Phil. Trans. R. Soc. Lond. B **326**, 119–157 (1989) [119]
Printed in Great Britain

THE PHYLOGENETIC REGRESSION

BY A. GRAFEN†
Animal Behaviour Research Group, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, U.K.

(*Communicated by W. D. Hamilton, F.R.S. – Received* 13 *February* 1989)

CONTENTS

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

A new statistical method called the phylogenetic regression is proposed that applies multiple regression techniques to cross-species data. It allows continuous and categorical variables to be tested for and controlled for. The new method is valid despite the problem that phylogenetically close species tend to be similar, and is designed to be used when information about the phylogeny is incomplete. Information about the phylogeny of the species is assumed to be available in the form of a working phylogeny, which contains multiple nodes representing ignorance about the order of splitting of taxa. The non-independence between species is divided into that due to recognized phylogeny, that is, to phylogenetic associations represented in the working phylogeny; and that due to unrecognized phylogeny. The new method uses one linear contrast for each higher node in the working phylogeny, thus applying the 'radiation principle'. For binary phylogenies the method is similar to an existing method.

A criterion is suggested in the form of a simulation test for deciding on the acceptability of proposed statistical methods for analysing cross-species data with a continuous y-variable. This criterion is applied to the phylogenetic regression and to some other methods. The phylogenetic regression passes this test; the other methods tested fail it.

Arbitrary choices have to be made about the covariance structure of the error in order to implement the method. It is argued that error results from omitted but relevant variables, and the implications for those arbitrary choices are discussed. One conclusion is that the dates of splits between taxa, even supplemented by rates of neutral gene evolution, do not provide the 'true' covariance structure. A pragmatic approach is adopted.

Several analytical results about the phylogenetic regression are given, without proof, in a mathematical appendix.

A computer program has been written in GLIM to implement the phylogenetic regression, and readers are informed how to obtain a copy.

1. Introduction

In this paper I present the phylogenetic regression, a new statistical method for analysing cross-species data in a regression framework with a continuous y-variable. The new method supplies the hypothesis testing facilities of multiple regression in a way that takes account of the special difficulty associated with comparative data, namely non-independence of species. Arbitrary combinations of continuous and categorical variables can be controlled for and tested for.

Section 2 establishes necessary preliminaries, then §3 develops the phylogenetic regression. Section 4 proposes a criterion for the acceptability of any statistical regression method of analysing comparative data with a continuous y-variable; §5 applies this criterion to the phylogenetic regression and some other methods. Section 6 discusses an arbitrary choice that

has to be made in assigning lengths to path segments in a representation of the phylogeny, and what attitude should be taken to it on biological grounds. In §7 the ahistorical nature of the method is asserted, and general conclusions are drawn about the advantages of the phylogenetic regression. Section 8 discusses the ease of implementation with a program written by the author for the purpose, and gives information on how to acquire a copy of it. The mathematical appendix (§10) defines formally the phylogenetic regression, and states without proof four theorems about the phylogenetic regression. The proofs have been omitted for reasons of space, but have been lodged in the archives of the Royal Society and the British Library Document Supply Centre†.

For a general introduction to the statistical problems of the comparative method, and reviews of extant methods, the reader is referred to Ridley (1983) and Pagel & Harvey (1989). Papers based on the phylogenetic regression are currently being prepared. The present paper concerns only the principles of hypothesis testing with comparative data in a regression framework with a continuous y-variable, and is not intended to be introductory.

2. Preliminaries

In this section, necessary preparation is made for developing the phylogenetic regression. In §2*a* the distinction is drawn between similarity due to recognized and unrecognized phylogeny, and a method of representing the first of these is developed. Similarity due to recognized phylogeny is treated by developing the *standard regression*, a generalization of Felsenstein's (1985) comparative method to the case of non-binary trees, which also incorporates a degree of flexibility in assumptions about the covariance structure. Similarity due to unrecognized phylogeny will be treated in §3 by using the *radiation principle* of Ridley (1983), which is therefore explained in §2*c*.

(*a*) *Similarity due to recognized phylogeny*

Some representation of the phylogeny of the species will be available when a comparative analysis is to be performed. This may only be the taxonomic divisions into genera, families, orders and classes, or it may be a more detailed attempt at a phylogeny. I shall assume that one will be taken as the best available representation of the phylogeny: call this the 'working phylogeny'.

The phylogenetic associations represented in the working phylogeny constitute *recognized phylogeny*. Most true phylogenies probably contain only binary nodes. If in the working phylogeny there are nodes with many daughter nodes, then this probably represents our ignorance about the order of splitting within the set of daughter nodes. Phylogenetic closeness present in the true phylogeny, but absent from the working phylogeny, constitutes *unrecognized phylogeny*. Both types of phylogeny cause statistical problems, and the solutions to them are quite distinct.

There is an important assumption that must be made about the working phylogeny in order to make progress. The groups defined by the working phylogeny must be monophyletic. In other words, the working phylogeny is a *valid coarsening* of the true phylogeny. A valid coarsening is obtained by uniting adjacent nodes in a phylogeny, an operation which corresponds to admitting ignorance about the order of splitting. Transferring one daughter

† Copies of the material deposited may be purchased from the British Library Document Supply Centre, Weatherby, West Yorkshire LS23 7BQ, U.K. (reference SUP 10052).

node from one parent to another is not allowed. The working phylogeny is a mixture of the true phylogeny and ignorance about the order of splitting, but is permitted to contain no downright errors.

The converse of *valid coarsening* is *compatible refinement*, a notion that will be useful in §§4 and 5. All the binary phylogenies that can be coarsened to give the working phylogeny are compatible refinements of the working phylogeny. The compatible refinements of the working phylogeny are all the possible true phylogenies.

This assumption has implications for how to choose a working phylogeny. Avoid being spuriously exact about the order of splits, and be content to express ignorance about the order of splitting as multiple nodes. It is unfortunate in this respect when phylogenies based on DNA–DNA hybridization, such as that of Sibley *et al.* (1988), are presented without any indication of the reliability of the positioning of taxa in the phylogeny.

The standard regression is a generalization of Felsenstein's (1985) comparative method. His method of representing phylogenetically based similarity will be adopted here for similarity due to recognized phylogeny. The aim of the exercise is to specify how similar each pair of species should be by obtaining a covariance from the tree of the working phylogeny.

Beginning with a tree, as in figure 1, we assign lengths to each path segment. The idea is that the covariance between two species is obtained as the shared path length in the paths from the top of the tree to the two species' nodes. The variance of a species is therefore given by the total length of the path from the top to the species' node. This procedure provides a variance for each species and a covariance for each pair of species. This set of variances and covariances is exactly what many statistical methods require to work correctly in the presence of non-independence.

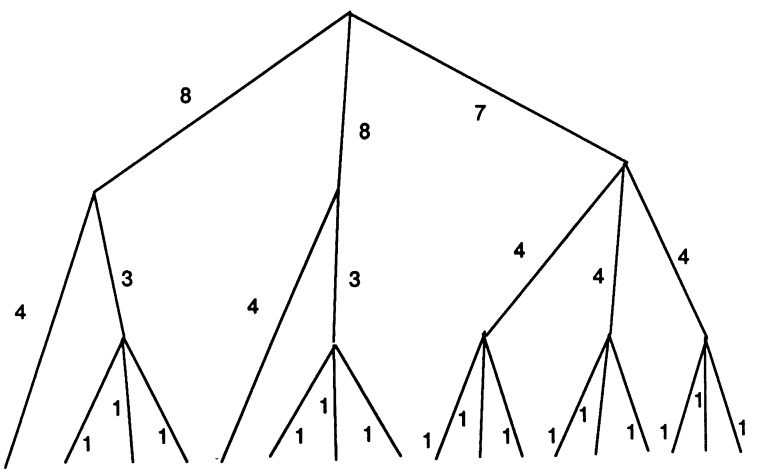

FIGURE 1. The figure shows a phylogeny with path-segment lengths indicated. The root of the tree is at the top, and species are represented by the tips at the bottom. Time therefore moves forward as we move down the diagram. Notice that the total length from the root to each species tip is the same. This reflects what will usually be a reasonable assumption that the error variance of each species is the same. The analyses developed in the text do not depend on this assumption, however.

Notice that the tree-based covariances have a good phylogenetic pattern. The closer two species are in the phylogeny, the higher the covariance between them.

This method of deriving covariances from the tree has one arbitrary element, namely the initial assignment of lengths to the path segments. To provide some flexibility in the tree, a

whole family of sets of lengths will be generated, which differ in the level at which most variation occurs. In the statistical methods being developed, the data themselves will be allowed to choose which member of the family fits best.

The generation of the family begins with the specification of an initial tree with lengths. First, I will explain how the initial tree might be found, and then how the whole family is generated from it. There are various ways to find this initial tree. One way is to assign a height to each node as one less than the number of species below or at that node. Then each path segment's length is the difference between the height of the upper and lower nodes. This method is illustrated in figure 2, and it has the advantage of treating taxonomic ranks as arbitrary: the ranking habits of taxonomists are excluded from our study. Another way is to assign arbitrary heights to given taxonomic levels, such as zero to species, one to genera, two to families, three to orders and four to classes. The length of each path segment is again given by the difference between the heights of the upper and lower nodes. This method would be appropriate if it was thought that taxonomic rank was a reasonable indicator of expected divergence in the omitted variables which constitute the error. There are many alternative methods for devising the initial set of lengths.

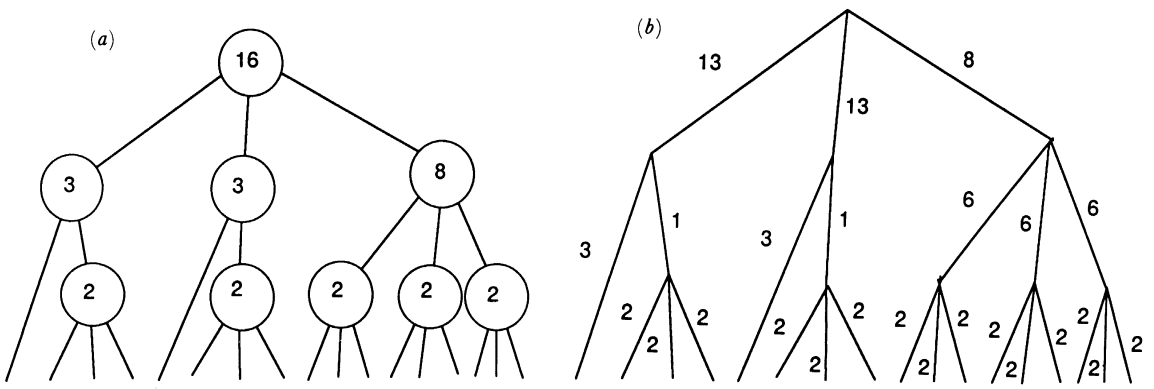

FIGURE 2. The figure demonstrates one way of assigning path-segment lengths to a tree. Each node is given a number, calculated as one less than the number of species below that node in the tree. All species nodes therefore receive a value of zero, which is not shown, but the 'heights' of all higher nodes are shown inside the circles in (*a*). The length of a path segment is computed as the difference between the heights of its upper and lower nodes. These lengths are shown in (*b*).

Both of these methods involve initially assigning a height to each node, and then finding the length of each path segment as the difference between the heights of its upper and lower nodes. The method for generating the family involves transforming the heights in the following way. First scale the heights so that species have a height of zero and the top of the tree has a height of one. Now we can apply a family of mathematical transformations to the heights in order to create the family of sets of lengths. The transformations that will be used in the methods being developed are just the power transformations. Each height will be raised to a given positive power. This leaves the species at zero and the top of the tree at one, but it distorts the tree either by compressing near the bottom and expanding near the top (if the power is greater than one), or compressing near the top and expanding near the bottom (if the power is less than one). Samples of these distortions are shown in figure 3. This power will be called ρ.

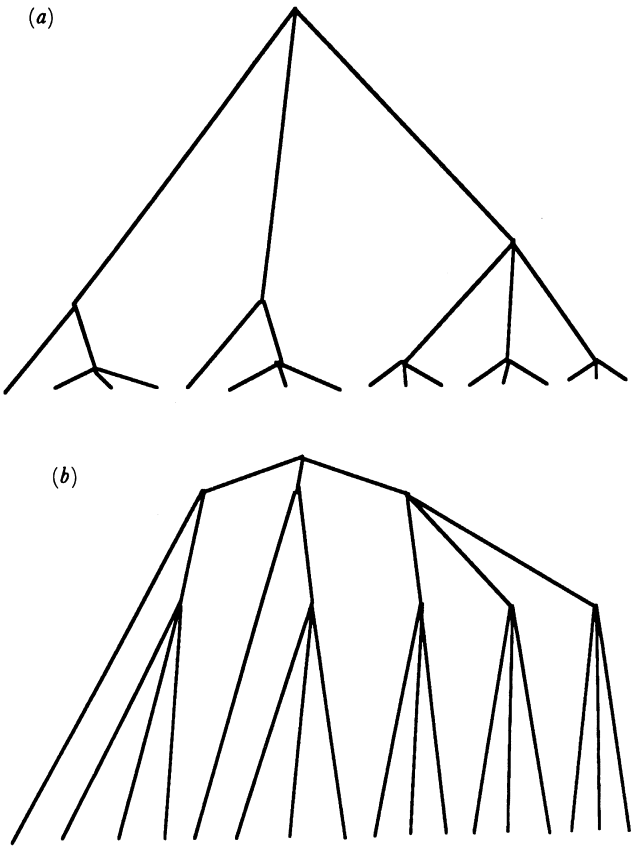

FIGURE 3. The figure demonstrates the distorting effect of the parameter ρ on the phylogeny of figure 1. A high value of ρ increases the lengths of segments high in the phylogeny, as in (*a*). In this case, the radiations low in the phylogeny would be more reliable evidence, as increased path-segment length means more error. A low value of ρ increases the lengths of segments low in the phylogeny, as in (*b*). In this case radiations high in the phylogeny would provide more reliable evidence.

(*b*) *The standard regression*

When using a method for assigning variances and covariances such as that just described, there are standard statistical techniques available that provide valid analyses in the presence of non-independence due to recognized phylogeny. Multiple regression can be performed in the usual way provided the pattern of variances and covariances is specified (see, for example, Johnston 1972). This technique is sometimes called Generalized Least Squares.

