J. theor. Biol. (1992) **156,** 405–423

The Uniqueness of the Phylogenetic Regression

ALAN GRAFEN

Plant Sciences Department, University of Oxford, South Parks Road, Oxford OX1 3RA, U.K.

(*Received on* 18 *March* 1990, *Accepted in revised form on* 10 *December* 1991)

The phylogenetic regression provides the hypothesis testing facilities of general linear models for comparative data with incompletely known phylogenies. It applies Ridley's radiation principle. It takes a thorough-going regression approach and so applies a Brownian motion model not to all the variables, but just to the error in the regression model. The phylogenetic regression is effectively unique as a regression method for comparative data—any substantially different method must be statistically unsound. The phylogenetic regression has been justified by analytical work and by computer simulations. Methods lacking explicit justification should be treated with suspicion. The statistical way forward for comparative biology lies in well-defined, well-justified methods.

Introduction

The phylogenetic regression (Grafen, 1989; for applications see Stone & Willmer, 1989; Boomsma & Grafen, 1990; Kirk, 1991; Promislow, 1991) is a recent advance in statistical technique that provides the hypothesis testing facilities of general linear models for comparative data in a way that overcomes the problems of non-independence that had bedevilled the comparative method. The advance was the transplanting of the principles underlying Ridley's (1983) method for discrete variables to a method for continuous variables, based on Felsenstein's (1985) method, but taking a thorough-going regression approach. Felsenstein's method is perfectly satisfactory when we know the true, probably binary, phylogeny of the species in the dataset. The focus of this paper is how to cope when the phylogony is imperfectly known. Felsenstein takes a correlational approach in which the variables themselves are assumed to be subject to Brownian motion. The new method assumes only that deviations from the regression line are subject to Brownian motion. It is very similar to Felsenstein's method when the phylogeny is known.

Ridley's central idea is that each radiation in the phylogeny should contribute one independent datapoint to the final test—I have called this the radiation principle. The difficulty is designing a method to do this, while at the same time using all the data, and providing all the hypothesis testing facilities of general linear models. For example, it is desirable to allow variables to be controlled for and to be able to test for the significance of categorical variables. The phylogenetic regression succeeds in all these ways.

There may seem to be many ways in which these goals could be achieved. In fact there are natural properties that any reasonable method must have which make the

0022-5193/92/120405+19 $03.00/0

method of the phylogenetic regression effectively unique as a way of applying the radiation principle. Harvey & Pagel (1991) review comparative methods, and present their own method and the phylogenetic regression as alternatives. My aim in this paper is first of all to explain why the phylogenetic regression is effectively unique, pointing out by way of illustration flaws in the method of Harvey & Pagel (1991), and then to explain further reasons why the phylogenetic regression should be preferred.

Methods that implement the radiation principle are conveniently explained as two transformations that are applied in turn to the dataset, and then as the performance of a statistical test on the final form of the data. First I will explain these transformations. In later sections the design choices to be made at each step will be pointed out. The dataset begins with one datapoint for each species, as shown in Table 1.

TABLE 1

The species dataset. There is one datapoint for each species. This is the form in which the data is collected or abstracted from the literature

Node	Y	Constant	$X1$	$X2$
1	−1·668	1	−1·9643	−0·72860
2	−2·680	1	−1·6285	0·56893
3	−2·394	1	−0·7114	0·55530
4	−4·090	1	0·1473	2·51710
5	−6·950	1	−2·4279	1·18156
6	5·131	1	2·2122	−1·26446
7	2·521	1	0·7464	0·02422
8	7·185	1	3·1639	−0·70283
9	−5·135	1	−1·0424	1·35052

The phylogeny is illustrated in Fig. 1. The same two transformations are applied to each variable in the dataset. The first transformation begins by working out the average value for a variable at each higher node in the phylogeny. Then each node is re-expressed as a difference from its parent node. This reconstitutes each variable as a series of phylogenetically arranged differences in a process I call "hanging on the tree". The root of the tree is the only node in the tree not to have a value. This new dataset can be written down as a sequence of datapoints, as illustrated in Table 2. The datapoints are in groups. Each group corresponds to a set of sister nodes. There are now more datapoints than in the original dataset, so we may call this the long dataset in contrast to the species dataset. (Any analysis of the long dataset must take into account the set of constraints that each variable adds to zero over the members of any group.)

Each group in the long dataset corresponds to a radiation in the phylogeny. To apply Ridley's radiation principle, we need to reduce the data in one group to a single datapoint, and so form what may be called the short dataset. This is done by taking linear contrasts. A linear contrast is a weighted sum of datapoints in which the weights themselves add up to zero. One linear contrast is taken across the datapoints in each group, and each linear contrast produces one datapoint for the short

TABLE 2

The long dataset. There is one datapoint for each node except the root (no. 14). The nodes are divided into groups depending on their parent node. For example, species 1 to 3 belong to the same group because their parent is node 10. Nodes 9, 12 and 13 belong to a different group because their parent is 14, so groups may not be adjacent in the list of datapoints. Notice that the sum of any variable within a group is zero. Notice also that the constant is now uniformly zero. Each node's average for the constant is 1, and so the deviation from its parent node, whose average is also 1, must be zero. The length column gives the length of the path segment leading down to that node in the phylogeny

Node	Parent	Length	*Y*	"Constant"	*X*1	*X*2
1	10	2	0·5793	0	−0·5296	−0·86048
2	10	2	−0·4327	0	−0·1938	0·43705
3	10	2	−0·1467	0	0·7233	0·42342
4	11	1	1·4300	0	1·2876	0·66777
5	11	1	−1·4300	0	−1·2876	−0·66777
6	12	2	0·1853	0	0·1714	−0·61677
7	12	2	−2·4247	0	−1·2944	0·67191
8	12	2	2·2393	0	1·1231	−0·05514
9	14	8	−3·8985	0	−0·9081	0·84593
10	13	2	1·4152	0	−0·1273	−0·74268
11	13	3	−1·8575	0	0·1671	0·97477
12	14	6	6·1822	0	2·1752	−1·15228
13	14	4	−2·4260	0	−1·1731	0·36997

dataset. There are therefore as many datapoints in the short regression as there are groups in the long dataset, which is the number of radiations in the phylogeny. The short dataset for the example is shown in Table 3. A conventional statistical test using general linear model methods is then performed on the short dataset. In a moment we turn to fill in the details of how these transformations and analyses must be carried out.

