Syst. Biol. 59(1):9–26, 2010

DOI:10.1093/sysbio/syp074
Advance Access publication on November 4, 2009

Phylogenetic Logistic Regression for Binary Dependent Variables

ANTHONY R. IVES[1,*] AND THEODORE GARLAND JR.[2]

[1]*Department of Zoology, University of Wisconsin-Madison, Madison, WI 53706, USA; and*
[2]*Department of Biology, University of California, Riverside, Riverside, CA 92521, USA; E-mail: tgarland@ucr.edu;*
[*]*Correspondence to be sent to: Department of Zoology, University of Wisconsin-Madison, Madison, WI 53706, USA; E-mail: arives@wisc.edu.*

Received 25 March 2008; reviews returned 15 September 2008; accepted 19 September 2009
Associate Editor: Todd H. Oakley

Abstract.—We develop statistical methods for phylogenetic logistic regression in which the dependent variable is binary (0 or 1) and values are nonindependent among species, with phylogenetically related species tending to have the same value of the dependent variable. The methods are based on an evolutionary model of binary traits in which trait values switch between 0 and 1 as species evolve up a phylogenetic tree. The more frequently the trait values switch (i.e., the higher the rate of evolution), the more rapidly correlations between trait values for phylogenetically related species break down. Therefore, the statistical methods also give a way to estimate the phylogenetic signal of binary traits. More generally, the methods can be applied with continuous- and/or discrete-valued independent variables. Using simulations, we assess the statistical properties of the methods, including bias in the estimates of the logistic regression coefficients and the parameter that estimates the strength of phylogenetic signal in the dependent variable. These analyses show that, as with the case for continuous-valued dependent variables, phylogenetic logistic regression should be used rather than standard logistic regression when there is the possibility of phylogenetic correlations among species. Standard logistic regression does not properly account for the loss of information caused by resemblance of relatives and as a result is likely to give inflated type I error rates, incorrectly identifying regression parameters as statistically significantly different from zero when they are not. [Analysis of covariance; ancestor reconstruction; comparative methods; generalized least squares; independent contrasts; morphometrics; phylogeny; regression for binary outcomes.]

Comparative biologists have come to accept the idea that multispecies data sets (e.g., for the relation of brain size to body size) should be analyzed with methods that account for "phylogenetic signal," the tendency for related species to resemble each other (Blomberg and Garland 2002). Under an assumption of Brownian motion–like character evolution, any hierarchical (i.e., nonstar) phylogenetic tree implies that some amount of phylogenetic signal should exist for the phenotypes of a set of species (Freckleton et al. 2002; Blomberg et al. 2003). In general, this resemblance of relatives will violate one or more assumptions of most common statistical methods, such as residuals from a regression model being independent and identically distributed (Felsenstein 1985, 2004; Harvey and Pagel 1991; Garland et al. 2005).

Several approaches have been developed to deal with the statistical issues caused by phylogenetic nonindependence (reviews in Harvey and Pagel 1991; Martins and Hansen 1996; Rohlf 2001; Garland et al. 2005). If the only concern is hypothesis testing, then it is possible to apply conventional statistical methods (e.g., analysis of variance to compare ecologically defined groups of species), compute familiar test statistics (e.g., F ratios), and then compare those test statistics with null distributions that have been derived by simulating or randomizing data in accordance with a specified phylogenetic tree and assumed model of character evolution (Martins and Garland 1991; Garland et al. 1993; Lapointe and Garland 2001). Alternatively, a "known" phylogenetic tree (topology and branch lengths) and an assumed model of character evolution can be used to transform the tip data to make them have equal expected variances and remove correlations related to phylogenetic signal. This is the basis for Felsenstein's (1985) well-known method of phylogenetically independent contrasts (PIC) (Garland et al. 1992). Data containing phylogenetic correlations can also be analyzed by the techniques of generalized least squares (phylogenetic GLS or PGLS: Grafen 1989; Martins and Hansen 1997; Pagel 1997; Duncan et al. 2007). In fact, PIC is an algorithm that constitutes one way of solving such models and is a thus a special case of PGLS methods. Under the assumption of Brownian motion character evolution, PGLS and PIC calculations yield the same parameter estimates and statistical tests (Garland and Ives 2000; Rohlf 2001).

Grafen (1989) first noted that it is often statistically advantageous to estimate a transformation of phylogenetic branch lengths (especially if the branch lengths used are entirely arbitrary) simultaneously with estimation of other parameters in a statistical model (e.g., regression slopes); in effect, this involves estimating the strength of phylogenetic signal in the residuals at the same time as estimating other parameters. Although such methods have often been lumped under the rubric of PGLS, they cannot actually be solved with GLS methods per se because they contain one or more parameters governing the phylogenetic variance–covariance matrix (Forsyth et al. 2004; Huey et al. 2006; Duncan et al. 2007; Lavin et al. 2008). To incorporate the estimation of phylogenetic signal along with other parameters, models must be built on specific assumptions about an evolutionary process that gives the expected variance–covariance structure of the residuals (e.g., an Ornstein–Uhlenbeck [OU] evolutionary process intended to mimic stabilizing selection; Felsenstein 1988). Once these models are formulated, techniques—such as maximum likelihood (ML) or

restricted ML—that estimate both the mean and the variance components of the model can be applied. An example of this type of model is RegOU, which implements regression for continuous-valued traits under the assumption that the residual variation is described by an OU process containing a parameter to estimate the strength of phylogenetic signal (Huey et al. 2006; Lavin et al. 2008).