The family of tree-lengths can be used by finding the best-fitting value of the positive power and the best-fitting regression parameters simultaneously by maximum likelihood. This is a satisfactory method of dealing with similarity due to recognized phylogeny, and uses only routine statistical techniques. It is the least that can reasonably be done when performing a regression on comparative data. Because this is a satisfactory maximum likelihood method of dealing with similarity due to recognized phylogeny, it will be of interest to compare it with the new method I propose in §3. I shall refer to it as the *standard regression*, because it uses only standard techniques. It is described more formally in §3 and defined in §10.

The standard regression is a generalization of Felsenstein's (1985) method. It can be applied to non-binary trees, and it allows a degree of flexibility in the covariance structure. The aim

of these extensions is to create a statistical method that can be routinely applied to the kind of data commonly encountered in comparative biology. The case of a binary tree is special, for there is no unrecognized phylogeny, and so the standard regression is a fully satisfactory method. The extension of Felsenstein's method to non-binary trees requires a way of dealing with unrecognized phylogeny.

(*c*) *The radiation principle and similarity due to unrecognized phylogeny*

The problem of similarity due to recognized phylogeny, then, has an easy and acceptable solution which I have just named the standard regression. If the phylogeny is known in full, this solves the whole problem. However, in most cases full phylogenies are not known, and working phylogenies have multiple nodes. The unrecognized phylogeny can have the same effect on the standard regression as recognized phylogeny has on the naïve regression in which all species are taken as independent. Some pairs of species will be more closely related than the statistical method assumes, and other pairs less so. Points treated as independent by the statistical method will not be independent. The new method I will propose aims to give robust conclusions in the presence of similarity due to unrecognized phylogeny.

Each higher (that is, non-species) node together with its immediate daughter nodes can be considered to be a 'radiation'. Originally one species, it has branched into a number of taxa. The principle which allows similarity due to unrecognized phylogeny to be dealt with is to use each radiation as an independent data point. If the same relationship between two variables arises in two radiations, this cannot result from phylogenetic similarities, either recognized or unrecognized. The same relationship must have arisen independently in the two cases. Observing in an informal way that a relationship has arisen frequently in evolution has no doubt given justifiable confidence to a number of biologists that the relationship they are studying is real. But it was Ridley (1983) who advanced the techniques of comparative biology by inventing (what I call) the radiation principle, which is that formal statistical tests should be constructed by using each radiation as an independent data point.

Two radiations in a working phylogeny can have one of two relationships with each other. The first is that neither is ancestral to the other. These radiations cannot be dependent, as they depend on quite different data. The second relationship is that one radiation is a descendant of the other. The fact that a positive correlation exists between two characters across the genus means within a family does not imply that within any one genus there will be a similar correlation. If there is, then this is independent evidence of an association between the two characters. Freedom from the taint of similarity due to unrecognized phylogeny is achieved by counting each radiation only once. If a radiation has ten daughters, it is tempting to assign ten degrees of freedom to it in total, leaving eight for the assessment of the slope of one character on another. But ten degrees of freedom are appropriate only if the daughters are independent, or if the extent of interdependence of each pair of daughters can be exactly specified. In ignorance of the true phylogeny, this is exactly what we cannot do. The evidence from each radiation can be safely used, however, if only one degree of freedom is assigned to it.

The principle we shall follow, then, is that radiations can be taken as independent. To ensure independence in a technical sense requires that the averages at higher nodes are correctly weighted, and this is discussed further in §3*a* and 5*e* and 10. Several existing methods are based on the radiation principle: the method of Ridley (1983), applied by Ridley (1983, 1986, 1989*a*, *b*), and Sillén-Tullberg (1988); the method of Read (1987); a hybrid method applied

by Krebs *et al.* (1989); and the method of Felsenstein (1985), applied by Sessions & Larson (1987). Applying the radiation principle in a general regression framework is the purpose of the next section.

3. The phylogenetic regression

In this section the phylogenetic regression is developed, by applying the radiation principle to the standard regression. This involves a transformation of the standard regression, called 'hanging on the tree', explained in §3*a*. This is followed by a process which extracts one data point for each radiation in the working phylogeny, by using linear contrasts, as explained in §3*b*. The phylogenetic regression is a straightforward multiple regression applied to this reduced data set. To my knowledge, the process of extraction is statistically original. The method as a whole has an alternative interpretation, developed in §3*c*, as a restricted randomization test, which non-statisticians may find helpful. Section 3*d* explains the statistical properties of the phylogenetic regression, and §3*e* elucidates the degrees of freedom associated with tests using the phylogenetic regression.

It is convenient now to establish some notation. The method is a regression method with one continuous y-variable, and this will be called y. The x-variables being controlled for will be called X, and the x-variables being tested will be called Z. y is a vector, and X and Z are matrices. Let the parameters for X and Z be β and γ. The deterministic part of the model can now be expressed as

$$\boldsymbol{E}(y) = X\beta + Z\gamma,$$

and the null hypothesis that Z is irrelevant to y is written formally as $\gamma = 0$. (For convenience in this section, the constant is taken to be one of the columns of X. Elsewhere in the paper it is written separately.)

The deterministic part is the same for the naïve species regression and the standard regression described in §2*b*. They are distinguished by the stochastic parts of the model. If the error term is ϵ, then

$$\epsilon = y - X\beta - Z\gamma$$

and the species regression assumes that

$$\boldsymbol{E}(\epsilon\epsilon^T) = \sigma^2 \boldsymbol{I},$$

that is, the errors for each species are equal and there are no covariances between the errors. The standard regression assumes that ϵ is multi-normally distributed with zero mean and that

$$\boldsymbol{E}(\epsilon_i \epsilon_j) = \sigma^2 V_{ij}(\rho) = \sigma^2 (1 - h_{ij}^{\rho}).$$

The power ρ represents the power to which the heights are raised before computing the path segment lengths. The variances and covariances therefore depend on ρ as well as on σ^2. h_{ij} is the height in the initial working phylogeny at which the paths to species i and j diverge. Species have a height of zero, and the top of the tree has a height of one.

In the standard regression introduced in §2*b*, ρ is fitted simultaneously by maximum likelihood with the regression coefficients β and γ.

(*a*) '*Hanging a variable on the tree*'

The phylogenetic regression works by extracting one data point from each radiation. It is necessary to explain first a transformation of a variable from its original form of one number

for each species, into a form in which the information is spread through the working phylogeny. Taking the working phylogeny, with the values of y for each species represented at the species tips, we can work up the tree writing at each higher node the mean value of y for the daughter nodes. Then, we can transform these values by representing each node's value as a deviation from its parent node's value. The values of y are now expressed incrementally down the tree. Figure 4 illustrates this process. The process can be carried out for any variable, and may be called 'hanging on the tree'.

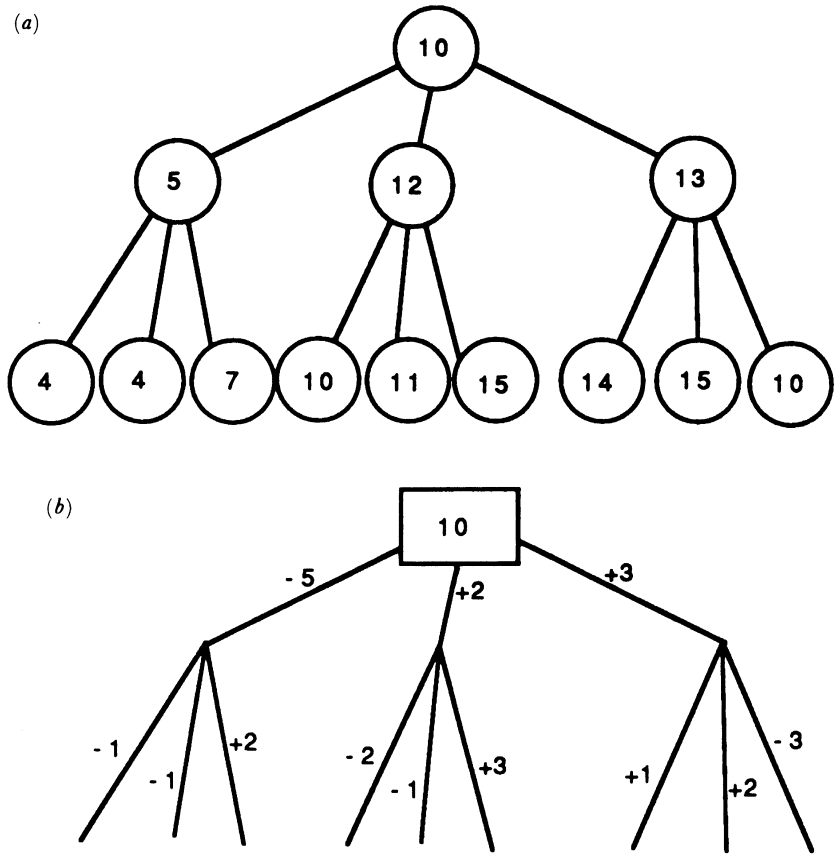

FIGURE 4. (*a*) The species measurements at the bottom have been averaged up the tree to yield means at the next set of nodes, and then those node means have been averaged to yield the highest node's mean. In a larger phylogeny, this process would continue up the tree until every higher node had a mean calculated for it. The means in this example are unweighted. This is appropriate if the path segment lengths are all equal. In general, the means are weighted so as to be efficiently estimated according to the variance–covariance structure implied by the path-segment lengths. The weights are defined in the matrix L in § 10. In (*b*), each path segment has been assigned a value by subtracting the upper node's value from the lower node's. The mean for any node can now be calculated by starting with the grand mean and adding the values found on the path from the root of the tree to that node. Thus the original species data has been re-represented as a hierarchy of phylogenetically arranged differences from a grand mean.

Each radiation can now be drawn as in figure 5, with one point for each of the daughter nodes. Only two variables can be conveniently represented, but the principle applies to any number.

The means used are weighted means. The weights are derived using the path segment lengths of the working phylogeny. They are chosen to make each mean an efficient estimate,

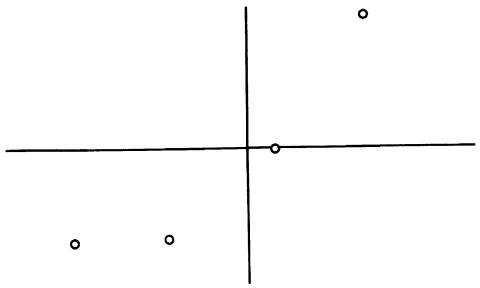

FIGURE 5. The axes represent two continuous characters. Each point represents one of the four daughter nodes of a higher node. A daughter node's point is plotted as the deviation of the mean for all species below the daughter node from the mean for all species below the parent node. This is why the mean value of the points must equal zero on both axes. If a regression through the origin gives a non-zero slope, this is evidence from within the radiation for a relationship between the two variables. The principle of this two-dimensional plot extends to many dimensions.

on the assumption that the variable concerned is normally distributed with a variance–covariance structure given by the path segment lengths in the working phylogeny. The error of a mean at a node therefore comprises a part due to the path segment lengths directly below the node, and a part due to the sampling errors of the means of the nodes at the end of those path segments. The matrix L defined in §10 contains these weights.

(*b*) *Linear contrasts*

The phylogenetic regression reduces each radiation in the form of figure 5 to one data point by forming a linear contrast of the points on the graph. A linear contrast is a weighted sum, in which the weights themselves sum to zero. The same linear contrast can be applied to all the variables in a data set, so that the resulting data point has as many variables as the original data point. Figure 6 illustrates the regression formed from the linear contrasts, which contains one data point for each radiation.

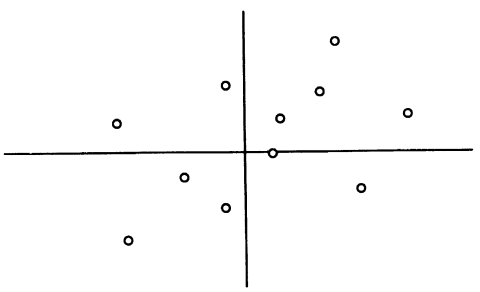

FIGURE 6. The axes represent two continuous characters. Each point represents one radiation in the working phylogeny. Each point is derived from a plot like figure 5 for its own radiation, and is a linear contrast of the points in that plot. A linear contrast is a weighted sum whose weights sum to zero. If all the points lie on the same line through the origin, then the linear contrast must also lie on that line. If all the radiations have a plot in which the daughters' points lie on the same line, then this is strong evidence for an association between two characters. This situation would result in all the linear contrasts also lying on that line, so the strength of evidence is passed onto the plot shown here. The regression analysis performed on a plot like this is called the short regression. The principle of transferring from radiation plots (as in figure 5) to the short regression plot (in this figure) is shown for two characters, but is performed in exactly the same way for any number of characters.

In this final data set, the effects of unrecognized phylogeny have been eliminated by condensing each radiation into one data point. The variance–covariance matrix of the errors

in this reduced, or 'short', regression, can be computed from the errors in the standard regression and the linear contrasts used. Conventional regression techniques can therefore be applied to this short regression, and their results will be valid despite similarity due to unrecognized phylogeny. The test of the hypothesis $\gamma = 0$ in this short regression is the eponymous test of this paper: the phylogenetic regression.

The choice of linear contrasts has not yet been explained. They are defined precisely in the Appendix as the matrix product GC^{-1}. The linear contrasts are derived from the residuals in the standard regression of y on X. These residuals are 'hung on the tree'. After separate scaling within each radiation, these residuals provide the linear contrasts. (The scaling is used, purely for convenience, to ensure that the variance–covariance matrix of the short regression has a particular form.) Any choice of linear contrasts which did not depend on Z or ϵ would ensure validity. However, to achieve good power, and to avoid making arbitrary choices, it was desirable to have the contrasts depend on ϵ. This particular choice was made to ensure that despite this dependence the test is still valid (see theorem 2 of §10). These linear contrasts weight those points more heavily which are less well explained by X alone.

An alternative interpretation established in theorem 3 of §10 is that, in the test for Z, we condition on certain properties of the residual from the regression of y on X. These properties are the 'pattern' of the residual at each radiation, to be explained in §3*c*. The process of extracting one data point from each radiation is, to my knowledge, statistically novel. It is hoped that readers will find at least one of the three interpretations readily understandable.

The phylogenetic regression therefore applies the radiation principle in having one data point in its data set for each radiation. The method of condensation allows a data point to have any number of variables, and so arbitrary numbers of variables can be controlled for and tested for in the short regression.

(*c*) *A randomization test*

The phylogenetic regression can also be explained as a randomization test, and biologists have found this alternative perspective helpful. Readers who have found the weighted sums and linear contrasts perfectly clear may wish to skip this subsection. In the standard regression, the test statistic for Z, controlling for X, would be an F-ratio in the usual way as follows:

$$F = \mathrm{MS}_Z/\mathrm{MS}_{\mathrm{error}},$$

in which MS stands for mean square. This F-ratio would be looked up in tables. The point of using radiations as replicates is that the distribution of this F-ratio is unreliable, and probably too optimistic about the significance of Z.