The word contrast is used ambiguously in statistics to refer to the vector of weights to be used in the weighted sum, and the result of the summation. Where there is danger of ambiguity, I shall refer to the contrast vector and the contrast result, respectively.

Recently, Lynch (1991) has published a very interesting method for analysing comparative data. It is based on the analogy of quantitative genetics, and attempts to distinguish between a kind of phylogenetically heritable component to traits, and an adaptive component. As Lynch points out, such a technique has potential applications in estimation and hypothesis testing, just as quantitative genetics does. For the purposes of the present paper, Lynch's method has no direct relevance. It takes account of recognized phylogeny (*sensu* Grafen, 1989), but does not attempt to take account of the fact that multiple nodes represent our ignorance about the

TABLE 3

The short dataset. There is one datapoint for each higher node. It is derived from the group of points in the long dataset corresponding to its daughter nodes. For example, the datapoint for node 14 *in the short dataset is derived from the datapoints for nodes* 9, 12 *and* 13 *in the long dataset. Once we arrive at the short dataset we can apply ordinary multiple regression and trust the significance levels we obtain. Notice, however, that the "constant" is uniformly zero in the short dataset, and so we must perform a multiple regression without a constant. The F-ratio testing the hypothesis that neither* X1 *nor* X2 *is related to* Y *comes to* 54·1 *on* 2 *and* 3 *df. This test has a total of* 5 *df. which is the number of higher nodes in the phylogeny, and the number of datapoints in the short dataset. (For this example, the fitting of* ρ *was suspended, so that the branch lengths given in Table* 2 *were taken as fixed and known. This is equivalent to setting* ρ = 1. *This makes it easier for readers to check the arithmetic)*

Node	*Y*	"Constant'	*X*1	*X*2
10	1·476	0	−0·8920	−2·032
11	5·720	0	5·1504	2·671
12	6·611	0	3·4396	−1·130
13	3·728	0	−0·3354	−1·956
14	8·343	0	2·8535	−1·576

exact order of probably binary splits among a group. In a phylogeny with multiple nodes, ignoring this unrecognized phylogeny can lead to grossly inflated Type I error rates (Grafen, 1989). Quantitative geneticists are committed by their technique to the "correlational" view according to which the characters themselves (or some transformation of them) must possess a multi-Normal distribution, and this would pose problems for the inclusion of categorical variables in the model. The "regression" approach has no difficulty with categorical *X* variables. It may be that further development will incorporate an allowance for unrecognized phylogeny into Lynch's principled technique.

As a historical note, I am grateful to Dr William Kirk for drawing to my attention a technique used in entomology that seems to be the first example of a fully valid statistical method in the face of all the problems of phylogeny, but to have escaped

the notice of Grafen (1989) and Harvey & Pagel (1991). It consists of calculating correlation coefficients within each of a number of genera, and then employing a sign test on the signs of those coefficients. This method was first employed by Wasserman & Mitter (1978), and then later by Kirk (1990). Although fully valid, this test does not employ all the information, and is restricted in controlling for third variables. However, it antedated by 11 years the next method I am aware of (Grafen, 1989) that is valid in the presence of recognized and unrecognized phylogeny (*sensu* Grafen, 1989) for continuous data.

Where the Harvey & Pagel Method is Described

Harvey & Pagel (1991) provide the first published description of their method, although they credit it to two earlier works. Pagel & Harvey (1989) gives a very brief outline, promising future elaboration. An unpublished and undated manuscript of Pagel is also mentioned.

The Pagel & Harvey (1989) paper makes various claims for the superiority of their method over the phylogenetic regression. These are not repeated by Harvey & Pagel (1991), though not explicitly disavowed. It is also noteworthy that the earlier work contains an offer to distribute a computer program implementing their method. Although not explicitly withdrawn, the later work makes no mention of this offer.

Both works contain errors in the description of the phylogenetic regression, and readers are cautioned against believing those accounts.

For the purposes of this paper I shall take the book by Harvey & Pagel (1991) as the primary source for the method. The method is described in print only there, the method has changed since the fragmentary description in Pagel & Harvey (1989; see Harvey & Pagel, 1991: 157), and the book is a more recent and elaborated expression of the author's views.

In view of the changeability of Harvey & Pagel's method, it will be important that some aspects criticized in this paper are fundamental parts of the structure of the method, and not surface details that can easily be amended. Any future versions of the method are likely to share those fundamental parts.

The Transition from Species Dataset to Long Dataset

We return to discuss the first of the three stages in the implementation of the comparative method. The transition from species to long datasets requires the calculation of a mean for each variable at each higher node in the phylogeny, and then the re-expression of each node's average as a deviation from its parent node. One important property of the long dataset is that each datapoint, although not independent of other datapoints in the same group, should be independent of datapoints in all other groups. In order to achieve this, we must be careful when forming the long dataset to use appropriate weights in the calculation of averages at higher nodes.

The difficulty can be seen by considering in Fig. 1 the genus (node 10) comprising the species at nodes 1, 2 and 3. The average value calculated for node 10 must be independent of the deviations of each of species 1, 2 and 3 from that mean. There is only one set of weights that achieves this, and they are the weights that give rise to the most efficient estimate of the mean at node 10. The most efficient estimate of the mean is the one with lowest variance, and the appropriate weights depend on the lengths of the path segments in the tree connecting node 10 with nodes 1, 2 and 3. When working out the most efficient weights for node 13, we need to use the lengths of the path segments between node 13 and nodes 10 and 11, along with the sampling variance of the means at nodes 10 and 11. The sampling variance of the mean for node 10, for example, can be calculated from the lengths of the path segments between 10 and nodes 1, 2 and 3.

