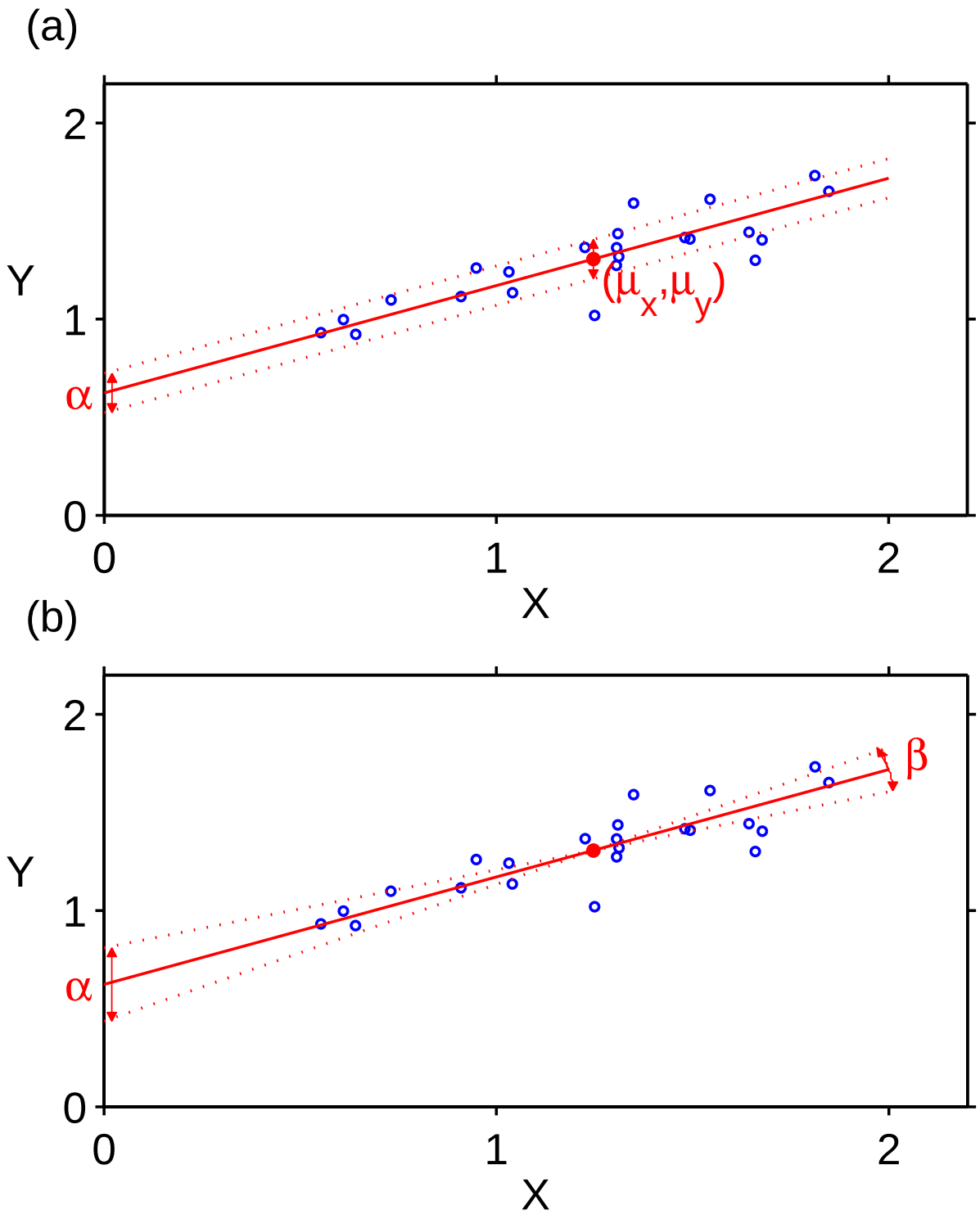

Fig. 7. Schematic diagram showing the two sources of error in estimating elevation, α. (a) Uncertainty estimating the centroid (μ_x, μ_y) affects elevation. (b) Uncertainty estimating the slope of the line β affects elevation (and has greater effect the further μ_x is from the Y-axis.)