The idea of the randomization test is to construct random 'alternative Zs', and to compare the explanatory power of the observed Z with that of the alternatives. Ordinary regressions can be interpreted as randomization tests of this sort, and the special properties of the phylogenetic regression will be easier to understand in that context. Let $F(y, X, Z)$ stand for the F-ratio obtained by testing for the significance of Z, controlling for X, in explaining the variation in y. The null hypothesis distribution for the F-ratio is conventionally thought of as being generated by replacing y with a normally distributed random variable, say ν. Then if ν is a normally distributed vector, the null hypothesis distribution is generated by $F(\nu, X, Z)$. This can be thought of as comparing Z's ability to explain y with Z's ability to explain random numbers.

An alternative view is to keep y fixed, but to substitute random numbers for Z. Suppose that

Z has n_Z degrees of freedom associated with it. Then let Ξ be a random matrix, each of whose n_Z columns is identically and independently distributed in a normal distribution. Then comparing $F(y, X, Z)$ with the distribution generated by $F(y, X, \Xi)$ is comparing Z's ability to explain y with the ability of an equivalent number of random variates to explain y. Provided v and each column of Ξ all have the same variance–covariance matrix as that assumed in the computation of the sums of squares, these two tests are the same. The F-ratio would have the same probability distribution in the two randomization tests.

The reason for explaining this interpretation of ordinary regressions as a randomization test is that the phylogenetic regression can also be understood as a randomization test, which imposes a restriction on the possible values which the Ξs, the randomized alternative versions of Z used to generate the null hypothesis distribution of the test statistic, may take.

To explain this restriction, it is necessary to decompose Z into two different kinds of information. By hanging Z on the tree, each radiation can be drawn as in figure 7, with one point for each of the daughter nodes. The relative values of the daughter's deviations may be called the *pattern* of the radiation. Each Z-variable has its pattern at each radiation. By conserving the patterns, all correlations within a radiation are also conserved. This includes correlations between one Z-variable and another, and between one Z-variable and the y-variable. Correlations including only a subset of the daughters are also conserved if the pattern is conserved. These intra-radiation correlations are the untrustworthy part of the information about Z, because they may be contaminated by unrecognized phylogeny. Two clumps of values may represent distinct sister groups within the radiation.

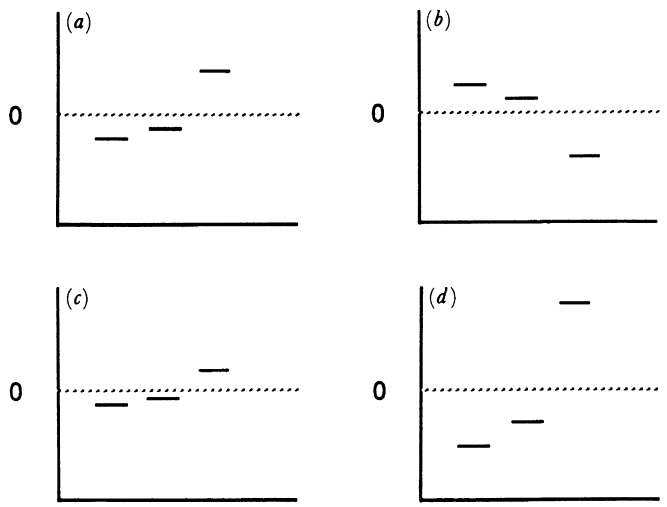

FIGURE 7. Part (*a*) Values attached to the path segments at the middle node of figure 4, which are deviations of −2, −1 and 3. Parts (*b*), (*c*) and (*d*) have the same pattern as the original values, because they are in the same ratio, but their magnitudes are different.

Apart from pattern, there is only one degree of freedom left (for each Z-variable) in describing the daughter's deviations at a radiation, their *magnitude*. The magnitude is the absolute size of the deviations, and is scaled so that it equals one if the variance of the daughter's values equals the variance of the residuals in the corresponding radiation after the regression on y on X.

Each Z-variable in each radiation has a pattern and a magnitude. The restriction which the

phylogenetic regression places on each Ξ is that it should share the same patterns as Z at all radiations. Only the magnitudes differ. The rationale for this is that each Ξ has the same set of correlations within each radiation with itself and with y as Z does. As all the contamination of unrecognized phylogeny is contained in those intra-radiation correlations, each Ξ has the same advantage from unrecognized phylogeny in explaining y as Z does. It follows that if Z explains a significant fraction of the variation in y, compared with the null hypothesis distribution created using the restricted Ξ, then this cannot be due to unrecognized phylogeny. That is the basis of the randomization interpretation of the phylogenetic regression.

Figure 8 shows an example of a Z, and of Ξ values formed which share its patterns but differ in their magnitudes. It illustrates that they share the general 'phylogeneticness' of Z, and also share the same correlations within each radiation.

Theorem 4 of §10 states formally the equivalence of the randomization test with the phylogenetic regression.

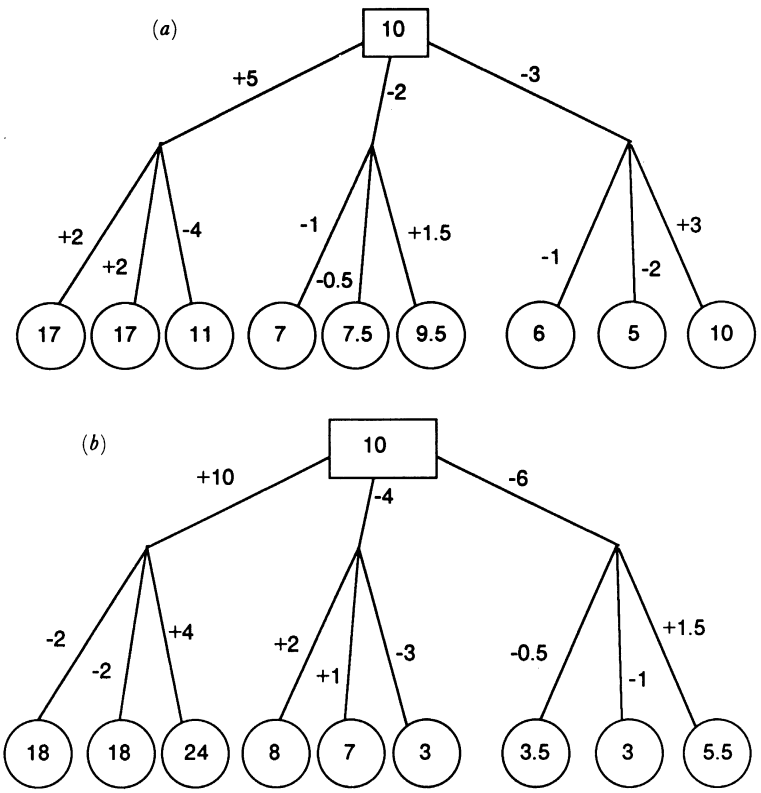

FIGURE 8. This represents two instances of randomized versions of the phylogeny in figure 4, in which the magnitudes of the deviations have been altered. In (*a*), the factors are -1 for the top node, and -2, 0.5 and -1 for the lower nodes. In (*b*) the factors are -2 for the top node, and 2, -1 and -0.5 for the lower nodes. The data in each tree therefore shares the same degree of 'phylogeneticness' as the original.

(*d*) *Statistical properties*

The purpose of this section is to explain briefly what is known about the statistical properties of the phylogenetic regression. It is proved in §10 that if the working phylogeny is the true phylogeny, and if the value of ρ is taken as fixed and known, then both the standard regression and the phylogenetic regression are exact tests. It follows from general results on Generalized

Least Squares (Johnston 1972) that the standard regression has various optimality properties. It is best linear unbiassed, and uniformly most powerful. Once ρ is unknown and must be estimated, these exact results no longer apply, but give confidence that the tests are reasonable.

A larger step is to relax the assumption that the working phylogeny is the true phylogeny. In the test to be proposed in §4 for comparative methods, the 'true phylogeny' is selected as a random compatible refinement of the working phylogeny. For data created in this way, in which both the error and the Z-variable that is being tested for are created using the same true phylogeny, the standard regression can be expected to be invalid. Correlations between the error and the Z-variable will be created which are due to error correlated with phylogeny, but the standard regression will attribute them to a true relationship.

This problem for the standard regression is dealt with by the phylogenetic regression by using only one data point for each radiation. Although the radiations are in a loose sense 'independent', the points in the short regression are not completely independent in a statistical sense, and so the phylogenetic regression can have no pretensions to exactitude. The reasons for this will be considered in more detail in §5*e*. The behaviour of the standard and phylogenetic regressions in the more complex cases have been investigated by simulations reported in §5.

(*e*) *Phylogenetic degrees of freedom*

The final topic in describing the phylogenetic regression is degrees of freedom in the phylogenetic regression, and the points can be made most easily with reference to table 1. The table is drawn up for an example in which a pair of variables Z is tested for its effect on y, controlling for the constant and a set of three variables X. The data set has 49 species, and the working phylogeny has 23 higher nodes. The total degrees of freedom will be considered first. The most obvious point is that the total number of degrees of freedom in the standard regression is the number of species, but is at most the number of higher nodes in the phylogenetic regression. There is no constant in the phylogenetic regression, so one degree of freedom is not allotted to the constant as it is in the standard regression. The main consequence of this reduction is that not so many explanatory variables can be studied. An absolute limit to the number of explanatory variables in the regression is the number of degrees of freedom in it. It is advantageous to have many more degrees of freedom than explanatory variables. This shows the importance of knowing the phylogeny and so increasing the number of degrees of freedom in the phylogenetic regression. It is natural and desirable that the more ignorant we are about the phylogeny the less we should be able to infer about cross-species relationships between variables.

The total degrees of freedom for the phylogenetic regression may be less than the number of radiations, as illustrated in the PHY (ii) column of table 1. This occurs when the residuals from the standard regression of y on X, once 'hung on the tree', are exactly zero at all the daughters of a higher node. This can arise in two ways. The first way is that the X-variables contain a subset which vary only in the radiation in question, a subset sufficiently numerous to ensure that the fitted y-values exactly equal the observed y-values. This is analogous in a non-phylogenetic regression to having a categorical variable that takes one value for all data points except one. That special data point is effectively deleted from the regression because it can be fitted exactly. The second way in which all the residuals in a radiation can be exactly zero is if the observed values of y just happen to lie on the fitted line. This 'just happening' will have probability zero if we assume that the error in the standard regression is normally distributed.

TABLE 1. EXAMPLES OF POSSIBLE DECOMPOSITIONS OF DEGREES OF FREEDOM IN THE PHYLOGENETIC REGRESSION

(Testing for the effect on y of Z, controlling for the constant and X. The data set has 49 species and 23 higher nodes. In the standard regression (STD), the total degrees of freedom is therefore 49, whereas in the usual phylogenetic regression (PHY(i)) the total is 23. Notice that there is no constant in the phylogenetic regression. PHY(ii) shows the case in which there is no variation at one higher node in the working phylogeny after the standard regression of y on the constant and X. That node is dropped, and so the total number of degrees of freedom is 22 not 23. By subtraction, the residual degrees of freedom are also reduced by one. PHY(iii) shows the case in which three variables which are not collinear in the standard regression, namely the three columns of X, are collinear in the phylogenetic regression. This reduces the degrees of freedom associated with X, but leaves the total unchanged. The residual degrees of freedom are therefore increased by one.)

source	STD	PHY(i)	PHY(ii)	PHY(iii)
constant	1	0	0	0
X	3	3	3	2
Z	2	2	2	2
residual	43	18	17	19
total	49	23	22	23

But regression methods are routinely and reasonably used in cases where this cannot be true, for example where the y-variable is an artificial variable containing a ranking of some attribute of the species into a small number of groups. The case where the errors are exactly zero therefore needs to be considered.

No matter how it arises that all the residuals at a radiation are zero, the consequences are the same. As explained in §3*b*, the linear contrasts used to form the short regression are the residuals at the radiations. But if all the residuals at a radiation are zero, then there is no variation left for Z to explain. Hence that radiation contributes no information about the explanatory power of Z. That radiation must therefore be dropped from the short regression. Such radiations would arise if the phylogenetic regression were applied to binary data of the kind to which Ridley's (1983) method is applicable, and would be those radiations which have no variation below them. This establishes a close formal parallel between Ridley's method and mine.

The degrees of freedom associated with a variable will usually be the same in the tests derived from the standard and phylogenetic regressions. The exceptional case, exemplified in the PHY (iii) column of table 1, will now be considered. Two explanatory variables which are not collinear in the standard regression may be collinear in the phylogenetic regression. In the simplest situation, two variables once 'hung on the tree' are zero except in the highest radiation in the tree, but differ there. In the standard regression, the differences within that radiation suffice to prevent collinearity. However, in the phylogenetic regression these two variables will be zero except at the highest radiation, and so will be collinear as one will be a multiple of the other. A welcome consequence is that testing for one of the variables while controlling for the other is not possible, because of the collinearity. This is natural, as the information separating the two variables comes from within one radiation, and by the radiation principle this information cannot be trusted. More complex cases involving more variables and more radiations arise analogously.

These complications were omitted from the discussion earlier in §3 for the sake of clarity. They arise naturally in the formal treatment of the phylogenetic regression in §10.

4. How a solution to the phylogenetic problem can be recognized

Discussions of the merits of different proposed solutions to the statistical problems of comparative analyses have usually been vague in their criteria of acceptability of a method. Most proposals can be classified as at best informal ameliorations of methods that treat species as independent. Only Ridley (1983) and Felsenstein (1985) have made any attempt to prove that a method is correct. I now establish a criterion by which the proper working of the hypothesis testing of any proposed regression method for comparative data can be recognized. The criterion is applied in §5.

There are two basic properties of any statistical test: size and power. The reader is referred to any book on the theory of statistics (e.g. Cox & Hinkley 1974) for discussion of these properties. Size is the chance that significance will be achieved if the null hypothesis is true. A test is valid if the size equals the nominal p-value for every p-value. Tests must be valid to be acceptable. Among the class of valid tests, the preferred test is the one with maximum power. Power is the ability to give significance when the null hypothesis is not true. The power will usually depend on which significance level is chosen, and also on which alternative hypothesis is used.