Most commonly, statistical analyses of comparative data involve dependent variables that are continuously distributed (e.g., home range area, metabolic rate) or at least ordinal (e.g., scale counts of squamate reptiles). Less commonly, the dependent variable may exist in only 1 of 2 possible states, such as which sex is heterogametic. The statistical problem presented by categorical (discrete) dependent variables is distinct from the problem of discrete independent variables. In fact, as with conventional statistical analyses, independent variables that are categorical traits are easy to deal with using dummy variables when applying PGLS (including PIC) or RegOU-type analyses (Grafen 1989; Garland et al. 1993; Martins and Hansen 1997; Pagel 1997; Duncan et al. 2007; Lavin et al. 2008).

Binary dependent variables analyzed by comparative biologists are of many types, including whether a species has temperature-dependent sex determination, whether a female primate advertises her estrus, whether an insect has wings, whether a parasite has an intermediate host, whether an animal builds a nest, whether the organism inhabits lakes or the ocean, and whether the geographic distribution of a species is restricted to a single island or not. However, most of these traits still must evolve through normal microevolutionary mechanisms during which a population will transition from individuals being 100% of one type to being 100% of another. In some cases, the evolutionary transition may occur so rapidly that few populations with "mixed" phenotypes are to be found. Nonetheless, the evolutionary transition of a population (or entire species) will take a finite number of generations, and we would anticipate the occurrence of such binary traits to reflect the phylogenetic history of the species that exhibit them (i.e., they should show phylogenetic signal).

Binary dependent variables also arise for traits that are inherently continuously valued, yet nonetheless are best, or at least most conveniently, scored as distinct categories. For example, bird species might be categorized as either sedentary or migratory, even though species exhibit a range of migratory behaviors in terms of both the proportion of the population migrating and the distance of migration. To deal with this, Boyle and Conway (2007) analyzed data on bird migration first using the dichotomous categorization of sedentary versus migratory and then for the subset of species showing some migration, treating migratory behavior as a continuous variable (see also Thom et al. 2004). When available, continuous-valued data are preferred because they should typically increase statistical power to detect relations in the data (Garland et al. 1993; Al-kahtani et al. 2004; Munoz-Garcia and Williams 2005). Of course, when quantitative information is not available, treating continuous-valued traits as categorical is unavoidable.

Binary dependent variables might also be used when a continuous-valued trait is bimodally distributed or often takes the value 0, thus violating statistical assumptions of standard tests. For example, even though diet composition is a continuous variable, in many lineages of animals most species are either carnivorous or herbivorous, with few being omnivorous. Similarly, in the fish genus *Poeciliopsis* most species have a low matrotrophy index, indicating little placental transfer of nutrients, but some have very high values and almost none have intermediate values (Reznick et al. 2002); the high-matrotrophy index species are almost discretely different from others. In some cases, reasonable hypothesis testing (adequate type I error rates) with a binary dependent variable might be accomplished by using ordinary regression in combination with Monte Carlo simulations using a phylogenetic tree to obtain appropriate statistical distributions (Martins and Garland 1991; Garland et al. 1993; Diaz-Uriarte and Garland 1996). However, better parameter estimates and greater power should be obtainable through the explicit construction of a statistical model for binary dependent variables.

Here, we develop a model of evolution and a corresponding approach to statistical estimation for phylogenetic logistic regression in which there is a binary dependent variable (Y) and zero, one, or more independent variables (X). The independent variables can be continuous and/or discrete, even when there is only a single independent variable. This sets our method apart from existing methods for analyzing binary dependent variables that do not allow for continuous-valued predictors (Maddison 1990; Pagel 1994; Ridley and Grafen 1996; Grafen and Ridley 1997; Pagel 1997; Schluter et al. 1997; Cunningham et al. 1998; Lorch and Eadie 1999; Schultz and Churchill 1999; Perez-Barberia et al. 2002; Lindenfors et al. 2003; Pagel and Meade 2006). Our approach involves generalized linear models (GLMs) that can be used to analyze data from the exponential family of statistical distributions, including Gaussian (normal), Poisson, and binomial distributions (McCullagh and Nelder 1989). The use of GLMs is well established in a phylogenetic context (Martins and Hansen 1997), although most of this work has addressed Gaussian-distributed, continuous-valued dependent variables. Paradis and Claude (2002) proposed operationalizing phylogenetic GLMs for binary dependent variables using generalized estimating equations (GEEs), and Forsyth et al. (2004) independently implemented phylogenetic GEEs to analyze the invasiveness of species, with the binary dependent variable invasive versus noninvasive. Although related to the approach we develop, these methods using GEEs require the specification of an expected variance–covariance matrix reflecting phylogenetic associations. As we describe here, the matrix constructed from the branch lengths of phylogenetic trees under the assumption of Brownian motion evolution (Martins and Hansen 1997) does not give the correct

correlation structure for evolutionary models of binary traits.

In addition to correctly specifying a structure of the variance–covariance matrix, our approach also includes an estimated parameter that governs the strength of phylogenetic signal in the dependent variable. Therefore, just as recent methods for phylogenetic regression estimate phylogenetic signal of residuals simultaneously with regression coefficients (Grafen 1989; Hansen 1997; Freckleton et al. 2002; Huey et al. 2006; Duncan et al. 2007; Lavin et al. 2008), our phylogenetic logistic regression does not require the a priori assignment of phylogenetic signal but instead allows the data to dictate its magnitude in the statistical model. Thus, the outcome of fitting the model to data may be an indication that the residuals contain no phylogenetic signal, so that the trait in question can be viewed as having evolved along a star phylogeny. In such cases, however, it is important to realize that statistical tests will not be the same as for conventional (nonphylogenetic) analyses because the additional parameter for the strength of phylogenetic signal has been estimated and thus there is added uncertainty from this estimation in the model.