Efficient weights should be used because efficiency is itself a good thing and because only the efficient weights ensure that two datapoints in different blocks in the long regression are independent. The phylogenetic regression uses these efficient weights in calculating the means for higher nodes (see Grafen, 1989: theorem 1), as does Felsenstein (1985).

Considering the independence of datapoints logically requires the specification of a covariance structure on the data. Following Felsenstein (1985), we adopt Edwards & Cavalli-Sforza's (1964) Brownian motion model in which the lengths of the path segments in the phylogeny represent variances. The error variance of one datapoint is the summed length of the path from the root of the tree to the species node in question. The covariance between species equals the shared path length in the paths from the root to the two species. Without some model of the error, no progress can be made with parametric methods. (See Felsenstein, 1988 for a discussion of non-parametric methods.)

There is one very important difference between Felsenstein's (1985) method and the phylogenetic regression that concerns how this covariance structure is interpreted. Felsenstein takes a correlational approach in which he assumes that the characters

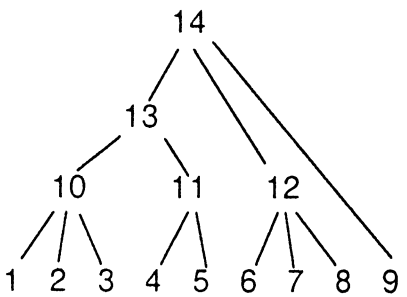

FIG. 1. This is the phylogeny used in the example dataset of Tables 1, 2 and 3. Note that like many phylogenies with which comparative biologists have to deal, there are multiple nodes. There is no neat set of levels in the phylogeny: species have from none to two nodes intervening between them and the root. The path segments are assigned lengths to represent the associated variance of error. For the example here, the lengths were assigned using the "Figure 2" method of Grafen (1989), and are shown in Table 2.

themselves evolve according to Brownian motion. This implies that the characters as measured at the species tips will have a multi-Normal distribution, with the variance–covariance matrix implied by the branch lengths. The phylogenetic regression takes a thorough-going regression approach, in which no assumptions are made about the X variables. It is only the error in the model which is required to be Normally distributed and which is assumed to have the variance structure represented by the branch lengths. As with correlation and regression in conventional statistics, this makes little difference to the methods themselves, but important differences to the interpretation of the assumptions. On the correlation view, it would be stretching the assumptions to test for a discrete X variable, because a variable that can take the values 0 and 1 cannot possibly be evolving by Brownian motion, which would imply that the values at the species tips are Normally distributed. On the regression view, however, it is perfectly satisfactory because it is only the deviations from the regression line that are attributed to Brownian motion. The thorough-going regression approach leads to a greater variety of tests, and allows the phylogenetic regression to offer all the hypothesis testing facilities of general linear models. General linear models include all fixed effects analyses of variance, analyses of covariance and multiple regression.

The Transition from Long Dataset to Short Dataset

The long dataset itself could be taken as the dataset for analysis. Unless the phylogony is binary, this would break Ridley's (1983) radiation principle that each higher node should contribute one datapoint. Simulations (Grafen, 1989: 135–43, in which this method is called the "standard regression") show that breaking the radiation principle really does matter, affecting significance levels in just the expected way. The Type I error rates for a nominal 5% p-value in four separate simulations were 6·3%, 10·7%, 7·9% and 28% based on 1000 runs each. Transforming from the long to the short dataset in order to reduce the number of datapoints to 1 per higher node is therefore a statistical necessity. The transition from long to short datasets is achieved by using linear contrasts. For example, the short datapoint for node 13 is formed by a linear contrast across the long datapoints for nodes 10 and 11. The short datapoint for node 10 is formed by a linear contrast across the long datapoints for nodes 1, 2 and 3.

The radiation principle asserts that one independent datapoint can be obtained from each radiation. The use of linear contrasts creates one datapoint for each radiation, but these will be independent only if each datapoint in the long dataset is independent of the datapoints in all groups except its own.

The radiation principle is silent on how to choose the linear contrasts, raising the spectre of many different tests with equal claims to be the implementation of the radiation principle. Most choices in the statistical development of regression-type methods turn out to be resolvable by appeal to simple properties that a reasonable method must have. In line with this expectation, by adding some commonsense requirements, it turns out that the linear contrast vectors are effectively unique. These requirements can be seen by concentrating on *annihilation*.

Annihilation is a technical term that just means setting to zero, simultaneously for all the values of one variable. A variable in the long dataset is transformed into a variable in the short dataset by a set of linear contrasts, as we have seen. The variable in the long dataset is said to be annihilated by the set of linear contrasts if the variable produced for the short dataset turns out to contain zero in every element. (An example of annihilation will arise later in the paper, when we will see that by definition a set of linear contrasts always annihilates the constant.) Any set of linear contrast vectors must annihilate some potential variables, and we can choose between different possible choices of linear contrasts by studying which variables they would annihilate. One important feature of annihilation is that an annihilated variable cannot be tested for. In an ordinary bivariate regression an X variable that has all its values equal is annihilated, and it is clearly impossible to estimate a slope if there is no variability in X.