The quite general notions of size and power are made concrete in a particular case by adopting a model for generating hypothetical data. Proposing a test for phylogenetic methods of analysis is a matter of proposing a model for generating data, which can then be used to find the size and power of any method. To make data in accordance with recognized phylogeny is easy, using the path segment lengths as illustrated in figure 1. Each path segment is allotted an independent normally distributed random variable with variance equal to its length. The data for a species is then the sum of the random variables on the path from the top of the tree to that species' node. This will create species data with the variances and covariances implied by the tree.

In order to include unrecognized phylogeny, I propose instead the following modification of this model. Starting with the working phylogeny, choose at random (following a procedure to be described directly) a binary phylogeny that is a compatible refinement (see §2*a*) of the working phylogeny. Assign path lengths to the binary phylogeny by following the method of figure 2. Then generate the data from the binary phylogeny by using the path segment lengths. This is a two-step process, and each step is to be performed each time a data set is to be generated. This will produce data with patterns resulting from recognized and unrecognized phylogeny.

To be well defined, this method requires a specification of exactly how the compatible refinement is chosen. The particular method I have chosen is as follows. Each non-binary node is dealt with separately and independently, representing the independence of our ignorance of the order of splitting at each multiple node. Suppose a node has n daughters. The first step is for one of the daughters, each having a $1/n$ chance of being selected, to be assigned to group 1. Then another daughter, with each of the remaining n-1 daughters having a $1/(n\text{-}1)$ chance of being selected, is assigned to group 2. Each of the remaining n-2 daughters has an independent chance of one half of joining group 1 rather than group 2. The phylogeny can now be modified as follows. If a group has more than one daughter, then a new node is created as a daughter of the original parent node, and as the parent node of all the daughters in that group. If a group has one daughter, then that daughter remains a daughter of the original parent node. The

original parent node is now a binary node, and by applying the process recursively to all newly formed non-binary nodes, the original multiple node will be replaced by a set of binary nodes which are a compatible refinement of the original multiple node. This procedure implicitly defines a probability distribution over the set of binary compatible refinements of any working phylogeny.

This two-step random procedure involves choosing a phylogeny at random, and then, given the phylogeny, the values of the variables. The first step introduces phylogenetic similarities between species which are not recognized by the working phylogeny. A method of analysis that is valid for data created in this way will therefore have to deal properly with similarity due to unrecognized phylogeny.

The test I propose for regression methods for comparative data is then as follows. Many data sets are generated by using the procedure just described. It is important that both the error and the x-variable being tested for are created phylogenetically, so that the similarity due to unrecognized phylogeny affects them both in the same way. The method being tested is then used to analyse these many data sets. The validity of the test is checked by observing on what fraction of trials nominal p-values are exceeded when the data are constructed under the null hypothesis. The power is investigated with data constructed under an alternative hypothesis. This test is applied in §5 to the phylogenetic regression and some other methods.

There are two arbitrary elements in the test I have proposed. Some other choice could have been made for the probability distribution over the compatible refinements, and some other rule could have been used to assign path lengths to the binary trees. There is no natural choice in either case; this fact reflects our ignorance about phylogeny and omitted variables (see §6*a*). Although it has some arbitrary elements, this model is much more satisfactory than a model with no similarity due to unrecognized phylogeny, and immensely more satisfactory than no model at all.

5. The criterion applied

The simulations reported here implement the simulation test for phylogenetic methods proposed in §4. The behaviour of neither the standard nor phylogenetic regressions is known analytically for this situation, and only the soundness of the general principles underlying the phylogenetic regression give confidence that it will perform well in this simulation test. Some other methods will also be subjected to the simulation test.

A minor purpose of this section is to show that the phylogenetic regression is unqualifiedly better than various rivals, but this much is almost obvious from its derivation. The major purpose is to investigate how well behaved the phylogenetic regression is in absolute terms. One practical problem is that if the phylogenetic regression is only slightly better than a rival, it may not be worth the extra effort to use the phylogenetic regression. It will be important to see by what margin the phylogenetic regression outdistances its simpler competitors. The difficulty of use depends on the software available for implementation; this aspect is discussed in §8. Technically, there are four reasons why the phylogenetic regression is not exact for the kind of data simulated in this section. These reasons will be discussed in §5*e* after the magnitude of the difficulties they cause has been assessed.

One superficially attractive type of simulation has not been performed, and that is to find the effect of mis-specification of the path segment lengths. For example, it would be possible to find the effect of using the path segment lengths derived by the method of figure 2 in the

analysis, but creating data by using some other method. The reason this has not been done is that the aim of the present paper is to provide the same kind of justification for the phylogenetic regression as already exists for multiple regression with non-comparative data. Misspecification of the variance–covariance matrix invalidates multiple regression in that case, just as it would invalidate the phylogenetic regression with comparative data. In both cases, these difficulties are an intrinsic part of the problem, not failures of the solution to the problem. There is therefore no point in trying to prove that mis-specification does not invalidate the phylogenetic regression, because it must. It is possible to investigate whether the phylogenetic regression is likely to be more invalidated by mis-specification than the multiple regressions which biologists perform routinely with other data, but this imprecise problem has not been tackled here. It is important to realize that every regression method for comparative data comes up against the problem of potential mis-specification, whether or not this is recognized by its practitioners.

(*a*) *The working phylogenies*

There are two working phylogenies used in the simulations. The first has 100 species divided into genera of ten species each, all belonging to one family. The second is illustrated in figure 9, and has 72 species and 23 higher nodes. The path segment lengths were generated by using the method of figure 2, and then transformed. Both working phylogenies were simulated with $\rho = 1$ and $\rho = 0.2$.

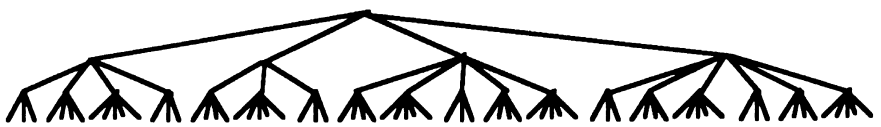

FIGURE 9. This figure shows the second working phylogeny used in the simulations. There are 72 species and 23 higher nodes.

(*b*) *The methods of analysis*

(i) The naïve species regression. Each species is treated as an independent data point with equal weight.

(ii) The standard regression, the maximum likelihood method described in §2(*b*) and §3, and defined in §10. The value of ρ was estimated by maximum likelihood in each trial, using the regression of y on X. Then that estimated value was taken as fixed to test for the addition of Z. The number of degrees of freedom in the denominator of the F-ratio was diminished by one on account of the estimation of ρ. The ill-conditioned nature of the estimate of ρ (discussed in §6*b*) and the primacy of hypothesis testing make it reasonable to take the first value of ρ as fixed when adding Z to the regression, rather than re-estimating ρ simultaneously with the estimation of γ. This method is also computationally less burdensome.

(iv) The phylogenetic regression, the method described in §3 and defined in §10. The estimated value of ρ from the standard regression of y on X was taken as fixed in the calculation of the phylogenetic regression. The number of degrees of freedom in the denominator of the F-ratio was diminished by one on account of the estimation of ρ.

(iv) The 'genus means' analysis. This method is applied only to the first working phylogeny. The method treats the genus means as independent data points of equal weight.

(v) The 'nested species within genus' analysis. This method is applied only to the first

working phylogeny. The method treats the deviations of species around genus means as independent data points of equal weight. It is implemented like the species regression, but each model includes a categorical variable which has a different level for each genus.

(*c*) *How the data are created*

The data generated are y, X and Z. Throughout the simulations, there are three variables being controlled for, so X is a matrix with three columns. There is one variable being tested for, so Z is a matrix with one column. The pseudo-random number generator seeds are set differently at the beginning of each series. GLIM's standard pseudo-random number generator was employed.

X is given values, which are constant through a series of trials. They are assigned values using the random number generator, and are normally distributed with zero mean and unit variance. X is therefore not phylogenetically distributed. In each trial, the working phylogeny is used to generate a binary compatible refinement. This binary phylogeny is then used to create Z and ϵ by using the path-segment method of figure 2. A standard normal random number is drawn for each path segment in the binary phylogeny, and multiplied by the square root of the path segment length, so that its variance equals the path segment length. Then the value for each species is derived as the sum of all the random variables on the path from the top of the tree to the species' node; y is then constructed by using

$$y = X\begin{bmatrix} 5 \\ -3 \\ 1 \end{bmatrix} + Z\gamma + \epsilon,$$

where γ is zero for series in which the null hypothesis is true, and 0.3 for series in which the null hypothesis is false. The crucial point is that the same binary phylogeny is used to create ϵ and Z, thus causing similarity between species due to phylogeny not represented in the working phylogeny.

Then the data set y, X and Z is subjected to the methods being tested. In each series there are 1000 trials. For each series, and for each type of regression, the results are presented as a table showing (i) the actual frequency with which the nominal 0.1, 0.05, 0.025, 0.01, 0.005 and 0.001 p-values were exceeded; and (ii) the root mean square error of the estimate of γ.

(*d*) *Results*

The counts in tables 2 and 3 can be assessed statistically as they have a binomial distribution. This is reasonably well approximated by a Poisson distribution, so that the standard error may be estimated as the square root of the observed value. Note that counts in the same columns are not independent as a count includes all counts below it in the table. The differences between successive counts are independent subject to the total of 1000 trials in each column. The results from different quarters of a table are independent, as they used different data. Each quarter of table 3 used only one set of data. Each quarter of table 2 used two sets of data, one for the standard and phylogenetic regressions and the other for the species, genus means and nested regressions. Much smaller differences are likely to be reliable when a comparison is made between products of the same data set than would be suggested by Poisson variability.

Simulations are reported in table 2 for the first working phylogeny, and in table 3 for the second. The left-hand side of each table shows results under the null hypothesis ($\gamma = 0$), and

TABLE 2. RESULTS OF SIMULATIONS WITH WORKING PHYLOGENY 1

(P is the P-value, and E is the expected number of times it is exceeded under the null hypothesis in 1000 trials. SP is the species regression, STD is the standard regression and PHY is the phylogenetic regression. GM is the analysis which treats the genus means as independent. ND is the nested analysis in which the species deviations from their (true) genus means are assumed to be independent. The top half of the table is for $\rho = 0.2$, the lower half for $\rho = 1$. The left-hand side is for $\gamma = 0$, the right for $\gamma = 0.3$. Within each quarter, the top six rows of numbers contain the actual number of times in 1000 trials the nominal P-values were exceeded. RMS (root mean square of deviations from the true value) is computed from the 1000 estimates of the value of γ. The relative power index (see text for details) of the phylogenetic regression is 80% for $\rho = 1$, and 77% for $\rho = 0.2$.)

P	E	SP	STD	PHY	GM	ND	E	SP	STD	PHY	GM	ND
$\rho = 1$												
0.1	100	748	362	112	238	348	100	785	761	483	327	735
0.05	50	705	280	47	158	259	50	750	699	326	234	674
0.025	25	658	214	20	97	194	25	713	645	204	151	625
0.01	10	608	149	7	54	121	10	679	584	108	83	556
0.005	5	567	118	3	36	90	5	642	541	60	51	502
0.001	1	506	61	0	7	41	1	593	423	11	15	397
RMS	—	0.517	0.184	0.975	0.711	0.183	—	0.509	0.183	1.042	0.692	0.192
$\rho = 0.2$												
0.1	100	329	148	93	129	112	100	802	874	740	216	843
0.05	50	239	79	39	69	61	50	746	795	583	148	757
0.025	25	186	40	15	34	30	25	693	724	424	77	663
0.01	10	139	19	3	11	12	10	596	613	256	41	545
0.005	5	96	11	0	5	4	5	529	532	148	21	458
0.001	1	62	2	0	1	0	1	418	367	36	5	297
RMS	—	0.173	0.112	1.335	0.549	0.113	—	0.174	0.113	1.640	0.561	0.118
				$\gamma = 0$						$\gamma = 0.3$		

TABLE 3. THE RESULTS OF SIMULATIONS USING WORKING PHYLOGENY 2

(Column headings, and content and arrangement as for table 2. The relative power index of the phylogenetic regression is 105% for $\rho = 1$, and 95% for $\rho = 0.2$.)

P	E	SP	STD	PHY	E	SP	STD	PHY
$\rho = 1$								
0.1	100	712	186	125	100	760	734	642
0.05	50	644	107	58	50	710	630	524
0.025	25	584	66	31	25	666	553	420
0.01	10	533	30	10	10	617	456	283
0.005	5	491	20	6	5	576	383	205
0.001	1	415	5	1	1	524	233	83
RMS		0.513	0.147	0.402		0.509	0.156	0.518
$\rho = 0.2$								
0.1	100	245	116	97	100	703	786	745
0.05	50	178	63	44	50	631	675	623
0.025	25	124	29	19	25	557	570	517
0.01	10	82	14	5	10	462	451	362
0.005	5	64	4	4	5	409	366	263
0.001	1	31	2	2	1	275	207	106
RMS		0.178	0.125	0.416		0.181	0.128	0.620
		$\gamma = 0$				$\gamma = 0.3$		

the right hand side the results under an alternative hypothesis ($\gamma = 0.3$). The data in the upper half of each table was created with $\rho = 1$, thus introducing strong phylogenetic effects. In the lower half, the data were created with $\rho = 0.2$, introducing much weaker phylogenetic effects.

The first important conclusion can be drawn from the left-hand ($\gamma = 0$) side of the tables. The phylogenetic regression is the only method that is approximately valid in every case. Every other method is seriously invalid somewhere; all are in the top left quarter of table 2. On the crucial grounds of validity the phylogenetic regression is wholly superior to the other methods employed here.

Its validity was investigated formally by fitting a logistic regression to the independent successive differences in the left-hand ($\gamma = 0$) side of the tables. The expected value of those differences is used as an offset in the model. The analysis shows that the results taken together differ significantly from validity ($\chi^2_{24} = 43.3$, $p < 0.01$). The first main effect is that the numbers of p-values falling below the 0.1 level in the four categories differ significantly from the expectation of 100 ($\chi^2_4 = 16.7$, $p < 0.01$). Inspection of the $P = 0.1$ row in the left-hand ($\gamma = 0$) side of the tables suggests that this is mainly caused by the values for high ρ being too high. The second main effect is that within the categories of p-value of 0.1 and below, there is a tendency for the more significant levels to be less numerous than expected, and for the less significant levels to be more numerous (a linear contrast taking consecutive integer values for the successive column differences gives $\chi^2_1 = 10.7$, $p < 0.01$), although the strength of this tendency does not seem to depend on ρ, or on the working phylogeny ($\chi^2_3 = 3.03$, $p > 0.25$). The residual variation ($\chi^2_{16} = 12.8$) is satisfactorily small and consistent with the binomial nature of the data. These significant deviations from validity are consistent with the approximate nature of the test; the probable reasons for the deviations are discussed in §5(*e*).