Below, we first develop a model of evolution for a single binary variable without considering independent variables. This leads to a statistical test for phylogenetic signal of a binary trait analogous to the tests developed by Freckleton et al. (2002), Blomberg et al. (2003), and Housworth et al. (2004) for continuous-valued traits. We illustrate these methods by analyzing the data set of Brashares et al. (2000) on antelope antipredator behavior (scored in a binary fashion) and perform simulations to check the statistical properties of the parameter estimators. We then consider independent variables, illustrating the statistical methods again with the data set of Brashares et al. (2000) and performing simulations to investigate the statistical properties of the estimators. Complete documentation of the methods is given with the Matlab (MathWorks 1996) computer code "PLogReg.m" (see Supplementary Material available from http://www.sysbio.oxfordjournals.org).

Univariate Case: Phylogenetic Signal

The univariate case corresponds to phylogenetic logistic regression applied in the absence of independent variables, so there is only a single parameter that determines the mean. For this case, we construct a model of phylogenetic change of a binary trait by assuming the trait evolves up a phylogenetic tree. During each small increment of time, there is some probability α_1 that the trait switches to 1 if it is currently 0 and some other probability α_0 that the trait switches to 0 if it is currently 1; thus, evolution up the phylogenetic tree takes the form of a Markov process, as has been used in previous models of evolution of binary traits (e.g., Pagel 1994). This process of evolution leads to a probability distribution for the trait values at the tips of the phylogenetic tree. The absolute magnitudes of α_0 and α_1 set the rates of transitions between 0 and 1 and hence affect the strength of phylogenetic correlations observed among tip species. For example, if α_0 and α_1 have large values, then transitions between 0 and 1 occur rapidly, and this will break down the tendency for closely related species to resemble each other.

Although we make these specific assumptions about the evolutionary process to produce a statistical model, we recognize that the evolution of a real trait through time is unlikely to follow this process precisely. For example, the transition probability might vary among branches of the phylogenetic tree. Nonetheless, basing our analyses around a specific (and rather simple) model of evolutionary change makes it possible to derive an explicit statistical distribution for the values of a binary trait among species.

To specify the model precisely, let Y_i ($i = 1, 2, \ldots, n$) denote a random variable for a trait taking values 0 or 1 for a collection of n phylogenetically related species and let $\mathbf{Y}$ denote the vector of random variables Y_i. (Here, we follow the standard conventions of using uppercase italics Y for a random variable corresponding to trait Y, lowercase italics y for a realization of the random value Y, uppercase bold $\mathbf{Y}$ for a vector of random variables, and lowercase bold $\mathbf{y}$ for a vector of realizations.) Let the $n \times n$ matrix $\mathbf{W}$ describe the phylogenetic tree, with diagonal elements w_{ii} giving the distance from the base to tip i and off-diagonal elements w_{ij} giving the length of the shared branch leading to the last common ancestor of species i and j. Assume trait Y evolves up the phylogenetic tree and that the transition rates α_0 and α_1 are constant, so that branch lengths are proportional to time. Here, we assume that the tips of the tree are contemporaneous and set the diagonal elements of $\mathbf{W}$ to 1, although we discuss the case of noncontemporaneous tips below. With this restriction on $\mathbf{W}$, the matrix $2(\mathbf{1} - \mathbf{W})$ gives the distance between each pair of tips on the phylogenetic tree, where $\mathbf{1}$ is the $n \times n$ matrix with all elements 1. The correlation matrix for $\mathbf{Y}$ given an overall rate of transitions $\alpha = \alpha_0 + \alpha_1$ and assuming the process is at stationarity is (e.g., Martins and Hansen 1997, equation 6c)

$$\mathbf{C}(\alpha) = \exp(-2\alpha(\mathbf{1} - \mathbf{W})), \tag{1}$$

where the exponential is applied individually to each element of the matrix. For a given tree size and shape, the parameter α is associated with phylogenetic signal because the larger α, the greater the rate of transitions and hence the lower the phylogenetic correlations among species (see, e.g., Blomberg et al. 2003). Specifically, as α approaches infinity, $\mathbf{C}(\alpha)$ approaches the identity matrix. If transition rates α were very high then the asymptotic probability of being in State 1 is $\mu = \frac{\alpha_1}{\alpha_0 + \alpha_1}$. Thus, the model can be formulated in terms of 2 parameters: μ that gives the asymptotic expectation of Y_i and α that gives the rate at which phylogenetic correlations among species are lost.

We have chosen to use μ and α as parameters in our statistical model because they have intuitive interpretations and increase the correspondence between the

model and standard logistic regression. Nonetheless, the model could also be formulated in other parameters, for example, α_0 and α_1 (i.e., the transition rates). An important statistical limitation, however, is that only 2 pieces of information are available from data sets, the mean value of Y and the correlation in Y among species. Therefore, it is only possible to estimate 2 parameters. This limitation explains some strategic decisions we made in model formulation. For example, in deriving the correlation matrix $\mathbf{C}(\alpha)$ we assumed that the process is at stationarity, so the probability that the trait at the base of the phylogenetic tree has State 1 equals μ, the same as at the tree tips. If we were to assume that the process were not at stationarity, then the correlation matrix would be $\frac{m(1-m)}{m_0(1-m_0)}\mathbf{C}(\alpha)$, where m is the expected trait value at the tips and m_0 is the expected trait value at the base of the phylogenetic tree. However, this model now has 3 parameters (m, m_0, and α) and only 2 can be estimated, so nothing is gained by this formulation. Therefore, because it leads to no loss of generality, we have used the assumption that the process is at stationarity.