It may at first sight seem an undesirable property of a comparative method that any variable should ever be annihilated. An analogy with the comparative method used by Read (1987) will show this is not so. This method calculates the slope between two variables within each of a number of genera. Then the population of slopes is considered as a dataset in its own right, and a one sample t-test is performed to test the hypothesis that the mean slope is zero. This test will "refuse" to work if each of the slopes is exactly zero. Why is this? Because a dataset in which all the slopes are exactly zero is "infinitely close" to datasets that should show an overwhelmingly significant positive association, and is also "infinitely close" to datasets that should show an overwhelmingly significant negative association, and also to datasets that should show an exactly zero association between the two variables. There is therefore no option but for a sensible statistical method to find some way of "refusing" to give an answer. (The neighbouring "significant" datasets have a zero variance but a non-zero mean, while the neighbouring datasets that show zero association have a zero mean but a non-zero variance. By making the non-zero value smaller we can make these datasets as close as we like to the dataset in which mean and variance are both zero.) The one sample t-test shows its refusal when the mean has to be divided by the standard error to calculate a t-value. Zero over zero is undefined. Returning to the phylogenetic regression, the refusal is shown when one of the X variables turns out to contain exactly zero in all its elements in the short dataset.

Annihilation of variables in the move from the long to short dataset in a comparative method should arise in the same kind of circumstances as in the method used by Read, namely when there are infinitely close datasets that should have completely different outcomes. Analogy with the method used by Read tells us where that happens. The groups of datapoints in the long dataset are analogous to Read's genera. When the association between Y and the X variable is exactly zero within each group, then the X variable should be annihilated. In that case, it is possible to construct infinitely close datasets that should show a highly significant association (positive or negative by choice), and other infinitely close datasets that should show a zero association.

Our commonsense requirement is therefore that an X variable should be annihilated when it is uncorrelated with Y in each radiation. This requirement tells us that

the linear contrast vectors must be obtaind from Y. Specifically, the linear contrast for a group must be (proportional to) the values of Y in that group in the long dataset. This tells us which linear contrasts to use in moving from the long to the short regression.

The importance of the annihilation requirement does not lie just in datasets in which annihilation occurs. The set of annihilated variables is like a fence in a field. It forms part of the "dividing line" between variables with a positive association with Y and those with a negative association. The wrong positioning of the set of annihilated variables implies an incorrect assignment of variables to those categories. The wrong positioning of a fence is not so important for the land actually under the fence, as for all the land in between the correct and the incorrect positions.

The same commonsense requirement can be developed in another way, deriving from the fact that if one of the linear contrast vectors is exactly zero or undefined, then the corresponding higher node must be dropped from the short dataset. I shall not pursue this here, but simply remark that appropriate dropping of short datapoints also requires that the linear contrasts be based on Y, and not on one of the X variables.

The discussion so far has assumed that all the X variables are test variables. If some X variables are being controlled for in the analysis, then the part of Y in the above discussion is taken by the residuals in the regression of Y on the control variables, rather than by Y itself. The rationale is simple. A test variable should be annihilated if, within each group of points in the long dataset, its correlation is zero with the so far unexplained component of Y. (The regression of Y on the control variables must be performed to obtain these residuals, and it is done using the long dataset, taking account of its non-diagonal variance–covariance matrix.) In fact, Y is always regressed at least on the constant before being used to form the linear contrasts.

The linear contrast vectors, then, are effectively uniquely determined by requiring appropriate annihilation. They must be based on the residuals in the regression of Y on the control variables. The phylogenetic regression (Grafen, 1989) uses these linear contrasts. The method of Harvey & Pagel uses linear contrasts based instead on the X variable to be tested for assocation with Y (Harvey & Pagel, 1991: 157). The contrast is not actually X itself. They assume that only one X variable is to be tested at a time. The X variable is never annihilated (because their contrast result is effectively the mean of X values above the mean of X minus the mean of X values below the mean of X), and the other X variables, those being controlled for, are annihilated in the wrong circumstances. Another X variable is annihilated when within each radiation it has the same mean when the special X variable is higher than average as it does when the special X variable is lower than average.

Harvey & Pagel's method as they describe it is restricted to test for only one X variable at a time. This excludes various kinds of possible tests, including testing for a categorical variable where there are more than two categories. The phylogenetic regression allows any number of X variables to be tested for simultaneously, and the unsatisfactoriness of basing the linear contrasts on an X variable can be seen in this more general context. With more than one X variable to be tested, an arbitrary

choice would need to be made of *which* X variable is to provide the linear contrasts. There is no satisfactory rule for choosing such a special X variable. A natural "linearity requirement" in general linear models is that the joint significance of a pair of X variables should be the same as the significance of any two linearly-independent linear combinations of the pair (e.g. the sum and difference of the X variables). This requirement cannot be satisfied if one of the test variables is used to form the linear contrasts. A parallel ambiguity arises if there are two control variables, one of which is to be used as the special X variable.

Use of Y provides naturally unique linear contrasts that do satisfy this "linearity requirement", and it is in the nature of the structure of the general linear model that only Y can do this. The use of Y to form linear contrast vectors raises certain technical statistical problems, because the linear contrasts then depend on the error in the model. The usual formulae for the properties of linear contrasts assume that the contrasts are independent of the error. Use of Y therefore requires a very careful justification in technical statistical terms, which it receives in Grafen (1989: section 10, particularly theorem 2).

Two other approaches lead to the same method. Indeed it was these two approaches, and their common result, that led me to develop the phylogenetic regression. The interpretation in terms of linear contrasts is a later view, convenient for computation and analysis, but logically less appealing. The first approach is to add dummy variables to a regression performed on the long dataset, to ensure that the residuals after adding the test variables are proportional, within each radiation separately, to the residuals before adding the test variables (Grafen, 1989: theorem 3). The second approach is a randomization test that involves fixing the relative values of the residuals within each radiation in a regression on the long dataset (theorem 4). The natural method is likely to have various interpretations and to be derivable from various principled approaches.

In conclusion, the linear contrasts used to move from the long to the short dataset are uniquely determined by commonsense requirements. This can be seen technically in terms of annihilation, or more informally by considering that if there is a right set of linear contrasts, it must be unique.