The strong statistical significance of the deviations should not be mistaken for strong magnitude of effect. The data set has a total count of 4000, so even small effects can be detected. One way to measure the magnitude of the deviations is comparison with the other methods. Another is to consider whether one's inference from an analysis would be much affected by knowing that a p-value that claimed to be 5% was in fact a p-value of 3.9% or 5.8%, which are the most extreme values the phylogenetic regression produces in the two tables. These values are probably well within the range of invalidity that would be produced in usual sorts of regression by mild relaxation of assumptions of normality, constant variance and independence of data points.

The invalidity of the other methods is not minor in scale. In the top left ($\gamma = 0$, $\rho = 1$) quarter of table 2, the species regression gives $p < 0.001$ over 50% of the time, and the standard regression gives $p < 0.001$ over 6% of the time. These are major discrepancies. It is tempting to think that if an effect is very highly significant, then however bad a method is, we can be sure that there is some evidence of an effect. These results show that this is far from being the case.

The advantage in validity of the phylogenetic regression is stronger in the top ($\rho = 1$) halves of the tables than in the bottom ($\rho = 0.2$) halves. This is to be expected, as higher values of ρ correspond to stronger effects of phylogeny in general, leading to an advantage over the species regression; and to stronger effects of unrecognized phylogeny as well, leading to an advantage over the standard regression, and the genus mean and nested methods.

The invalidity of other methods makes it difficult to assess the power of the phylogenetic regression. In order to get some measure of its power, I have constructed a 'relative power index' as follows. My aim was to compare the power of the phylogenetic and standard regressions at the 5% p-value. If the standard regression were valid, this would be done by comparing the entries in the right hand ($\gamma = 0.3$) side of each table corresponding to the 5% nominal p-value for the phylogenetic and standard regressions. As the standard regression is wildly invalid, this would be a nonsensical comparison. Instead I have tried to find the actual 5% p-value for the standard regression by fitting a curve between the number of times the standard regression exceeded the nominal p-value under the null hypothesis ($\gamma = 0$), and the number of times it exceeded the nominal p-value under the alternative hypothesis ($\gamma = 0.3$).

There are six data points for this curve from the top and bottom halves of each table, one point for each nominal p-value. Each point takes its y-value from the right-hand ($\gamma = 0.3$) side and its x-value from the left-hand ($\gamma = 0$) side. By interpolation on this curve, I can find the power corresponding to the actual 5% p-value. The curve was fitted by using GLIM. The y-axis was first transformed to be 1000 minus the entry in the right-hand ($\gamma = 0.3$) side of the table. A reciprocal link with a Poisson error was found to give a good fit. (The correlation of the errors and binomial nature of the data vitiate this as a statistical exercise, but the purpose is simple curve-fitting.) Each half-table's data was fitted separately. This yielded an estimate of the number of times the actual 5% p-value was exceeded by the standard regression for each value of ρ. Each kind of regression should exceed the 5% p-value 50 times by chance, so 50 is subtracted from the observed figure for the phylogenetic regression and from the estimated figure for the standard regression. These excesses measure the power of the methods. The relative power index is then calculated as the observed excess of the phylogenetic regression divided by the estimated excess of the standard regression, and expressed as a percentage. 100% means that the phylogenetic regression and standard regression are estimated to be equally powerful.

It is to be expected that the phylogenetic regression will be less powerful because it throws away information whose value it cannot assess. The standard regression uses this information and is invalid as a consequence. But by interpolating to the actual 5% p-value, we have assessed after the event how reliable that information was, and so the standard regression should have good power properties when treated in this way. (Of course this is only possible in a simulation with repeated data sets; in the ordinary situation of using the method, with a unique data set and in ignorance of whether the null hypothesis is true or not, it would be impossible to correct for invalidity in this way.) The relative power indices were 80%, 77%, 105% and 95% for the four half-tables. These indicate that the phylogenetic regression has remarkably good power properties. Even the lower figures of 80% and 77% are excellent, in view of the fact that the phylogenetic regression is the only valid method. It may be that the lower values for the first working phylogeny arise from the small number of degrees of freedom in the denominator of the F-ratio, though of course these values are much too few to claim this as a conclusion.

The species regression is particularly unpowerful when $\rho = 1$, as can be seen by comparing its right- and left-hand columns in the upper halves of tables 2 and 3. Whether the species regression attains significance is principally influenced by chance, and only a little influenced by whether there is a real effect to be detected.

The genus means method comes closest of the other methods to validity in the top left ($\gamma = 0$, $\rho = 1$) quarters of tables 2 and 3. This is done by discarding most of the information, as revealed in its very poor power shown in the top right ($\gamma = 0.3$, $\rho = 1$) quarters of tables 2 and 3.

Next, I discuss the estimates of γ. The tables show the root mean square deviation from the true value of γ, which is 0 in the left-hand side of each table and 0.3 in the right. All of the methods have unbiased estimates except for the phylogenetic regression under the alternative hypothesis. In each case, the estimate of the bias, the mean estimate over the 1000 trials, was less than twice its standard error, the root mean square divided by the square root of 1000. The phylogenetic regression was biased when $\gamma = 0.3$. In fact, the mean estimates of γ in table 2 were 0.823 ($\rho = 1$) and 1.527 ($\rho = 0.2$), and in table 3 were 0.397 ($\rho = 1$) and 0.529 ($\rho = 0.2$).

These correspond to biases of 0.523, 1.227, 0.097 and 0.229. The bias is most easily explained by reference to the short regression. The linear contrasts used to form the short regression contain squared terms in Z if the null hypothesis is false, as they are derived from the errors of a regression of y on X. When these squared terms meet Z itself in the calculation of the estimate of γ in the short regression, a cubic term in γ appears. This bias does not affect the hypothesis testing properties, which are fully captured by the behaviour of the p-values.

Among the other methods, the standard regression has the best estimates of γ. This is to be expected, as the standard regression is the maximum likelihood method for the case in which the true phylogeny belongs to the class generated by ρ-distortion of the working phylogeny. In table 2, the nested method has very similar root mean square errors. The nested method loses information compared with the standard regression because it ignores the genus means; on the other hand it gains because it does not have to estimate ρ. These two effects seem to cancel each other out roughly. The species regression is worse than these. As it uses the wrong variance–covariance matrix, it should be worse. And indeed, the difference between the standard and species regressions is much stronger when ρ is higher and so there is a greater discrepancy between their variance–covariance matrices. The genus mean method is worst of all, reflecting the fact that it throws away most of the information.

It is important to realize that these root mean square errors are measured in the repeated trials of the simulations, and are not the standard errors which the various regression methods would claim to have according to the standard formulae. Where the tables show a method to be invalid, this implies that the actual standard error will exceed the claimed error. It is therefore not true that by adopting the species regression, for example, one is simply obtaining a less precise estimate of γ. The precision of the estimate of γ claimed by the species regression will be hopelessly optimistic.

The conclusion from this comparison of root mean square errors is that if an estimate is desired, it is best to use the estimate of the standard regression. Of course, the standard error claimed by the standard regression cannot be trusted, and hypothesis testing must still be carried out with the phylogenetic regression.

(*e*) *Discussion*

The first point for discussion is why the phylogenetic regression is only approximately valid. What factors influence the approximation? There are at least four. It is most convenient to discuss these factors in terms of the short regression.

The F-distribution of the F-ratio from the short regression depends on the error's being normally distributed. This is true when the working phylogeny is the true phylogeny, as the sum of normals is normal. However, in each simulation trial, there is a random choice of binary phylogeny, and then normal errors are constructed for that phylogeny. The result of the double-stage randomization is that the error is a superposition of normals. It follows that the error is not itself normal. This cause of approximate validity is not a major cause for concern, as the assumption of normality is probably hardly ever met in applications anyway. However, it may cause a degree of invalidity in the test.

The second reason for invalidity concerns the averages of species values used in constructing the mean values for higher nodes, as illustrated in figure 4. These are weighted averages, and the independence of radiations which are nested depends on those weights being the right weights. They will be right if the variance–covariance matrix of the standard regression is

exactly right for the double stage randomization of the stimulation test. However, the 'true' variance–covariance matrix will not in general be contained in the family of ρ-distortions of the working phylogeny. This is because the true variance–covariance matrix is the average of the variance–covariance matrices of the random compatible refinements selected in the first random step of the simulation. The estimation of ρ will seek out the closest member of the family, and it seems likely that the closest member will be fairly close, but there is no reason to expect this average to be a member of the family of ρ-distortions. This means that the standard regression contains a mis-specification for the data-creating process of the simulation. This is an additional reason for inexactitude of the phylogenetic regression.

The third reason is involved, but is connected to the most crucial aspect of the superiority of the phylogenetic over the standard regression. In both simulations, Z as well as ϵ was randomly created. The reason is that to create similarity due to unrecognized phylogeny between Z and y it was not possible to have Z fixed in all the trials of a simulation. The similarity due to unrecognized phylogeny had to be represented by having Z as a random variable. If the working phylogeny was the true phylogeny, this would be irrelevant. The distribution of ϵ would be independent of the distribution of Z, and so the normal theory would apply. In the simulations, on the other hand, the distribution of ϵ is not independent of the distribution of Z. If two of the ten species in an unsubdivided genus have very different values of Z, then it is likely they are distant in the current trial's binary phylogeny, and so the values of ϵ are likely to be different too. If two species have very similar values of Z, it is likely that the current trial's binary phylogeny has placed them close together, and so the values of ϵ will be similar too. This problem works through to the short regression, creating non-independence between the y-variable and the test variable under the null hypothesis.

A fourth, general, reason is that the method makes only an approximate allowance for the effect of the uncertainty in the estimate of ρ namely, the reduction by one in the denominator degrees of freedom. I am grateful to Professor P. Armitage for pointing this out to me. In each trial of the simulation, ρ is re-estimated, and so the simulations do take this into account.

These four causes are responsible to unknown degrees for deviations from strict validity of the phylogenetic regression. None is easily fixed, or naturally avoidable. The approach I have taken is to assess how well the method works with these imperfections. At least in the examples studied, no serious damage is done.

Additional simulations not reported here were conducted in which the working phylogeny was taken as the true phylogeny, and ρ was taken as fixed and known. These confirmed the analytical results of §10 that the standard and phylogenetic regressions are exact tests in this situation.

Conclusions can also be drawn about the other methods. Unrecognized phylogeny is very important. The standard regression, which is a maximum likelihood method and therefore has various optimality properties when there is no unrecognized phylogeny, performs very poorly in the simulations. Moving to higher-level means, to eliminate undeserved multiplicity of data points, helps against recognized but not against unrecognized phylogeny. This move, represented in the simulations by the genus means method, is the basic philosophy of the 'higher taxonomic level methods' reviewed by Pagel & Harvey (1989). Nested analyses, represented in the simulations by the nested species within genus method, have the same pattern of success and failure. The nesting technique is the basis of the 'nested analysis of covariance methods' reviewed by Pagel & Harvey (1989).

Many more simulations could have been performed to explore the behaviour of the

phylogenetic regression. One minor reason why more are not shown is that the simulations are computationally onerous. The main reason, however, is that no other method I know of provides serious competition for the phylogenetic regression. It is clear from its derivation that it is likely to be superior to any method that does not cope with unrecognized phylogeny. If another serious method were proposed, then more substantial simulations would be necessary to decide between the rivals.

The chief reason to trust the phylogenetic regression is the combination of the analytical results of § 10, showing that it is exact for a known phylogeny, and its application of the radiation principle. The simulations are important extra support, but are not the mainstay of the case for the phylogenetic regression.

(*f*) *Conclusions on the basis of the simulations*

The conclusions I draw from the simulations are that the phylogenetic regression is by far the best of the methods tested, and that it is good enough to be worth using. Section 8 considers how easy it is to apply the method.

6. The arbitrary choice of path-segment lengths

This section discusses what attitude should be taken on biological grounds to the arbitrary choice that needs to made in the phylogenetic regression about the path-segment lengths in the phylogenetic tree. Section 6 (*a*) develops a view of what is meant by error in a comparative study, and draws conclusions contrasting with those of Felsenstein (1985, 1988) about what assumptions it is reasonable to make about it; § 6 (*b*) discusses degrees of flexibility that are and might be allowed in path-segment lengths in the phylogenetic regression. Finally, § 6 (*c*) sums up on this arbitrary choice, arguing that it is an unavoidable feature of any method that does the same job as the phylogenetic regression, and that a strictly analogous arbitrary choice is made in all regressions and analyses of variance on non-experimental data.

(*a*) *A model of the error*

Various reasons have been suggested for why phylogenetically closer species should be more similar, such as 'phylogenetic inertia' (Wilson 1975). This complex nexus of ideas will not be reviewed here. Instead I will suggest what seems to me to be the natural explanation of why phylogenetically close species tend to be similar. It is so natural that it cannot be original with me, but I do not know where it has been suggested before.

Suppose that speciation and extinction are driven by niche creation and destruction. When a niche is created, it is likely to be filled by speciation from a species in a similar niche. Such a species can survive better initially in the vacant niche, and evolve sooner to exploit it fully, than a species in a more distant niche. This implies that similar niches will tend to be occupied by phylogenetically close species. The reason for the similarity of the two species after speciation, and full adaptation to the new niche, is that the niches are very similar.

On this view, species are perfectly adapted to their current niche, *while at the same time* phylogenetically closer species are more similar. No mysterious forces are invoked. It implies that phylogenetically close species should be similar in all sorts of ways, many more than our statistical analysis will deal with at once. We usually study only a few aspects of a niche at a time.

The important implication of this view is that the cross-species pattern in one character can be

understood as an adaptive response to the niche and to other characters. It is not necessary to know about ancestral states, or about how fast evolution of the character has occurred. The current state of a species and its niche is sufficient to explain its value of a given character. To be sure, the distribution of species in the multi-dimensional space of all characters has a phylogenetic pattern which does require historical elements in its explanation, and this explanation is a worthwhile exercise. But it does not interfere with an ahistorical approach to the explanation of the distribution of one set of characters conditional on the distributions of other characters and on properties of niches.

There are other supposed problems for comparative methods, such as the existence of multiple peaks, and characters being modified in different directions in different taxa by the same selective force. These general philosophical problems are relevant to Ridley's method and mine in exactly the same way, and the stalwart defence of his comparative method (Ridley 1983, pp. 28–34) applies equally to mine.