The correlation matrix $\mathbf{C}(\alpha)$ has a different structure from the correlation matrix that is used for phylogenetic regression of continuous-valued traits (Martins and Hansen 1997; Garland and Ives 2000; Lavin et al. 2008). For continuous-valued traits under Brownian motion evolution, the correlations in trait values between species are proportional to the lengths of shared branches (off-diagonals) given in the matrix $\mathbf{W}$, whereas for our evolutionary model of a binary process the correlations are given by $\mathbf{C}(\alpha)$. The structure of $\mathbf{C}(\alpha)$ is identical to that produced for continuous-valued traits following an OU model of evolution under the assumption that the process is at stationarity (Hansen and Martins 1996; Martins and Hansen 1997; Butler and King 2004). The derivation of the OU process given in Blomberg et al. (2003) differs from that given in these citations by assuming that the trait value at the base of the phylogenetic tree is known with zero variance; this has the advantage of producing a transform that returns the original tree (i.e., $\mathbf{W}$) when the parameter giving phylogenetic signal $d = 1$. This assumption is not an option for the case of binary variables because the variance is determined strictly by the mean. Although the matrix $\mathbf{C}(\alpha)$ is never identical to $\mathbf{W}$, when $\alpha = 1$ the strengths of phylogenetic correlations (off-diagonal elements) are of similar overall magnitude for $\mathbf{C}(\alpha)$ and $\mathbf{W}$, and therefore $\alpha = 1$ serves as a rough reference point to gauge the strength of phylogenetic signal. In other words, when $\alpha = 1$ the magnitude of phylogenetic correlations among tip values of the trait is approximately of the same magnitude as the phylogenetic correlations that one would expect for continuous-valued traits evolving in a Brownian motion fashion up the same tree. The relationship between $\mathbf{C}(\alpha)$ and $\mathbf{W}$, however, depends on the structure of $\mathbf{W}$ and therefore should be considered on a case-by-case basis; the program PLogReg.m outputs the matrix $\mathbf{C}(\alpha)$ so that it can be examined directly (see Supplementary Material).

Because the statistical model requires input of the matrix $\mathbf{W}$ that gives expected phylogenetic correlations among species, special consideration needs to be made when a phylogenetic tree has noncontemporaneous tips. The phylogenetic correlations in our model (equation (1)) depend on the branch-length (patristic) distances between tips on the phylogenetic tree given by off-diagonal elements of $2(\mathbf{1} - \mathbf{W})$. To preserve the relative distances for a tree with noncontemporaneous tips, let $\tilde{\mathbf{W}}$ be the matrix with elements $\tilde{w}_{ij}$ giving the shared branch lengths between tips i and j (measured on any scale, e.g., estimates of time, DNA divergence). If $\mathbf{T}$ is the matrix whose elements $t_{ij} = (\tilde{w}_{ii} + \tilde{w}_{jj})/2$ equal the average length from base to tips i and j, and if $\max(\mathbf{T} - \tilde{\mathbf{W}})$ is the maximum value of the elements in $\mathbf{T} - \tilde{\mathbf{W}}$, then $\frac{2(\mathbf{T}-\tilde{\mathbf{W}})}{\max(\mathbf{T}-\tilde{\mathbf{W}})}$ gives the tip-to-tip distances on the phylogenetic tree standardized so that the maximum distance between tips is 2. Thus, in equation (1) we let $\mathbf{W} = \mathbf{1} - \frac{\mathbf{T}-\tilde{\mathbf{W}}}{\max(\mathbf{T}-\tilde{\mathbf{W}})}$ to give a standardized way to incorporate phylogenetic trees with noncontemporaneous tips.

As an explicit example, consider the case in which species A and species B have base-to-tip length 2, species C has base-to-tip length 8, and species B and species C share branch length 1, thereby giving $\tilde{\mathbf{W}} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 8 \end{bmatrix}$. The species-to-species branch lengths are then $\begin{bmatrix} 0 & 4 & 10 \\ 4 & 0 & 8 \\ 10 & 8 & 0 \end{bmatrix}$, and hence the standardized distance matrix $2(\mathbf{1} - \mathbf{W}) = \begin{bmatrix} 0 & 0.8 & 2 \\ 0.8 & 0 & 1.6 \\ 2 & 1.6 & 0 \end{bmatrix}$. Here, species A and species B are nearest and therefore have the lowest corresponding element of $2(\mathbf{1} - \mathbf{W})$, even though in the initial tree $\tilde{\mathbf{W}}$, species B and species C are the phylogenetically related species.

Parameter Estimation

Although it is possible to derive the likelihood function for the evolutionary process we described above (e.g., Pagel 1994) and hence to estimate parameters μ and α using ML estimation, instead we use a procedure that is more flexible and numerically more efficient. Specifically, we estimate μ given α using the quasi-likelihood function and then estimate α given μ using least-squares estimation, repeatedly alternating between estimating μ and α until both values converge. Both quasi-likelihood and least-squares estimation require knowing only the first 2 statistical moments of the probability distribution of trait values among tip species; however, for a binomial process the first 2 moments fully specify the distribution, and therefore the estimation procedure uses all information provided by the data.

The quasi-likelihood function is derived from the expectation and variance of the distribution of $\mathbf{Y}$. Although for any distribution the quasi-likelihood function only approximates the likelihood function, quasi-likelihood estimates are the same as ML estimates, and the asymptotic properties of the estimators which are used to derive, for example, approximate confidence intervals are the same (McCullagh and Nelder 1989). In the evolutionary process described above, the expectation of all elements of $\mathbf{Y}$ is simply μ, and the correlation structure of the distribution of $\mathbf{Y}$ is given by $\mathbf{C}(\alpha)$ (equation (1)), which together define the quasi-likelihood function for a given value of α. Quasi-likelihood estimation underlies GEE (Liang and Zeger 1986; Zeger and Liang 1986; Zeger et al. 1988). The GEEs proposed for phylogenetic analyses of comparative data (Paradis and Claude 2002; Forsyth et al. 2004) have been first-order approximations (GEE1), whereas it is also possible to use second-order approximations (GEE2) that incorporate both the mean components of the models (regression coefficients) and the variance components (those that affect the covariance matrix, such as the parameter α) (Prentice 1988; Zhao and Prentice 1990; Liang et al. 1992). However, for our application the second-order GEE2 is prohibitively complex and the first-order GEE1 often had poor convergence properties (results not presented). We therefore used quasi-likelihood functions directly, employing simplex minimization to find the ML parameter values rather than Newton–Raphson minimization that is typically used in the GEE approach.