Performing the Test in the Short Dataset

The test in the short dataset can be performed as an ordinary unweighted general linear model, with three provisos. The first is that the linear contrast vectors have been standardized so that each short datapoint has the same error variance. The second is that the correct weighting was used in forming the averages of higher taxa, because otherwise the short datapoints are non-independent of each other. The third is that the regression must be performed without a constant term.

The omission of the constant term follows immediately from the use of linear contrasts to form the short dataset. Suppose we have a regression $Y = I\alpha + X\beta + \varepsilon$, in which I stands for the vector each of whose element is 1, and represents the constant term in the model. A set of linear contrast vectors can be written as the

rows of a matrix **L**, and the regression obtained from these linear contrasts is found by pre-multiplying the original regression equation by **L**. This yields

$$\mathbf{L}Y=\mathbf{L}\mathit{1}\alpha+\mathbf{L}X\beta+\mathbf{L}\varepsilon.$$

If we represent the terms in this new regression by stars, i.e. as a regression of Y^* ($=\mathbf{L}Y$) on X^* ($=\mathbf{L}X$) with error ε^* ($=\mathbf{L}\varepsilon$), we have the following

$$Y^*=X^*\beta+\varepsilon^*.$$

The constant term is missing from this regression because by definition of linear contrasts, $\mathbf{L}\mathit{1}$ equals zero (an example of annihilation), and so α does not enter into the new regression. This analysis results in unbiased estimates of β with the usual optimal properties associated with least squares methods (subject to using only the information present in the contrast results). The re-introduction of a constant is a statistical nonsense, and it is considered further here only because Harvey & Pagel (1991) recommend including the constant, albeit with some qualifications to which I will return. The re-introduction destroys some of the optimal properties of the least squares method. The estimate of β is still unbiassed, but its standard error is misestimated, which in turn vitiates statistical tests on the value of β. It is important to notice the implication that the inclusion of the constant is not a matter of taste or preference. Tests that re-introduce the constant, such as Harvey & Pagel's method, are invalid.

The relationship between the correlation coefficient based on the correct exclusion of the constant (r) and that based on its erroneous inclusion (r_{incl}) is

$$r_{\text{incl}}=\frac{r-\lambda_X\lambda_Y}{\sqrt{(1-\lambda_X^2)(1-\lambda_Y^2)}}$$

where $\lambda_X=\sqrt{m_X^2/(m_X^2+s_X^2)}$, m_X is the mean of X and $s_X^2=(1/n)\sum(X-m_X)^2$ with parallel definitions for λ_Y. λ is always between zero and 1. It is high when the variation is low relative to the mean, and low when the variation is high relative to the mean. If CV represents the coefficient of variation, but using $1/n$ instead of $1/(n-1)$ in the calculation of the sample variance, then $\lambda=1/\sqrt{1+CV^2}$.

r and r_{incl} are directly related to the F-ratios the methods would give in testing for an association. The formula reveals the relationship between the two methods. They give virtually the same answer when both λs are small, and otherwise only when r, λ_X and λ_Y satisfy a condition that is unlikely to occur by chance. It follows that we should expect the inclusion and exclusion of the constant to matter except when for both X and Y, the coefficient of variation of their values in the short regression is very high. This relationship between r_{incl} and r does not of course allow a correction for the other errors in Harvey & Pagel's method.

One way to see that the regression on the short dataset should pass through the origin uses the fact that the statistical properties of a linear contrast should be unaltered if its sign is switched, so that positive elements become negative and vice versa. This would have the effect of reversing the sign of X and Y for one datapoint only, resulting in that datapoint being reflected in the origin. The statistical test

applied to the short dataset should therefore give the same answer when any number of its datapoints are reflected in the origin. This is true when the constant is omitted, and certainly not true when the constant is included in the regression.

Harvey & Pagel present various graphs (in Harvey & Pagel, 1991: fig. 5.18) showing how the inclusion of a constant in the short regression can produce extremely misleading results. They suggest that the results of their method can be checked by a binomial test on the signs of the contrast results, or by performing a regression through the origin. The binomial test loses power, and in any event a method that needs checking in this way surely has something wrong with it. Where is the guarantee that the binomial test will pick up all the cases in which including the constant in the short regression is misleading?

They also advise that when the constant is omitted, "the slope of the line should not be interpreted". As the theory above shows, the slopes in both cases are unbiassed estimates of the true slope, but only without the constant is the standard error validly estimated. So in complete contradiction of Harvey & Pagel's advice, confidence intervals on the slope should be interpreted only if the constant is omitted. For real examples in which the omission of a constant has serious effects on an analysis, see Packer *et al.* (1992).

Finding a Set of Branch Lengths for the Tree

So far in this paper I have considered how comparative methods operate on a fixed working phylogeny, and have argued that the phylogenetic regression is effectively unique as the implementation of the radiation principle. Any method in the same general class, such as Harvey & Pagel's method, is simply in error when it differs from the phylogenetic regression in any of the areas so far discussed. The question of branch lengths is more complex.

The assignment of branch lengths to the working phylogeny affects how the weighted averages are computed, and how the contrast vectors are standardized. It is analogous to assigning weights in a weighted regression, except that the main purpose is not to represent how species differ in their trustworthiness, but to represent to what extent each pair of species tells the same story because of phylogenetic affinity. Some patterns of branch lengths will correspond to cases in which species are virtually independent, and other patterns to cases in which all the important variation lies at high taxonomic levels.

The task of assigning branch lengths is to give each path segment in the phylogenetic tree (e.g. Fig. 1) a positive real number which represents the expected (squared) amount of change associated with that path segment. The most general scheme that leaves flexibility to the user is therefore to allow a completely free choice of the positive real number for each path segment. This is what the phylogenetic regression does. In many cases, it will be convenient to assign branch lengths in a simpler way, when special information is not available for each path segment. The phylogenetic regression allows two such "default methods". One of them assigns a "height" to each node equal to the number of species below that node, minus 1. Then the branch length of a path segment is found by subtracting the height of the

lower node on the segment from the height of the higher node. This is a crude way of assigning branch lengths that will (i) work for any phylogeny (ii) always give a positive branch length for each path segment and (iii) always have the same total length from the root to each of the species. It is for these reasons a useful default, and is known as the "figure 2 method" after the figure in Grafen (1989) in which it is illustrated.