The adaptive explanation of phylogenetic similarity suggests the following model for a given character. Suppose that if the full truth were known about a character y, it would turn out that the following relationship existed between y and a number of x-variables, x_1 to x_{100}:

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + \ldots + x_{100}\beta_{100}. \quad (1)$$

When a biologist studies y, theories will suggest that some variables are important, say x_1 to x_{10}, but it is likely that only a subset of these have been measured, say x_1 to x_3. The statistical model that the biologist would fit in a linear regression would therefore be

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \epsilon,$$

where ϵ represents the error from the biologist's model. But with our advantage of knowing the truth, we know that ϵ is, by subtraction,

$$\epsilon = x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + x_8\beta_8 \ldots + x_{100}\beta_{100}.$$

This formula is useful when we wish to know what assumptions it is reasonable to make about the errors in the regression model. The error is equal to a weighted sum of characters. If most characters are phylogenetically distributed, then so will the error be in the regression. This conclusion is independent of the reasons why characters are phylogenetically distributed. Thus the model is suggested by the adaptive account of phylogenetic similarities, but does not logically depend on it.

This view of the error as arising from relevant omitted variables is not specially applicable to comparative data, but is probably appropriate more widely. An alternative view is that the error arises from measurement error, which is not a character that a biologist wishes to investigate, and will not show regularities from study to study. From the statistical point of view, error includes all reasons why data points do not lie exactly on the fitted line, and so omitted variables and measurement error will both contribute to error (as will mis-specification of the deterministic part of the model). In ignoring measurement error in what follows, I am exploring the possibility that measurement error is small compared to the effects of omitted variables. It is interesting to note that the identities and methods of research workers, which must introduce error (in the extended sense) into comparative data, are also likely to be phylogenetically distributed.

The method used to represent the correlations of the errors in the phylogenetic regression is

the path segment length method of Felsenstein (1985), which he based on a Brownian motion model. On a tree, any set of lengths will give phylogeny-like similarities, so how are those lengths to be chosen? This is the central problem of this section. Felsenstein (1985) suggested that dates of the splits should be used, and so lengths on the tree would be proportional to the intervals between splits. The error in the character is supposed to be evolving by random drift. Accordingly, Felsenstein also suggested that if evidence on differential speeds of evolution was available, for example from pseudogenes, then that should also be incorporated into the tree.

The approach I shall take is different. The variances and covariances are wanted to represent the error in a statistical analysis. The view of the error espoused here is that it results from important but omitted variables. There is therefore no reason to expect the rate of random drift to be relevant. On Felsenstein's view, the same lengths should be used for every character analysed in the same set of species. On my view, it is perfectly possible that different characters should have different sets of lengths. It is also possible (indeed almost necessary) that as one character is better understood, and more relevant characters are included as x-variables, the appropriate lengths should change.

On my view, then, the arbitrariness of the path segment lengths is an intrinsic part of the logical position in which we find ourselves in drawing inferences from phylogenetic data, and would not be removed even if we knew all the dates of splits and rates of neutral gene evolution between all the splits. Section 6*c* discusses what attitude should be taken to this arbitrariness.

One main way in which sets of path-segment lengths can vary is the extent to which variation is placed close to the species level, making species more or less independent, or placed close to the top of the tree, making the higher taxonomic levels quite distinct from each other. This decision is very important, and it would be unsatisfactory to have to choose one fixed set of lengths. The family of trees used in the phylogenetic regression, parametrized by ρ, allows flexibility in this direction.

(*b*) *Fitting parameters of the tree*

The precise pattern of variances and covariances between species is beyond our reach. There are more path-segment lengths than there are species, so it is impossible to estimate the lengths from the data. The best that can be done is to provide a family of trees which provides flexibility in the most important directions, and let the data choose among the members of the family. The high principle of the randomization test has trickled away into pleas to reasonableness now that weights are being discussed. The phylogenetic regression is just like most other statistical methods for non-experimental data in this respect.

Choice among members of the family is performed by maximum likelihood. The relevant formula is

$$\text{lik} = \frac{\exp(-\frac{1}{2}(y-X\beta)^T V^{-1}(y-X\beta))}{\sqrt{(\det(2\pi\sigma^2 V))}},$$

where 'lik' means likelihood, and 'det' means determinant. This formulation subsumes ρ into V; β and the parameters of V are fitted simultaneously by maximizing the likelihood in the standard regression.

The parametrization of the variances and covariances is now discussed. There are two parameters, σ^2 and ρ. σ^2 as usual simply scales all the variances and covariances in the tree. ρ distorts the tree in the way described in §2*a*. To recap, V is defined by

$$V_{ij} = (1-h_{ij}^{\rho}),$$

where h_{ij} is the height in the working phylogeny at which the paths from the root to i and j diverge.

In an example the adequacy of the parametrization of V can be checked against any stated alternative. The residuals in the short regression should be normally distributed with equal variance. Each residual corresponds to a higher node in the phylogeny. The absolute value of the residuals could be plotted against height of the corresponding node, to see if the stretching imposed by ρ has approximately the right relative strength at different heights in the tree. The residuals could also be inspected for phylogenetic patterns; it might turn out that some taxa have larger residuals than others, suggesting that the path-segment lengths might with advantage be increased in one portion of the tree. Of course overinterpretation of residuals is to be avoided, and only strong patterns are likely to be worth following up.

It is possible, though I have come across no real examples (in an admittedly small sample), for the fitted value of ρ to be zero, so that path segments with a species at their lower end have a length of one, and all other path segments have zero length. The variability imputed to the deviation of daughter means from their parent node's mean would not then be zero because sampling variation is passed up the tree. So the fitted value of ρ will be zero when the variation found between higher nodes is equal to or less than what would be expected from sampling variation alone. Equality would be consistent with the absence of phylogenetic effects, and the independence of species. It is imaginable that there could be less variability than sampling would lead us to expect. This could arise in body size if different genera within a family occupied different kinds of niche, defined by diet type, and each genus contained a whole range of species of different sizes. Then each genus could have approximately the same mean but contain a great deal of variation. This hypothetical possibility would cause some problems for the method. One simple *ad hoc* solution would be to omit the sampling variation from the variability imputed by the method to the means of higher nodes. I stress that no examples are known to me of this situation. It is the kind of thing that careless over-parametrization of V could easily bring about.

The estimate of ρ has very poor statistical properties. I believe that it is asymptotically biased and that its sampling variance does not decrease to zero asymptotically. The reason for bias is that the estimate of ρ fulfils the extra function of representing in the working phylogeny the positive correlations introduced between daughters by the random choice of compatible refinement. This force will place more weight higher in the tree, and so the estimate of ρ will be higher than it should be. The question of sampling variance depends on how sample size is increased 'asymptotically', and in particular on whether the number of daughters per higher node goes to infinity or not. If it does, then it seems likely that the sampling variance of ρ will indeed go to zero. But this is an irrelevant way of increasing sample size asymptotically. It is much more reasonable to assume that the number of daughters per higher node remains finite. In this case there will always be a finite weight attached to the finite number of daughters of the top node, if the method of figure 2 is used to create path-segment lengths. This will maintain a finite sampling variance for ρ even as the number of species increases indefinitely.

The poor statistical properties of the estimate of ρ are not of direct concern, as ρ is essentially a nuisance parameter. But it is important to realize that caution should be taken if the estimate of ρ is interpreted in any way. The poor behaviour of the estimate may prevent the standard and phylogenetic regressions, when ρ has to be estimated, from being efficient for the parameters which are of interest. However, the uncertainty of the form of the tree is part of the nature of the problem, and is not introduced by the method of analysis.

The parametrization of V is an area of the phylogenetic regression which experience may alter. The different ways of forming the basic tree, and whether sampling errors should be passed back up the tree, are all matters with no truly principled solution. These problems are not, however, special difficulties with the phylogenetic regression; they reflect general difficulties with making inferences from comparative data.

(*c*) *Summing up*

The arbitrary choice of path-segment lengths represents a limitation on the certainty of the conclusions that can be drawn from the phylogenetic regression. The purpose of this section is to argue that this limitation is imposed by fundamental uncertainties in drawing conclusions from comparative data, and is not unnecessarily introduced by the phylogenetic regression. The limitations can be avoided only for statistical analyses of a very special kind.

In choosing path-segment lengths we are choosing how to weigh up evidence from different parts of the phylogenetic tree. Is the positive relationship between y and Z found within a genus in one part of the tree stronger or weaker evidence than the positive relationship found within a family in some other part? Are they, taken together, enough to counterbalance the negative relationship found within the whole order? On the view of the error expounded in §6*a*, these questions are strictly unanswerable. We would need to know about the omitted but relevant variables.

But these questions have a strong parallel in regressions that biologists do routinely. When performing a regression on non-experimental data, it is necessary to choose how to weight the data points against each other. This choice is expressed in the choice of weights in a weighted regression. In the absence of a way to choose these weights, it is usual to perform an unweighted regression without any consciousness of having made an arbitrary choice. The logical position of choosing path-segment lengths is the same as that of choosing weights. Most statistical methods for non-experimental data rely unavoidably on underlying assumptions about the covariance structure, which have to be made arbitrarily.

The pragmatic approach is to choose a 'reasonable' set of weights, and perform the regression. Someone who challenges the weights can always repeat the analysis with another set of weights to see if it makes a difference to the results. If two 'reasonable' sets of weights give very different results, then it is genuinely doubtful what inferences should be drawn from the data. This pragmatic approach is essential to very many analyses in many areas of application of statistics, and is adopted here for the phylogenetic regression.

This contrasts with Felsenstein's (1985, 1988) counsel of perfection (and despair!) that the tree must be known before sensible statistics can be done on comparative data. The time intervals between splits are needed, he argues, before the covariance structure can be known. The pragmatic approach suggests that, even with an unknown covariance structure, it is reasonable to do statistics. More fundamentally, the view of error developed in §6*a*, based on relevant but omitted variables, suggests that the time intervals between splits do not in principle supply the true covariance structure. (More fundamentally still, I doubt that there is a true covariance structure except as a formal convenience in the mind of the theoretical statistician. It is part of mathematical technique, not a truth about the world.)

Some kinds of analysis can be done that avoid the arbitrary choice, but they must omit data. A data point cannot be used twice, once in its own right, and again as part of an average. For example, once the species within a genus had been used, the genus mean could not then be used for a higher-level contrast. An ingenious method of obtaining more non-overlapping contrasts

than might be thought possible is described by Felsenstein (1988, p. 457). Over and above this, there are strong restrictions on the kinds of test that can be performed on these contrasts without relying on weighting the evidence: for example, they must at least be non-parametric. A rank correlation will not be valid, because it will not be true that the contrasts on one variable are independent of the contrasts on the other. More distant contrasts will have a more dispersed distribution for all variables. Tests will have to be based on just the sign of the difference in a variable, leading to a sign test or 2^n contingency table test.

The conclusions, then, are that there are uncertainties involved in drawing inferences from comparative data which no general statistical method can avoid. The arbitrary choices of path-segment length reflect these uncertainties, and not additional uncertainties introduced by the techniques employed in the phylogenetic regression. Biologists are, rightly, going to continue to look at comparative data sets and draw inferences in the absence of knowledge of true phylogenies. They deserve a statistical method that helps them as fully as possible in the rational assessment of that evidence.

7. Discussion

The non-historical nature of the logic underlying the phylogenetic regression is discussed in §7*a*. The advantages of the phylogenetic regression are summarized in §7*b*.

(*a*) *Reconstruction of ancestors or conditioning on pattern?*

The logic of the phylogenetic regression has no truck with history. Ancestral states are not inferred, and fossils would not take priority in directing inferences. That the reconstruction of ancestral states is not merely hidden in the algebra of the phylogenetic regression may be seen from the following example. Consider the phylogeny and evolution of a character depicted in figure 10. The original species divides into two, then into four, then into eight. During recent times, the eight species have all changed dramatically in the value of the character, so that all eight species have values outside the range of all earlier forms. Now imagine a biologist who does not know the history of the group applying a statistical analysis to the modern species. If

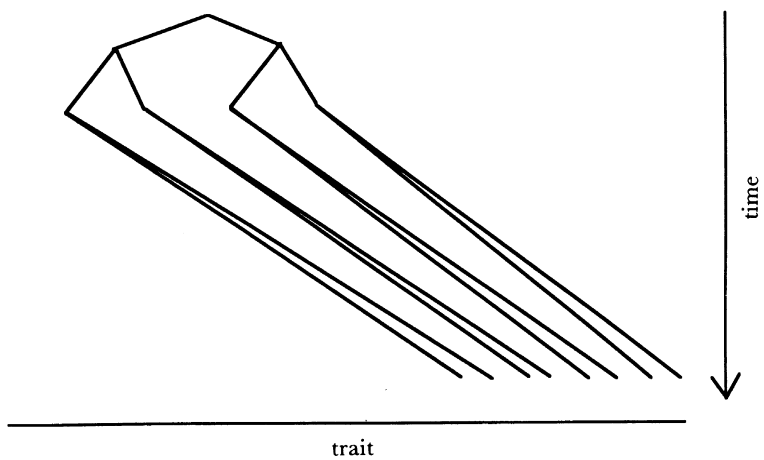

Figure 10. This figure illustrates a trait that changes in time so dramatically that extant species lie entirely outside the range of their common ancestors, yet retains a phylogenetic pattern in which more closely related species are more similar to each other than distantly related species.

the logic of the method requires the reconstruction of ancestral states, then we, with our extra information, know that the biologist's method will fail and will not believe his conclusions. My case is that the logic of the phylogenetic regression is still perfectly sound in such a situation, and our extra information would not lead us to doubt the biologist's conclusions. Or, putting ourselves in the position of the biologist, we do not rely on the assumption that ancestral states are averages of extant species' states.

How can this case be justified? The answer lies in my model of a character, expression (1) from §6*a*, which in this case we can write with only four x-variables as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4.$$

The logic of the method derives from the supposition that the omitted but relevant characters are phylogenetically distributed, not on any reconstruction of ancestral states. In particular, suppose that x_4 in the above formula has changed in all members of a group, perhaps because of some environmental change. The y-values of all species in the group will be changed, dramatically if β_4 is large. This does not alter the relationship between x_1, x_2, x_3 and y. The extent of this change due to β_4 does not alter how these relationships should be studied statistically; y will still be distributed phylogenetically if x_1, x_2, x_3 and x_4 are.

It will be seen that the justification for the phylogenetic regression then rests on the distribution among modern species of all the characters that determine y. Provided they are distributed phylogenetically, then the method is appropriate. Nothing need be assumed, or is tacitly asserted, about ancestral states. The change in x_4 may even have moved all modern species' y values outside the range of all ancestral species, as illustrated in figure 10.

(*b*) *Conclusions*

The phylogenetic regression has passed its tests well. It applies to comparative data with a continuous y-variable, and has the following satisfactory features.