Let $\hat{\mu}(\alpha)$ denote the estimate of μ conditional on α. In the parlance of GLMs (McCullagh and Nelder 1989), the quasi-likelihood function uses the link function g, defined such that

$$g(\mathrm{E}(\mathbf{Y})) = g(\boldsymbol{\mu}) = \mathbf{x}b_0, \tag{2}$$

where $\mathbf{x}$ is the $n \times 1$ vector of 1s, $\boldsymbol{\mu}$ is the $n \times 1$ vector of values μ, and g is the logit function,

$$g(\mu) = \log \frac{\mu}{1-\mu}. \tag{3}$$

Thus, the asymptotic mean of Y_i is $\mu = \frac{\exp(b_0)}{1+\exp(b_0)}$. The expectation μ is an increasing function of the logistic regression coefficient b_0, although unlike μ that is bounded between 0 and 1, b_0 is unbounded. For the general multivariate case where $\mathbf{x}$ is a $n \times (p+1)$ column vector containing ones in the first column and p independent variables in the remaining columns, the quasi-log-likelihood score or estimating equation is (McCullagh and Nelder 1989, p. 333)

$$U(\hat{b}(\alpha)|\alpha) = \sum_{p+1} \{(\mathbf{Ax})'\mathbf{V}(\alpha)^{-1}(\mathbf{y}-\boldsymbol{\mu})\} = 0. \tag{4}$$

Although here we only address the univariate case ($p = 0$), we provide equation (4) in its general multivariate form for use later when we discuss the multivariate case. In this equation, $\hat{b}(\alpha)$ is the vector of quasi-likelihood estimates of the regression parameters (in this case b_0) given α, $\mathbf{A}$ is the matrix containing along the diagonal $\boldsymbol{\mu} \bullet (\mathbf{1}-\boldsymbol{\mu})$, where $\bullet$ denotes the element-by-element (or Schur or Hadamard) product of 2 vectors, and $\mathbf{V}(\alpha)$ is the covariance matrix of $\mathbf{Y}$,

$$\mathbf{V}(\alpha) = \mathbf{A}^{1/2}\mathbf{C}(\alpha)\mathbf{A}^{1/2}. \tag{5}$$

The ML estimates of logistic regression coefficients are known to be biased away from zero. To reduce bias, we follow the procedure suggested by Firth (1993) and employed by Heinze and Schemper (2002) for standard logistic regression. Firth (1993) suggested penalizing the log-likelihood function for logistic regression, $\mathrm{LL}(b)$, by $1/2 \log |\boldsymbol{I}(b)|$, where $|\boldsymbol{I}(b)|$ is the determinant of the information matrix given by the second derivative of $\mathrm{LL}(b)$ with respect to the vector b. This leads to the penalized score equation for coefficient b_i of (Heinze and Schemper 2002)

$$U_i^*(\hat{b}(\boldsymbol{\alpha})|\alpha) = U_i(\hat{b}(\alpha)|\alpha) + 1/2\mathrm{trace}\{\boldsymbol{I}(b_i)^{-1}[\partial \boldsymbol{I}(b_i)/\partial b_i]\} = 0. \tag{6}$$

The information matrix $\boldsymbol{I}(b)$ equals $(\mathbf{Ax})'\mathbf{V}(\alpha)^{-1}(\mathbf{Ax}) = \mathbf{x}'\mathbf{A}^{1/2}\mathbf{C}(\alpha)^{-1}\mathbf{A}^{1/2}\mathbf{x}$. For the univariate case without any independent variables, the derivative of $\boldsymbol{I}(b)$ with respect to the coefficients b_0 can be obtained algebraically. However, for the case with independent variables described below, there is no simple algebraic form for the derivative of $\boldsymbol{I}(b)$, and therefore we computed the derivative numerically.

The least-squares estimate of α is obtained by minimizing the least-squares function

$$SS(\hat{\alpha}(\boldsymbol{\mu})|\boldsymbol{\mu}) = -\frac{1}{2}(\log|\mathbf{V}(\alpha)| + (\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}(\alpha)^{-1}(\mathbf{y}-\boldsymbol{\mu})). \tag{7}$$

This expression can be explained by noting that $(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}(\alpha)^{-1}(\mathbf{y}-\boldsymbol{\mu}) = [\mathbf{A}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})]'\mathbf{C}(\alpha)^{-1}[\mathbf{A}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})]$, where $\mathbf{A}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})$ gives the standardized or Pearson residuals that (for fixed $\boldsymbol{\mu}$) have variance equal to 1 (McCullagh and Nelder 1989, p. 37).

Statistical Inference

Confidence intervals of the estimate of b_0, and hence μ, can be obtained using either asymptotic results from the quasi-likelihood function or parametric bootstrapping. (Here, for clarity of presentation we refer to simulations used to obtain confidence intervals for parameter estimates as parametric bootstrapping. We also performed simulations to investigate the statistical properties of the estimators, which we refer to simply as simulations. Both of these procedures, however, involve simulating data from the statistical model.) Simulations (below) show that the likelihood-based

approximate confidence intervals for b_0 are often accurate even for small sample sizes. It is also possible to obtain approximate confidence intervals for α using likelihood-based methods, although these in general performed poorly, and therefore for α we recommend obtaining confidence intervals from parametric bootstrapping.