There are more branch lengths than there are species, so it is neither desirable nor possible to use the data in the analysis to estimate all the branch lengths. Most of the information must be simply assumed, when possible on the basis of other types of evidence. But we will often be able to afford one dimension of fitting, and there is one aspect of the tree in which flexibility is particularly worthwhile. This is the extent to which species are more or less independent, and the force of phylogeny is weak; or the species vary little within genera, genera vary little within families, and the force of phylogeny is strong. The regression view of the error (see above and Grafen, 1989: 143–5) is that it is only deviations from the regression line whose distribution need be considered. In particular, it is not directly relevant what pattern is shown by any of the variables in the analysis. As the deviations need not show the same tendency as any of the variables, and as they cannot be observed until after the analysis, it will rarely be sensible to assume in advance that we know the strength of phylogeny.

The phylogenetic regression includes a continuously varying parameter ρ ("rho") which is used to distort the initial branch lengths to create a family of possible phylogenetic trees that vary in the strength of phylogeny. $\rho = 1$ reproduces the initial specification. $\rho < 1$ increases the branch lengths near the species tips and decreases the branch lengths near the root, thus reducing the strength of phylogeny. $\rho > 1$ increases the branch lengths near the root and decreases the branch lengths near the species tips, implying that sister species will be very similar and that higher taxa will tend to be quite distinct. The phylogenetic regression then chooses a value of ρ by maximum likelihood, simultaneously with the regression parameters, in a regression of Y on the control variables using the long dataset. This value of ρ is then used to fix the phylogeny, and proceed to the short dataset and the final test.

The phylogenetic regression thus has a well-defined procedure for estimating the strength of phylogeny and taking it into account in the analysis. Harvey & Pagel's method is to assume initially that each branch length has equal length, and then to use inspection of the contrast results and residuals to take corrective action if necessary: "The only real difference in doing it at this stage is that the weighting is conditioned on patterns in the data, rather than in response to an assumed model" (Harvey & Pagel, 1991: 152). This is however quite an important difference.

First of all, the *ad hoc* nature of the adjustments would make the process virtually unrepeatable even by the same investigator at a later time, introducing a dangerous subjectivity into the statistical method. Of course, in any application of a statistical method there will be some parts that are well defined, and other parts that are not. Even in a simple experiment leading to a one-way analysis of variance, questions will arise about transformations, or the omission of outliers, or how exactly to measure the responses to treatment. But the inevitable questions of judgements are made very

difficult if the method of analysis itself is ill-defined. An ill-defined method pushes back the boundaries of judgment in an application to include varying the method itself. This requires that the user understand a great deal about the statistical justification of the method, and is therefore unsuitable in a method publicly promoted for widespread use.

The second important point is that informal data-dependent *ad hoc* adjustments by the user can in principle have serious effects on significance levels. Without reasonable understanding of the principles, a user might easily embark on a program of adjustment that had serious effects on the statistical properties of the test. Whether the method's significance levels are valid or not depends in principle on generating data according to a null hypothesis, and observing on what fraction of occasions that null hypothesis is rejected. So the method needs to be well defined in order to establish whether it provides a valid test or not. The validity of Harvey & Pagel's method is therefore not only unknown because the necessary work has not been done, but is actually unknowable to the extent that the method is ill-defined.

To summarize, the assignment of branch lengths is the one area in which the phylogenetic regression's approach is not unique, but only one reasonable method among many. Those many do not include Harvey & Pagel's (1991) method.

Which Method is to be Preferred?

In this section I give reasons why the phylogenetic regression is to be preferred over Harvey & Pagel's method for the analysis of comparative datasets. First of all there are the technical errors described above: the way in which the contrast vectors are formed, and the inclusion of a constant in the final analysis. Second there is the subjectivity of their method of handling branch lengths. Two more reasons will be given, Harvey & Pagel's "phylogenetic hypothesis" interpretation of contrasts will be considered, and a comparison of how the two methods have been justified will be made.

The third reason is the disjointed nature of their method. Harvey & Pagel recommend constructing the contrast results, and then choosing what to do with them. This separation is unsatisfactory, because the method of construction determines the contrasts' statistical properties and therefore the validity of the tests then performed. Two examples illustrate this point. Harvey & Pagel (1991: 152) suggested that if what they call "significant heterogeneity of variance" is found in the short regression, then "some form of weighted regression is required". To all appearances they are suggesting performing a weighted regression on the same set of contrast results. But in the circumstances in which weights are required in the short regression, the weights used to find the averages that are used to construct the contrasts in the first place should also be changed. To use an analogy, changing the weights half way through is like carrying out the calculations of a multiple regression by hand, starting off by employing formulae appropriate to an unweighted regression and then changing part way through to formulae appropriate to a weighted regression. The two uses of branch lengths should be altered in harmony to retain statistical coherence.

Another example also illustrates the disjointed nature of Harvey & Pagel's method. Harvey & Pagel (1991: 151) state that "formally, we should not use parametric statistics to analyse the comparisons derived from any of the three methods described above", namely those of Felsenstein (1985), Grafen (1989) and Harvey & Pagel (1991), "unless we know that the assumptions of those statistics have been met". The implicit claim seems to be that non-parametric statistics would be suitable, suggesting the possibility of "hybrid methods" in which parametric methods such as averaging are used to produce the contrasts and then non-parametric methods are used to analyse the contrasts. Such hybrid methods are open to considerable suspicion for two reasons. Roughly speaking, if assumptions justifying a parametric final analysis are not met, then the assumptions underlying the use of averages to calculate the contrasts will not be met either. Second, as Felsenstein (1985, 1988) has pointed out, when the parametric assumptions are not upheld, the contrast results will not have identical distributions on the null hypothesis, and most non-parametric tests assume identical distributions (they just allow that distribution not to be Normal). Only sign tests would be available, and then only if we are prepared to assume that the distributions, though different, are symmetrical.