(i) It uses all the data.

(ii) It provides the hypothesis-testing facilities of multiple regression, treating categorical variables and continuous variables, and allowing arbitrary numbers of variables to be controlled for and tested for. It handles interactions routinely.

(iii) It works with an arbitrary phylogeny. Further, it behaves correctly with a binary phylogeny, when it is equivalent to the standard regression; and with a null phylogeny (all species are daughters of the same higher node) when there are no degrees of freedom in the test.

(iv) It has been shown to be (approximately) valid, analytically and by simulation, in the face of similarity caused by both recognized and unrecognized phylogeny.

With these advantages, the phylogenetic regression will be the only available and acceptable method for many problems, and the best of the available methods for many more. The phylogenetic regression is the natural hypothesis-testing regression method for comparative data.

8. How the phylogenetic regression can be implemented

I have written a GLIM program that implements the phylogenetic regression, which is now in widespread use in the Zoology Department of Oxford University. Copies of the program are available free from me. The program is small (approximately 30 K) and requests should be sent in writing and accompanied by a formatted Macintosh floppy disc. GLIM (Generalized Linear

Interactive Modelling system) is available from NAG Ltd, Wilkinson House, Jordan Hill Road, Oxford OX2 8DR, U.K. GLIM runs on many micros (including Macintosh and IBM PC), and on most mainframes. The user interface is virtually uniform, and programs written on one machine run without alteration on others. The program is accompanied by documentation.

An analysis with 56 species and 42 higher nodes in the working phylogeny, in which one continuous variable is controlled for and one is tested for, takes under seven seconds of CPU time on a VAX 8700. The program's output includes the F-ratio of the test, parameter estimates from the standard regression, and a plot which shows the influence of individual radiations in the analysis. Continuous and categorical variables can be controlled for and tested for; ρ is automatically fitted by maximum likelihood. When complications arise from the loss of degrees of freedom described in §3*e*, they are correctly dealt with.

I am most grateful to Mark Ridley for his example and help. He has made useful comments on a number of previous drafts of this paper. Michael Bulmer made useful suggestions about the simulations and insisted that I include the estimation of ρ in the phylogenetic regression. He also listened patiently to informal and consequently incomprehensible accounts of the results that can now be found in §10 and suggested major improvements to the organization of the paper. Olof Leimar pointed out a notational error in the Appendix. Paul Harvey, Tim Guilford and Austin Burt commented on earlier versions of the paper. Mark Ridley, Sue Healy and Tim Guilford allowed me to practice using my GLIM program on their data sets. Mark Pagel and Paul Harvey allowed me sight of a manuscript of their paper which is currently in the press. Oxford University Computing Service provided computing facilities. I am supported by a Royal Society 1983 University Research Fellowship.

9. References

Cox, D. R. & Hinkley, D. V. 1974 *Theoretical statistics*. London: Chapman and Hall.

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

Felsenstein, J. 1988 Phylogenies and quantitative characters. *A. Rev. Ecol. Syst.* **19**, 445–471.

Johnston, J. 1972 *Econometric methods* (2nd edn). Tokyo: McGraw-Hill Kogakucha.

Krebs, J. R., Sherry, D. F., Healy, S. D., Perry, V. H. & Vaccarino, A. L. 1989 Hippocampal specialization of food-storing birds. *Proc. natn. Acad. Sci. U.S.A.* **86**, 1388–1392.

Pagel, M. D. & Harvey, P. H. 1989 Recent developments in the analysis of comparative data. *Q. Rev. Biol.* **63**, 413–440.

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.

Ridley, M. 1986 The number of males in a primate troop. *Anim. Behav.* **34**, 1848–1858.

Ridley, M. 1989*a* The timing and frequency of mating in insects. *Anim. Behav.* (In the press.)

Ridley, M. 1989*b* The incidence of sperm displacement in insects: four conjectures, one refutation. *Biol. J. Linn. Soc.* (In the press.)

Sessions, S. K. & Larson, A. 1987 Developmental correlates of genome size in plethodontid salamanders and their implications for genome evolution. *Evolution* **41**, 1239–1251.

Sibley, C. G., Ahlquist, J. E. & Monroe, B. L. 1988 A classification of the living birds of the world based on DNA–DNA hybridization studies. *Auk* **105**, 409–423.

Sillén-Tullberg, B. 1988 Evolution of gregariousness in aposematic butterfly larvae: a phylogenetic analysis. *Evolution* **42**, 293–305.

Wilson, E. O. 1975 *Sociobiology*. Massachusetts: Belknap Press.

10. Appendix

(*a*) *Preliminary remarks*

The purpose of this appendix is to state results of importance to the phylogenetic regression. The mathematics done is all quite simple, but to express it economically it has been necessary to adopt a rather formal approach. A major notational problem is the formal treatment of an arbitrary phylogeny. In these preliminary remarks, the meaning and relevance of the four theorems is discussed. Throughout the Appendix, formal remarks are made to explain the direction of the developing argument. In many cases the actual objects of interest are not mentioned in the mathematics at all. These objects are statistical tests. The first is the *standard regression*. The variables involved are a y-variable y, a set of x-variables X to be controlled for, and a set of x-variables Z to be tested for. The regression is defined by

$$\boldsymbol{E}(y) = \mathbf{1}_t u + X\beta + Z\gamma \quad (y - X\beta - Z\gamma) \propto \boldsymbol{N}(0, V),$$

where $\mathbf{1}_t$ is the constant term, V is defined by

$$V_{ij}(\rho) = (1 - h_{ij}^{\rho}),$$

and h_{ij} is the height in the initial working phylogeny at which the paths to species i and j diverge. The symbol $\propto$ is used to mean that the variance–covariance matrix of the error is assumed only to be proportional to V, not necessarily equal to it. For the purpose of this appendix the path-segment lengths are fixed, so that ρ is considered known.

The first theorem states that the standard regression is equivalent to the *long regression*, defined by

$$\boldsymbol{E}(Ly) = S\delta + LX\beta + LZ\gamma \quad (Ly - S\delta - LX\beta - LZ\gamma) \propto \boldsymbol{N}(0, C),$$

in which L, S and C are matrices defined formally later. The two regressions are shown to be equivalent in the sense that the residual sum of squares of the long regression, concentrated for β and δ, is the same function of y, Z and γ as the residual sum of squares of the standard regression concentrated for μ and β. This shows that the significance tests for $\gamma = 0$, controlling for $\mathbf{1}_t$ and X in the case of the standard regression and for LX and S in the case of the long regression, will yield the same test statistic with the same distribution. Each data point in the long regression represents the deviation of a node's value from its parent node's value. The data in this form is suitable for defining the randomization test explained in §3*c*. It is important that C is a diagonal matrix, so that this theorem allows the standard regression to be fitted by a package that cannot handle non-diagonal variance–covariance matrices. GLIM is such a package. The reason it is necessary to prove this first theorem is to show that the formulae for L and C are correct; their forms are far from obvious *a priori*. L represents the process of 'hanging on the tree' described in §3*a*.

The second and third theorems concern the *short regression*, defined by

$$\boldsymbol{E}(GC^{-1}Ly) = GC^{-1}LX\beta + GC^{-1}LZ\gamma \quad (GC^{-1}L(y - X\beta - Z\gamma)) \propto \boldsymbol{N}(0, \boldsymbol{I}).$$

The distribution of $(GC^{-1}L(y - X\beta - Z\gamma))$ is understood as a distribution conditional on G, as G is a random matrix because it depends on the value of y. The second theorem states that the process of performing the long regression, defining the random linear contrasts GC^{-1} and forming the elements of the short regression does indeed result in the same, standard, statistical test as the short regression. This is shown by proving that conditional on G, the residual in the

short regression after regression of $GC^{-1}Ly$ on $GC^{-1}LX$ has the same probability density whether the randomness arises through ϵ, the error in the standard regression, as transmuted by construction of G and the formation of the short regression; or whether the randomness is assumed to arise as a $\boldsymbol{N}(0, \boldsymbol{I})$ variable in the short regression itself. The first reason it is necessary to prove this theorem is to show that the formula for G is correct. The second reason is that G is a random matrix, as it depends on ϵ. In general, using contrasts G that depend on ϵ will violate the standard formulae for the variances and covariances of the contrasts, which rely on fixed G. As was seen in the simulations in §5, the short regression has high mean square error in its parameter estimate under the null hypothesis, and has biased estimates under the alternative hypothesis. It is therefore not at all obvious that the short regression will be valid but, as the theorem shows, it is.

The third theorem states that the short regression is equivalent to the *long regression with* T, defined by

$$\boldsymbol{E}(Ly) = S\delta + LX\beta + T\tau + LZ\gamma \quad (Ly - S\delta - LX\beta - T\theta - LZ\gamma) \propto \boldsymbol{N}(0, C).$$

T is a matrix representing a set of artificial variables added to the long regression to ensure that no matter what value Z may take, the residuals after regression on S, LX, T and LZ will remain proportional, within each radiation separately, to the residuals after regression on S and LX alone. T therefore depends on y, and like G is a random matrix. Equivalence means that the residual sum of squares for the long regression with T, concentrated for β, δ and θ, is the same function of y, Z and γ as the residual sum of squares of the short regression concentrated for β. The theorem is proved to show that the phylogenetic regression can be interpreted as conditioning within the standard regression on the patterns of the residual in each radiation, in the sense of 'pattern' explained in §3*c*.

The fourth theorem shows that the randomization test, explained in §3*c* and defined formally below, is equivalent to the short regression in the sense that the null distribution of the test statistic of the randomization test is also an F-distribution with the required degrees of freedom.

As well as these four results, the mathematical development defines the matrices used to construct the long and short regressions, and so formally defines the phylogenetic regression.

The proofs of the theorems have been omitted, but have been lodged in the archives of the Royal Society.

(*b*) *Definitions and theorems*

A preliminary note on matrix notation. I shall define matrices as $A \times B$, where A and B are finite sets, rather than as $m \times n$, where m and n are integers. An $A \times B$ matrix D will have elements D_{ab}, where $a \in A$ and $b \in B$. Where a matrix is defined as $m \times n$ or $A \times n$, the integers m and n should be understood as shorthand for the sets $\{1, 2 \ldots m\}$ and $\{1, 2 \ldots n\}$. The advantage of this notation is that if A' and B' are subsets of A and B, respectively, then a submatrix D' can be concisely defined as the $A' \times B'$ submatrix of D.

Definition of Π, Π_{h}, Π_{s}, Π_{t}, Π_{i}, Π_{di}. These definitions are made with respect to the working phylogeny. Let Π be the set of all nodes, Π_{t} the set of species nodes, Π_{h} the set of higher (i.e. non-species) nodes and Π_{s} the set of all nodes except the root. Let Π_{i}, $\mathrm{i} \in \Pi$, be the set of species nodes which are descendants of (or equal to) node i. Let Π_{di}, $\mathrm{i} \in \Pi_{\mathrm{h}}$, be the set of daughter nodes of node i. Associate each node with a distinct integer, to establish an arbitrary ordering over Π.

Definition of P, P_i. Let P denote the partition $\{\Pi_{di}\}_{i\in\Pi_h}$ of Π_s, and let P_i, $i\in\Pi$, denote the partition $\{\Pi_j\}_{j\in\Pi_{di}}$ of Π_i.

Definition of n, n_t, n_s, n_h. Let n be the number of nodes in the working phylogeny, n_t be the number of species nodes, n_s be $n-1$, and n_h be the number of higher nodes. Note that every species is either a species node or a higher node but not both, so that $n_h+n_t=n$. It follows that $n_s-n_h=n_t-1$.

Definition of $'$. Let $i'\in\Pi_h$ denote the parent node of i, $i\in\Pi_s$.

Definition of κ_i, h_i. Let κ_i be arbitrary non-negative real numbers representing the length of the path segment between i and i', $i\in\Pi_s$, with $\kappa_i>0$ if $i\in\Pi_t$. Let h_i be the summed length of the path segments between the root and i, $i\in\Pi$.

Definition of $\alpha(i,j)$. Let $\alpha(i,j)$ be the lowest common ancestor of i and j, $i,j\in\Pi$.

Remark. The h_i just defined are related to the h_{ij} used in the body of the paper by the relationship $h_{\alpha(i,j)}=1-h^{\rho}_{ij}$. The working phylogeny as used in the Appendix is taken as having already undergone transformation by ρ.

Definition of Ω_t, Ω_s, Ω_{ti}, Ω_{si}. Let Ω_t be the set of column vectors with real elements indexed by Π_t, and let Ω_{ti}, $i\in\Pi$, be the subspace of Ω_t with only those elements indexed by Π_i. Let Ω_s be the set of column vectors with real elements indexed by Π_s, and let Ω_{si}, $i\in\Pi_h$, be the subspace of Ω_s with only those elements indexed by Π_{di}.

Remark. The definitions of the various Π allow means at higher nodes to be dealt with in the same way as species values. Ω_t and Ω_s are the data spaces of the standard and long regressions, respectively.

Definition of $\mathbf{1}_{ti}$, I_{ti}, $\mathbf{1}_{si}$. Let $\mathbf{1}_{ti}\in\Omega_{ti}$, $i\in\Pi$, be the vector each of whose elements equals one. Let I_{ti} be the identity matrix over Ω_{ti}. Let $\mathbf{1}_{si}\in\Omega_{si}$ be the vector each of whose elements equals one.

Definition of U_i. Let U_i, $i\in\Pi$, be the $\Pi_i\times\Pi_t$ matrix defined by

$$(U_i)_{jk}=\begin{matrix}1 & j=k\\ 0 & j\neq k\end{matrix}$$

Remark. U_i is a matrix which picks out from a vector $x\in\Omega_t$ those elements indexed by elements of Π_i. $(U_i x)\in\Omega_i$, and equals x over those elements held in common. U_i^T transforms a vector $x\in\Omega_i$ into a vector which is an element of Ω_t, equals x in those elements indexed in common, and equals zero elsewhere.

Definition of V. Let V be the $\Pi_t\times\Pi_t$ matrix defined by

$$V_{ij}=h_{\alpha(i,j)},\quad \text{for}\quad i,j\in\Pi_t.$$

Extension of subscript notation for V. As an extension of the usual subscript notation, let V_{ij} also be defined when i and j are not necessarily species nodes, as the $\Pi_i\times\Pi_j$ submatrix of V. Further, let V_i denote V_{ii}.

Definition of σ_i^2. Let $\sigma_i^2=(\mathbf{1}_i^T V_i^{-1}\mathbf{1}_i)^{-1}$, $i\in\Pi$.