To derive the asymptotic approximations for confidence intervals, let $\mathbf{b}$ be the $1 \times (p+1)$ vector of regression coefficients. (Here, for generality, we have assumed that there can be p independent variables, although for the present univariate case $\mathbf{b} = b_0$.) The variance of the estimator of $\mathbf{b}$ can be approximated from GEE under the assumption that α is fixed at its least-squares estimate. Let $\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{b}_j}$ denote the $n \times 1$ vector of derivatives of $\boldsymbol{\mu}$ with respect to any column j of $\mathbf{b}$,

$$\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{b}_j} = \frac{\partial g^{-1}(\mathbf{b}_j)}{\partial \mathbf{b}_j} = \boldsymbol{\mu} \bullet (1 - \boldsymbol{\mu}). \tag{8}$$

Then the GEE estimate of the variance in the estimator of $\mathbf{b}$ is (Liang and Zeger 1986)

$$\hat{\sigma}^2(\hat{\mathbf{b}}|\hat{\alpha}) = \frac{\partial \boldsymbol{\mu}'}{\partial \mathbf{b}} \mathbf{V}^{-1}(\hat{\alpha}) \frac{\partial \boldsymbol{\mu}}{\partial \mathbf{b}}. \tag{9}$$

Parametric bootstrapping is performed by simulating data from the process leading to equation (1) using the values of $\mathbf{b}$ and α estimated from the data. Rather than perform the estimation of α directly, instead we estimated $a = -\log \alpha$; the estimator of $-\log \alpha$ was generally less skewed than that for α (or other transforms such as $1/\alpha$), and reversing the sign makes a increase with increasing magnitude of phylogenetic correlations among species. Because $a = -\log \alpha$ is unbounded below, we require a threshold for a below which we conclude that phylogenetic correlations are sufficiently small to be negligible. We set this threshold as $a < -4$ because for values of a below -4, the correlation matrix $\mathbf{C}(\alpha)$ is essentially equal to the identity matrix (provided our standardization procedure is used for trees with noncontemporaneous tips); for example, a correlation in the matrix $\mathbf{W}$ of 0.95 (equation (1)) becomes less than 0.01 in the corresponding matrix $\mathbf{C}(-4)$. Furthermore, at $a = -4$ the regression coefficient estimates are very close to those obtained using conventional logistic regression with the Firth correction. As described previously, a value of $\alpha = 1$ corresponds roughly to the strength of correlations in trait values between species that are of similar magnitude to those obtained for a continuous-valued trait under Brownian motion evolution with covariance matrix $\mathbf{W}$ used in our analyses. Thus, we use the case of $a = 0 = -\log 1$ as a reference point and consider values of a both less than and greater than 0.

Example

To illustrate these methods, we use data provided by Brashares et al. (2000) on 75 species of African antelope. We test the hypothesis that antipredator behavior, specifically whether they hide from predators ($Y = 0$) or flee/fight ($Y = 1$), shows phylogenetic signal. The test reveals highly significant phylogenetic signal (Table 1); the bootstrap 95% confidence interval for a is $(-1.91, 1.32)$, and of the 2000 bootstrap (simulated) data sets, none had values of $a < -4$ that would indicate no phylogenetic signal. Note that even though none of the 2000 simulated values of a equaled -4, we do not report this as significant at the $P < 0.0005$ $(=1/2000)$ level because even if there were a 1 in 2000 chance of a value of $a = -4$, this event might not be realized. Instead, we report $P < 0.005$, in which case the probability of obtaining no value of $a = -4$ would be very small $(<10^{-4})$. Thus, we use an arbitrary but conservative rule of thumb; if one bootstraps m data sets and none satisfies a null hypothesis, then we report a P value for rejecting the null hypothesis as $10/m$.

Properties of the Estimator

To investigate the properties of the estimators in more detail, we simulated data using the phylogenetic tree of the 75 antelope species. We selected values of a (i.e., $-\log \alpha$) to give both weaker ($a = -1$) and stronger ($a = 1$) phylogenetic signal and selected values of b_0 to give on average an equal number of 0 and 1 responses ($\mu = 0.5$ when $b_0 = \log 1$, equation (3)) and on average 4 times as many 0 responses as 1 responses ($\mu = 0.2$ when $b_0 = \log 0.25$). When there are many more zeros than ones (or vice versa), estimates of phylogenetic signal a should be less precise because the data contain less information (see below). Because in the univariate case we are mainly interested in determining the strength of phylogenetic signal, we present only estimates of a and not those for the other model parameter, b_0; the coefficient b_0 (the logit of the expected value of Y, equation (3)) only sets the mean μ.

Figure 1 gives the distributions of estimates of a for 2000 simulated data sets under the different combinations of a and μ. When phylogenetic signal is weaker ($a = -1$), the estimator of a is unbiased. When on average the numbers of 0 and 1 responses of the dependent variable are equal ($\mu = 0.5$, Fig. 1a), phylogenetic signal (i.e., estimates of $a > -4$) is detected in all but 1 of 2000 simulated data sets (i.e., the apparent power to detect signal is $1999/2000 = 0.9995$). In contrast, when on average the numbers of 0 responses are 4 times greater than the numbers of 1 responses ($\mu = 0.2$, Fig. 1b), the estimator of a is less precise (higher variance) and no phylogenetic signal is detected in 2.5% of the simulated data sets (power = 0.975). When phylogenetic signal is stronger ($a = 1$, Fig 1c,d), 27% (for $\mu = 0.5$) and 55% (for $\mu = 0.2$) of the simulated data sets had either all zeros or all ones, making it impossible to estimate the phylogenetic signal. Estimates of a from the remaining simulated data sets are slightly downward biased and less precise (having higher variance) than the case of weaker phylogenetic signal ($a = -1$). This