Thus, Harvey & Pagel's method treats the construction of the contrast results as leading to a kind of crossroads in the analysis, at which one can look around and decide what to do next. In view of the close connection between the assumptions underpinning the two halves of the analysis, it seems preferable to treat the analysis as a single unit as the phylogenetic regression does. The contrasts are still produced by such a single analysis, as intermediate working values, and can be plotted and interpreted in just the same way. There is a related point here. It is argued above that the contrast vectors must be based on Y, or more precisely the residuals of Y after regressing on the control variables. If this is accepted, then any particular set of contrasts are quite specific in their usefulness, reducing still further the possibility of using them appropriately in more than one way.

The fourth reason for preferring the phylogenetic regression is that there is no natural generalization of their method that tests for more than one X variable. Not only is this important in itself when such an analysis is required, but it casts doubt on the logic underlying the existing method. The natural regression-type method must surely work for arbitrary numbers of X variables.

Having presented four reasons to prefer the phylogenetic regression to the method of Harvey & Pagel (1991), I turn to the arguments they themselves give about the values of different methods. They make no claim for the superiority of their own method. Harvey & Pagel (1991: 120) point out that assumptions about branch lengths, for example, usually have to be made tentatively with no firm justification. They conclude that this cause of doubt about the results makes it less important which particular statistical technique is used. One might equally well argue that, the more uncertainty about the results for other reasons, the more important it is to avoid extra uncertainty by using the correct technique.

In another passage whose conclusion is that the precise technique chosen is unimportant, Harvey & Pagel (1991: 168) show in their table 5.5 the results of a series of simulation results by Martins & Garland (1991). One conclusion they draw from the

results is that the Type I error rates are "not wildly elevated" for techniques known not to be analytically exact. A reader might reasonably be tempted to generalize this conclusion to a comparison of the phylogenetic regression with Harvey & Pagel's own method. However, the methods compared by Martins & Garland are all for binary phylogenies, so some differences between the phylogenetic regression and Harvey & Pagel's method are simply not present. Also the Martins & Garland results look encouraging because the very worst method, the naïve species regression, claims 5% significance only 16% of the time. The "not wildly elevated" error rates of the other tests are around 9%, ranging up to 14%. Nine percent can be seen as 4% above 5%, which can be seen as a small (but I would still argue important) difference; or as over a third of the way from the correct 5% to the species value of 16%. Without further simulations, it is not clear whether the additive or the proportional view will be a better generalization about the relationships between the methods. In my simulations (Grafen, 1989), the species regression had Type I error rates of between 24% and 75%. If the proportional view is even partly appropriate, Martins & Garland's conclusions might not be as encouraging for Harvey & Pagel (1991) as they suggest. It is worth noting that Martins & Garland themselves stress the importance of employing an appropriate technique.

My own simulations (Grafen, 1989) compared a different set of methods to Martins & Garland, and similarly did not contain Harvey & Pagel's method. They show very strong differences between methods, suggesting that it is important to employ the correct technique.

The phylogenetic regression and Harvey & Pagel's method differ in the contrast vectors used to reduce the long to the short dataset, and for technical reasons given above the phylogenetic regression employs the correct technique. Harvey & Pagel (1991: 156–7) adopt the view that the contrast vectors "represent a hypothesis about the branching pattern of the unknown phylogeny", and here I wish to consider if this view, which could easily lead to results conflicting with the technical arguments, is appropriate. Harvey & Pagel's contrasts within a radiation are assigned as follows: each daughter node whose mean is above the mean for the parent node is assigned the same positive value, and each daughter node whose mean is below the mean for the parent node is assigned the same negative value. The ratio of the positive and negative values is assigned by ensuring that the sum of all the values is zero. The absolute magnitude of the values is unimportant because "standardization" is later performed by dividing by the sum of the absolute assigned values. (It is not made clear why this method of standardization was thought to be appropriate. Statistical theory in combination with the Brownian motion model determines the variance of the contrast results, and therefore determines a quite different method of standardizing if the aim is to equalize the error variances of the datapoints in the short regression.)

Harvey & Pagel explain their view of the contrasts by saying that "tips that received a positive weight are implicitly being treated as more closely related to each other than to the tips that receive a negative weight" (Harvey & Pagel, 1991: 156). They interpret the phylogenetic regression's contrast vectors thus: "the observations that deviate from the regression line in the same direction are more closely related to each

other than those that deviate in opposite directions" (Harvey & Pagel, 1991: 157). One possibility is that the contrast vectors logically need to represent a hypothesis about phylogeny, in order for the methods to work. This is certainly not true—if it were, the fact that those hypotheses were wrong would invalidate the methods. Fortunately, there is no reason whatsoever why the contrast vector has to represent a hypothesis about the unknown phylogeny. Is it then wise to choose contrast vectors so that they do represent such a hypothesis? The role of phylogeny throughout the method is that of a cause of covariation that requires disentangling. If we do not need to consider it, and validity is ensured, it is natural to choose the contrast vectors to make the method as powerful as possible. The phylogenetic regression's contrast vectors give more weight to observations that are less well explained by the control variables, which might be expected to lead to reasonable power, and indeed reasonable power is attained in the simulations I report (Grafen, 1989: 140). So even if the contrasts were not dictated by logic, there would be good reason to be happy with them. It is probably no coincidence that purely technical properties lead to reasonable power: multiple regression itself can be derived as the unique way to implement certain technical conditions and then afterwards be shown to have very strong (indeed optimal) power properties.