Remark. σ_i^2 is the sampling variance of the mean of all the species below node i, of a variable whose variance–covariance matrix is V.

Definition of f_i. Let $f_i \in \Omega_t$, $i \in \Pi$, be defined by $f_i = U_i^T V_i^{-1} \mathbf{1}_i (\mathbf{1}_i^T V_i^{-1} \mathbf{1}_i)^{-1}$.

Definition of L, L_i, W *and* K. Let L be a $\Pi_s \times \Pi_t$ matrix whose ith row is denoted by L_i, and defined by

$$L_i = [f_i^T - f_{i'}^T]$$

and let W be a $\Pi_s \times \Pi_s$ matrix defined by $W = LVL^T$. Let K be a $\Pi_t \times \Pi_s$ matrix defined by

$$K_{ij} = 1 \quad \alpha(i,j) = j$$
$$0 \quad \text{o.w.}$$

Remark. L is the matrix of linear contrasts which transforms the variables of the standard regression into the corresponding variables of the long regression; f_i is a vector which maps (by taking the inner product) a vector of species values into the mean value for species below node i. So $f_i - f_{i'}$ produces the deviation of the mean of the species below node i from the mean of the species below the parent of node i. If the variance–covariance matrix of a random vector x is V, then that of Lx is W. K is a matrix with a row for every species, and a column for every node except the root. An element equal to 1 indicates that the column-node is an ancestor of (or is equal to) the row-node.

Notational convention of bracketed subscripts. Any array dimension indexed by Π_s can also be considered to be indexed by Π_h, according to the partition P of Π_s. It is convenient to be able to use both forms of indexing explicitly. Accordingly unbracketed subscripts will refer in the usual way to indexing by Π_s, and bracketed subscripts will refer to the partitional indexing. Thus W_{ij} is a single element of the matrix W, defined for $i, j \in \Pi_s$. $W_{(i)j}$ is a $\Pi_{di} \times 1$ vector defined for $i \in \Pi_h$, $j \in \Pi_s$. $W_{(ij)}$ is the $\Pi_{di} \times \Pi_{dj}$ submatrix of W, defined for $i, j \in \Pi_h$.

Definition of C. Let C be a $\Pi_s \times \Pi_s$ matrix defined by $C = \mathrm{diag}_{i \in \Pi_s}(\sigma_i^2 - h_{i'})$.

Definition of |. If A and B are two matrices with the same number of rows, then let $A \mid B$ denote the matrix formed by juxtaposing the columns of A and B.

Definition of M^t, N^t, M^s, N^s. If A is a $\Pi_t \times n_A$ matrix of full rank, $n_A \leqslant n_t$, then let $M_A^t = A(A^T V^{-1} A)^{-1} A^T V^{-1}$, and let $N_A^t = I_t - M_A^t$. If A is a $\Pi_s \times n_A$ matrix of full rank, $n_A \leqslant n_s$, then let $M_A^s = A(A^T C^{-1} A)^{-1} A^T C^{-1}$, and let $N_A^s = I_s - M_A^s$. In each case, $rk(M_A) = rk(A)$. In each case, if A is a null matrix, then let $M_A = 0$, and $N_A = I$.

Remark. The M and N are orthogonal projection matrices in the Ω space indicated by their superscript. M projects onto the columns of the subscripted matrix; N projects onto the space orthogonal to them. Orthogonality in Ω_t is taken with respect to V^{-1}, and in Ω_s is taken with respect to C^{-1}. The principal properties of projection matrices, which will be used without comment, are that $M_{A|B} A = A$, and $N_{A|B} A = 0$; that $M_A M_A = M_A$, $N_A N_A = N_A$ and $M_A N_A = 0$; that if the columns of A and B span the same subspace, then $M_A = M_B$ and $N_A = N_B$; and that if the columns of A are orthogonal to the columns of B then $M_A B = 0$ and $N_A B = B$.

Definition of S. Let S be a $\Pi_s \times \Pi_h$ matrix defined by

$$S_{ij} = 1 \quad i \in \Pi_{dj}$$
$$0 \quad i \notin \Pi_{dj}$$

Equivalently, when considered as a $\Pi_h \times \Pi_h$ matrix according to the partition P of Π_s, S is diagonal with $S_{(i)i} = \mathbf{1}_{si}$, $i \in \Pi_h$.

Remark. The following theorem says that the long regression, which has C as its variance–covariance matrix, is equivalent to the standard regression. C is a diagonal matrix. This allows GLIM, for example, to handle the standard regression, even though it does not allow covariances among the errors. The LHS of the statement of the theorem is the residual sum of squares in the standard regression, concentrated for the parameter vectors for $\mathbf{1}_t$ and X, as a function of y, Z and γ. The RHS is the residual sum of squares in the long regression, concentrated for the parameter vectors of S and LX, as a function of y, Z and γ. Before the theorem we formally define the data of the analysis. Note that the null hypothesis is implicit in the definition of y.

Definition of ϵ, y, μ, X, β, Z, n_X, n_Z. Let ϵ be a random variable over Ω_t, distributed as $N(0, V)$. Let μ be a scalar, X a $\Pi_t \times n_X$ matrix of full rank which is linearly independent of $\mathbf{1}_t$, β an $n_X \times 1$ vector, and Z a $\Pi_t \times n_Z$ matrix. Let y be a random variable defined by

$$y = (\mathbf{1}_t \mid X)\begin{bmatrix}\mu \\ \beta\end{bmatrix} + \epsilon.$$

THEOREM 1. *If* γ *is an* $n_Z \times 1$ *vector, then*

$$(y - Z\gamma)^T N^t_{1_t|X}{}^T V^{-1} N^t_{1_t|X}(y - Z\gamma) = (Ly - LZ\gamma)^T N^s_{LX|S}{}^T C^{-1} N^s_{LX|S}(Ly - LZ\gamma).$$

Definition of e, Π_g, n_g, Ω_g, I_g, τ, λ. Let e be a random variable over Ω_s, defined by

$$e = N^s_{S|LX} L\epsilon.$$

Let $\Pi_g = \{i \mid i \in \Pi_h, e_{(i)} \neq 0\}$. Let n_g be the number of elements of Π_g. Let Ω_g be the set of column vectors with real elements indexed by Π_g. Let I_g be the identity matrix over Ω_g. Let $j_i = \min\{j \mid j \in \Pi_{di}, e_j \neq 0\}$, $i \in \Pi_g$. Let τ be a random variable over Ω_s defined by

$$\tau_{(i)} \propto e_{(i)}, \quad \tau^T_{(i)} C^{-1}_{(ii)} \tau_{(i)} = 1, \quad \tau_{j_i} > 0 \quad i \in \Pi_g$$
$$\tau_{(i)} = 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad i \notin \Pi_g.$$

For $i \in \Pi_g$, the conditions define in turn the relative values of the elements of $\tau_{(i)}$, the magnitude of $\tau_{(i)}$ and the sign of $\tau_{(i)}$. Let λ be a random variable over Ω_g defined by

$$e_{(i)} = \lambda_i \tau_{(i)}, \quad i \in \Pi_g.$$

Remark. It is formally possible that $\Pi_g = \{\}$, if all of the variability in y has been explained by X. In what follows I shall tacitly assume that this is not the case. In practical terms, this situation will be obvious because of a zero sum of squares in the standard regression, and in theoretical terms it has no particular interest. There is no possibility of discovering if Z explains variability in y from such a dataset.

Definition of M^g, N^g. If A is a $\Pi_g \times n_A$ matrix of full rank, $n_A \leqslant n_g$, then let $M^g_A = A(A^TA)^{-1}A^T$, and let $N^g_A = I_g - M^g_A$. Note that $rk(M_A) = rk(A)$. In the case that A is a null matrix, let

$M^g_A = 0$, and $N^g_A = I_g$. M^g and N^g are orthogonal projection matrices over Ω_g, and orthogonality is taken with respect to I_g.

Definition of G. Let G be a $\Pi_g \times \Pi_s$ matrix defined by

$$G_{i(j)} = \begin{matrix} \tau^T_{(i)} & i = j \\ 0 & i \neq j \end{matrix}$$

Remark. e is the residual in the long regression after regression of y on X. Π_g is the set of higher nodes at which these residuals are not identically zero. The circumstances in which some of the residuals are identically zero is discussed in §3*e*. Usually, $\Pi_g = \Pi_h$. Ω_g is the data space of the short regression, τ is a vector containing the 'pattern' of the residuals, and λ contains the 'magnitudes' in the sense of §3(*c*). G is a matrix which in combination with C will form the linear contrasts GC^{-1} which transform the long regression into the short regression. G^TGC^{-1} is a projection matrix. The short regression can therefore be seen as a projection of the long regression onto the columns of G^T. The ith column of G^T has zero everywhere except in the radiation of node i, and there it is proportional to $e_{(i)}$, the residuals in the long regression of y on X. This projection ensures that all the residuals after regression on $GC^{-1}LZ$ must lie in the same space, and so must be proportional, within each radiation, to $e_{(i)}$.

Definition of X^g. Let X^g be a $\Pi_g \times rk(GC^{-1}LX)$ matrix defined such that the columns of X^g span the same subspace as those of $GC^{-1}LX$. X^g may be a null matrix.

Remark. This definition is needed in case $GC^{-1}LX$ is not of full rank even though LX is. See §3(*e*). $X^g = GC^{-1}LX$ will satisfy the definition when $GC^{-1}LX$ is of full rank.

Remark. The following theorem shows that the short regression is analytically valid in the case where the working phylogeny is the true phylogeny, and ρ is known and taken as fixed at its true value. It does this by giving the probability density of $N^g_{X^g}GC^{-1}Ly$, the residual vector after regression of $GC^{-1}Ly$ on $GC^{-1}LX$, based on the whole process of computing the long regression, conditioning on τ, and using the random linear contrasts GC^{-1} to form the short regression. The theorem shows that conditional on τ, the probability density is the same as it would have been if $GC^{-1}LX$ were taken as fixed, and the residual's density calculated on the basis of an error ψ distributed as $N(0, I_g)$ in the regression $y^g = GC^{-1}LX\beta + \Psi$. This equivalence of the residual density in the two cases establishes the exactness of the short regression for testing for the addition of $GC^{-1}LZ$. Note that conditional on τ, $GC^{-1}LZ$ is fixed and not random.

THEOREM 2. *Conditional on τ*

$$N^g_{X^g}GC^{-1}Ly \sim N(0, N^g_{X^g}).$$

Definition of T, n_T. Let $n_T = n_s - n_h - n_g$. Let T be a $\Pi_s \times n_T$ matrix of full rank which satisfies $T^TC^{-1}(S|G^T) = 0$, and $rk(T|S|G^T) = n_s$. If $n_s - n_h - n_g = 0$, then T will be a null matrix.

Remark. T will be null only when the working phylogeny is binary and as a consequence the phylogenetic regression and the standard regression are the same.

Remark. The following theorem shows that the long regression with T is equivalent to the short regression. The LHS is the sum of squares of the long regression, concentrated for the parameter vectors associated with LX, S and T. The RHS is the residual sum of squares of the short regression, concentrated for the parameter vectors associated with $GC^{-1}LX$. The

introduction of A allows for possible collinearity between T and LX, or in other words that $rk(X^g) < rk(LX)$.

THEOREM 3. *Let A be a $\Pi_s \times rk(LX|S|T)$ matrix of full rank whose columns span the same subspace of Ω_s as the columns of $(S|LX|T)$. Then conditional on τ,*

$$(Ly - LZ\gamma)^T N_A^{s\,T} C^{-1} N_A^s (Ly - LZ\gamma)$$
$$= (GC^{-1}Ly - GC^{-1}LZ\gamma)^T N_{X^g}^{g\,T} I_g^{-1} N_{X^g}^g (GC^{-1}Ly - GC^{-1}LZ\gamma).$$

Definition of F. If a is a $\Pi_g \times 1$ vector, B is a $\Pi_g \times n_B$ matrix of full rank, $n_B < n_g$, D is a $\Pi_g \times n_D$ matrix of full rank, $n_B + n_D < n_g$, $rk(B|D) = rk(B) + rk(D)$, and $a^T N_{B|D}^g a \neq 0$, then let $F(a, B, D)$ equal

$$\frac{a^T M_{N_B D^a}^g}{a^T N_{B|D^a}^g} = \frac{a^T M_{N_B D^a}^g}{a^T N_B^g a - a^T M_{N_B D^a}^g}.$$

Definition of Z^c. Let Z^c be a matrix of full rank formed by deleting columns from Z in such a way that $(X^g|GC^{-1}LZ^c)$ is of full rank and that $rk(X^g|GC^{-1}LZ^c) = rk(X^g|GC^{-1}LZ)$. Z^c may be a null matrix.

Definition of m_X, m_Z. Let $m_X = rk(X^g)$, and $m_Z = rk(Z^c)$.

Definition of Γ. Let Γ be defined almost surely as a $\Pi_g \times m_Z$ random matrix whose elements are independently distributed in normal distributions, with zero mean, and

$$\text{var}(\Gamma_{ij}) = \frac{e_{(i)}^T C_{(i)}^{-1} e_{(i)}}{(e_{(i)}^T C_{(i)}^{-1} (LZ^c)_{(i)j})^2}$$

This definition fails when the denominator is zero for any i, j.

Definition of Ψ. Let Ψ be a $\Pi_s \times m_Z$ random matrix defined almost surely by

$$\Psi_{(i)j} = \begin{cases} \Gamma_{ij}(LZ^c)_{(i)j} & i \in \Pi_g,\ j = 1 \ldots m_Z \\ 0 & i \in \Pi_h - \Pi_g,\ j = 1 \ldots m_Z \end{cases}$$

Remark. Ψ is the random alternative to Z of the randomization test described in §3(*c*).

Remark. The following theorem shows that the randomization test described in §3(*c*) is the same test as the short regression. Formally, the randomization test is to find the p-value for the null hypothesis that $\gamma = 0$ by finding

$$\Pr\{F(GC^{-1}Ly, X^g, GC^{-1}\Psi) > F(GC^{-1}Ly, X^g, GC^{-1}LZ^c)\},$$

in which the substitution of Ψ for LZ^c is the only difference between the two sides of the inequality. If this probability is low, it implies that randomly selected explanatory variables would rarely explain as much of the remaining variation in Ly as LZ^c does. The short regression's test statistic would also have an F-distribution with m_Z and $n_g - m_X - m_Z$ degrees of freedom.

THEOREM 4. *Conditional on y, $F(GC^{-1}Ly, X^g, GC^{-1}\Psi)$ has an F-distribution with m_Z and $n_g - m_X - m_Z$ degrees of freedom.*