TABLE 1. Phylogenetic analysis of the univariate case demonstrating strong phylogenetic signal in the antipredator behavior (0 = hide, 1 = flee or fight) of 75 antelope species (data and phylogeny from Brashares et al. 2000)

Parameter	Value	SE	Approximate confidence interval	Bootstrap mean[a]	Bootstrap confidence interval[a]	Bootstrap P value[a]
Phylogenetic logistic regression						
a[b]	−0.118			−0.290	(−1.91, 1.32)	<0.005
b_0	−0.078	0.898	(−1.88, 1.70)	−0.114	(−1.96, 1.64)	

[a]Two thousand data sets were simulated to obtain bootstrap means and confidence intervals. Parametric bootstrapping was also used to test the null hypothesis that there is no phylogenetic signal (i.e., $a = -4$, the lowest value we allow for a).

[b]Estimates of the parameter for phylogenetic signal a ($= -\log\alpha$) and b_0 (the logit of the expectation μ; equation (3)) determining the mean were obtained from equations (1)–(7), with standard errors (SE) and approximate confidence intervals for b_0 obtained using GEE formulae (equation (9)).

leads to a greater number of simulations for which no phylogenetic signal is detected (type II errors), with estimates of $a = -4$ in more than 2% of the simulated data sets for which a could be estimated.

These simulations reveal a seeming paradox: when the value of a is larger, leading to stronger phylogenetic correlations, it may be more difficult to statistically distinguish the value of a from -4; specifically, for the case with $\mu = 0.5$, phylogenetic signal ($a > -4$) was almost always detected when it was weaker ($a = -1$, Fig. 1a) but was not detected in 2% of the data sets (for which an estimate could be obtained) when it was stronger (Fig. 1c). A heuristic explanation for this apparently paradoxical performance of the estimator of a is that when there is strong phylogenetic signal, many simulated data sets have almost all 0 or almost all 1 values. Figure 2 graphs estimates of a versus the observed number of 1 values for the same simulations used to produce Figure 1. The estimates of a become more variable and lower for the data sets with nearly all 0 or all 1 values. All data sets in which no phylogenetic signal was detected contain 2 or fewer 0 outcomes out of 75 possible and therefore contain little information. Thus, this type of problem should be anticipated in real data sets that are sparsely populated by either zeros or ones. This issue should also be considered in parametric bootstrapping, where large numbers of data sets are simulated that may have

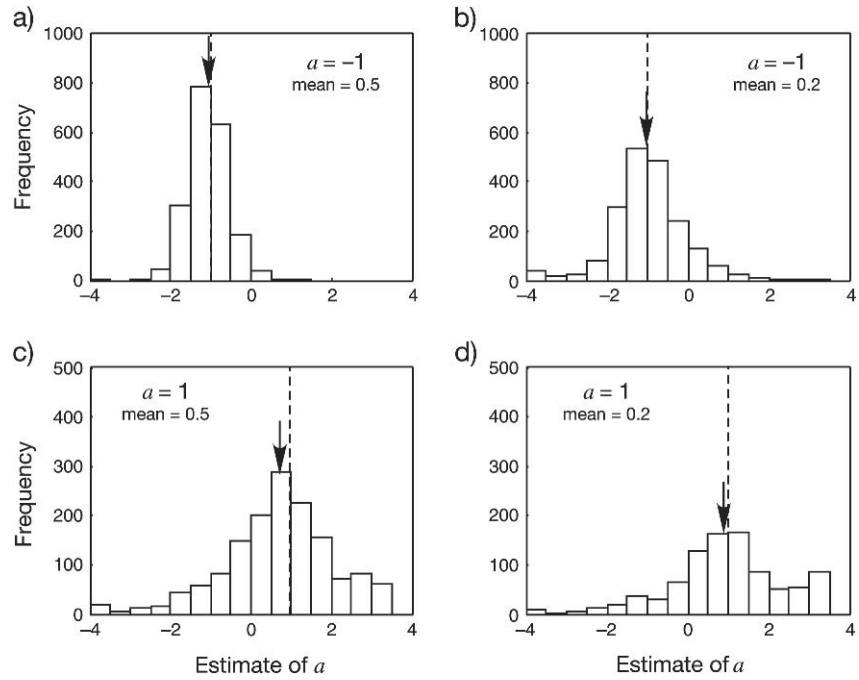

FIGURE 1. For the univariate case, simulations showing that phylogenetic logistic regression with the Firth correction (equations (1)–(9)) gives unbiased estimates of phylogenetic signal ($a = -\log\alpha$), although the precision (variability) of the estimates depends on both the true strength of phylogenetic signal (a) and the mean value of the dependent variable. Two thousand data sets were simulated using the $n = 75$ species phylogenetic tree given by Brashares et al. (2000) with true values of a giving weak phylogenetic signal ($a = -1$) in (a) and (b) and strong phylogenetic signal ($a = 1$) in (c) and (d). The value of b_0 was selected so that the mean value of the trait in (a) and (c) is $\mu = 0.5$ ($b_0 = \log_e 1$) and in (b) and (d) $\mu = 0.2$ ($b_0 = \log_e 0.25$). The values of a used to simulate the data are marked by vertical dashed lines, and the arrows give the mean of the estimates from the 2000 simulated data sets.