The division into two groups brought about by Harvey & Pagel's contrast vectors implies a discontinuity in the method. As one species value changes, and higher means change accordingly, there will be instantaneous movement of daughter nodes from below their parent node's mean to above, leading to discontinuous changes in the final F-ratio. This is a very unattractive feature. Fortunately, Harvey & Pagel's (1991) "phylogenetic hypothesis" view of the contrast vectors lacks any basis in logic: we do not need to take it, and there is no reason why we should choose to take it.

A final point for comparison between the methods is how they have been justified. A technique like multiple regression has its major results proved in many statistical textbooks (e.g. Cox & Hinkley, 1974), usually refined and generalized over the original justifications given when the test was first introduced. Fisher (1922) introduced multiple regression. Fisher is known to biologists as a population geneticist and evolutionist, but he also invented and developed analysis of variance and covariance, discriminant function analysis, probit and logistic regressions and many other statistical techniques. Even the familiar *t*-tests need and possess mathematically rigorous justifications (for clear and simple versions see Bulmer, 1979). The vast majority of the techniques that standard packages implement have highly technical justifications somewhere in the statistical literature. It is not considered necessary in biology to understand the justification in order to employ the method. Many biologists are so remote from these justifications, however, that they may be unaware that justifications exist, or they may simply take it on faith that any publicly recommended method must have a rigorous justification. I showed (Grafen, 1989) analytically that for a fixed and known phylogeny, the phylogenetic regression is an exact test. I performed simulations for the more complex case in which the phylogeny is not known. These simulations showed the phylogenetic regression to be approximately valid, and to have reasonable power. In other words I provided an analytical proof

that came as near as possible to the result that really mattered, and then bridged the gap by simulations. Felsenstein's (1985) method is analytically justified. In contrast, Harvey & Pagel (1991) provide no analytical work and no simulations to justify their method.

To conclude, Harvey & Pagel (1991) give two quite unconvincing reasons why choice of statistical technique may be unimportant. Four substantial reasons to prefer the phylogenetic regression are given above. The difficulty of the theorems about the phylogenetic regression in Grafen (1989) may be somewhat discouraging to potential users. But it is the fact that proofs exist which is important, and biologists who take proofs of other statistical techniques on trust are free to do the same for the phylogenetic regression.

Conclusions

A revolution is going on in statistical methodology for comparative data. It is now possible to apply principled statistical techniques to various kinds of data (Ridley, 1983; Felsenstein, 1985, 1988; Grafen, 1989). Any method that supplies the hypothesis testing facilities of general linear models, that is for continuous Y variable and arbitrary numbers of continuous and categorical X variables and their interactions, must be essentially equivalent, except as regards fitting parameters that vary branch lengths, to the phylogenetic regression (Grafen, 1989). The method of Harvey & Pagel (1991: 150–2, 157–62) is statistically incoherent.

I am grateful to Mark Ridley, Joe Felsenstein, William Kirk, Ted Garland, Wayne Maddison, Berni Crespi, Bill Hamilton, Richard Dawkins, Craig Packer, Andy Purvis, Paul Harvey and Mark Pagel for helpful comments and discussion.

REFERENCES

BOOMSMA, J. J. & GRAFEN, A. (1990). Intraspecific variation in ant sex ratios and the Trivers–Hare hypothesis. *Evolution* **44,** 1026–1034.

BULMER, M. G. (1979). *Principles of Statistics.* New York: Dover Publications.

COX, D. R. & HINKLEY, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

EDWARDS, A. W. F. & CAVALLI-SFORZA, L. L. (1964). Reconstruction of evolutionary trees. In: *Phenetic and Phylogenetic Classification* (Heywood, V. H. & McNeill, J., eds) pp. 67–76. London: Systematics Association.

FELSENSTEIN, J. (1985). Phylogenies and the comparative method. *Am. Nat.* **125,** 1–15.

FELSENSTEIN, J. (1988). Phylogenies and quantitative characters. *Ann. Rev. Ecol. Syst.* **19,** 445–471.

FISHER, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *J. R. Stat. Soc.* **85,** 597–612.

GRAFEN, A. (1989). The phylogenetic regression. *Phil. Trans. R. Soc. B (Lond.)* **326,** 119–157.

HARVEY, P. H. & PAGEL, M. D. (1991). *The Comparative Method in Evolutionary Biology.* Oxford: Oxford University Press.

KIRK, W. D. J. (1990). *Body Size in Flower Thrips.* Proceedings of the Third International Symposium on Thysanoptera. Poland: Kazimierz Dolny.

KIRK, W. D. J. (1991). The size relationship between insects and their hosts. *Ecol. Entomol.* **16,** 351–359.

LYNCH, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution* **45,** 1065–1080.

MARTINS, E. P. & GARLAND, T. JR (1991). Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. *Evolution* **45,** 534–557.

PACKER, C., LEWIS, S. & PUSEY, A. (1992). A comparative study of non-offspring nursing. *Anim. Behav.* **43,** 265–281.

PAGEL, M. D. & HARVEY, P. H. (1989). Comparative methods for examining adaptation depend on evolutionary models. *Folia Primat.* **53,** 203–220.

PROMISLOW, D. E. L. (1991). Senescence in natural populations of mammals: a comparative study. *Evolution.* **45,** 1869–1887

READ, A. F. (1987). Comparative evidence supports the Hamilton and Zuk hypothesis on parasites and sexual selection. *Nature, Lond.* **327,** 68–70.

RIDLEY, M. (1983). *The Explanation of Organic Diversity.* Oxford: Clarendon Press.

STONE, G. N. & WILLMER, P. G. (1989). Warm-up rates and body temperatures in bees: the importance of body size, thermal regime and phylogeny. *J. exp. Biol.* **147,** 303–328.

WASSERMAN, S. S. & MITTER, C. (1978). The relationship of body size to breadth of diet in some Lepidoptera. *Ecol. Entomol.* **3,** 155–160.