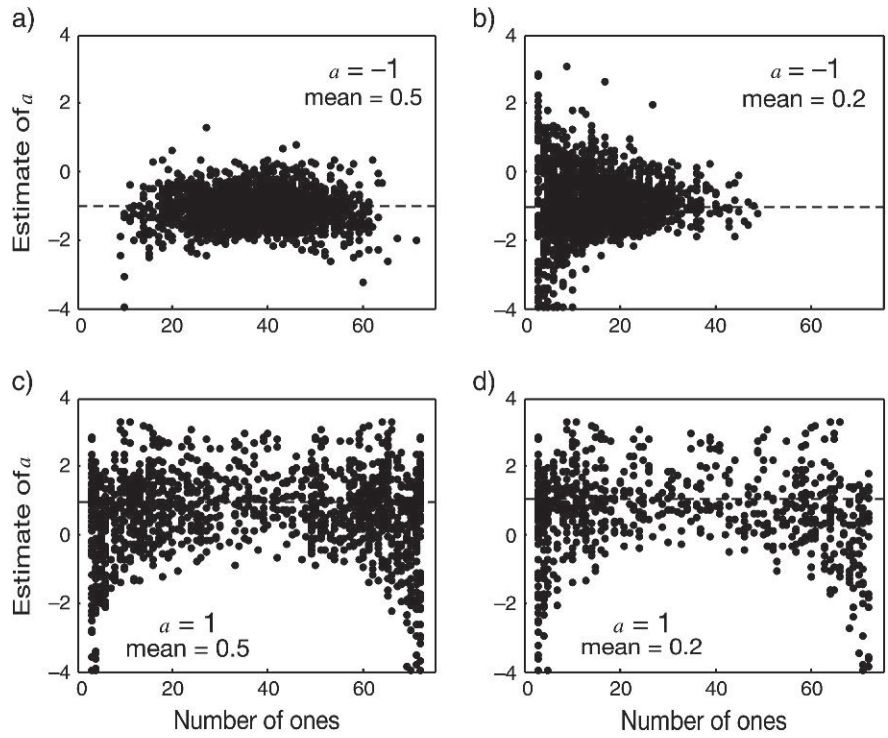

FIGURE 2. Effect of the observed number of zeros and ones in the data set on the estimated value of phylogenetic signal a in 2000 simulated data sets for $n = 75$ species. The simulations are the same as used to create the corresponding panels in Figure 1. In each panel, the horizontal dashed line indicates the location of the true value of a. When the dependent variable Y_i has a large number of either zeros or ones, the estimates of a tend to be lower and more variable. This occurs because data sets with large numbers of zeros or ones contain less information than those with more equal numbers of zeros and ones. The downward bias in the estimate of a observed in Figure 1c,d when $a = 1$ is caused by the large number of data sets with nearly all zeros or nearly all ones.

almost all zeros or ones. If bootstrapped estimates of a show clear bias, then the simulated bootstrap data sets should be investigated to understand the source of the bias.

These effects for data sets that are sparsely populated with zeros or ones can be understood at a more general and heuristic level. If a trait is invariant for a given set of species, then the question of phylogenetic signal becomes moot because we have no standard of comparison. That is, we do not know if the next species outside the clade containing the study species has the same or different value for the trait. If the sister clade for our study species were to be monomorphic for the opposite state, then taken as a whole (study clade plus its sister clade) the set of species would exhibit very strong phylogenetic signal. If the sister clade were monomorphic for the same state as for our study species, then it would beg the question of whether the trait might show phylogenetic signal in some broader phylogenetic sampling of species. Thus, the ability to estimate phylogenetic signal in a binary trait is much more complicated and less intuitive than for a continuous-valued trait (e.g., the K statistic of Blomberg et al. 2003), although even interpreting the meaning of phylogenetic signal in continuous-valued traits for evolutionary processes can be complicated (Revell et al. 2008).

MULTIVARIATE CASE: REGRESSION COEFFICIENTS AND PHYLOGENETIC SIGNAL

The multivariate case refers to phylogenetic logistic regression when there are one or more independent variables. In this case, the probability that Y_i for species i takes a value of 0 or 1 depends in part on independent variables X^j ($j = 1, \ldots, p$). As for the univariate case, we desire a model of an evolutionary process that generates a probability distribution for $\mathbf{Y}$. Although we do not expect all real data sets to be generated by the process we derive, it nonetheless gives a plausible model that can be used for statistical analyses.

Our model divides the process leading to the distribution of trait values among species into 2 components. One component is identical to the univariate case, in which values of Y evolve up the phylogenetic tree with asymptotic probability of being in State 1 equal to μ and transition rate α. At the end of this process, the value of Y for a given species is 0 or 1. In the second component, the values of Y are affected by the species-specific values of the independent variables X^j. For each species i, we assume the value of Y_i evolves toward either 0 or 1 depending on the values of X_i^j, with the rate of evolution no longer depending on the transition rate α but instead depending on the regression coefficients b_j for independent variables X^j. Thus, in the first component

of this process the expected correlation in trait values between species is determined by the parameter α, as in the univariate model. The expectation of the mean trait values (μ_i) in response to independent variables is set in the second component of the process.

An important point in understanding the construction of this model is that the value of α affects only the correlation structure of the residual variation, that is, the variation not explained by the independent variables. Even in the absence of phylogenetic correlation in the residual variation, there may be phylogenetic correlation in the dependent variable itself if the independent variables have phylogenetic structure and the independent variables do indeed affect (statistically speaking) the dependent variable. Nonetheless, it is phylogenetic correlation in the residual variation that is important in formulating the statistical model. As with standard regression and phylogenetic regression with continuously distributed dependent variables, the distributions of the independent variables do not enter into the statistical model. For example, in phylogenetic regression for continuous-valued traits (e.g., Lavin et al. 2008), the expectations for species trait values are set by the independent variables, yet residuals are correlated due to phylogenetic relationships. We derived our 2-component model of evolution with independent variables to give a statistical model that is comparable to phylogenetic regression for continuous-valued traits.

To implement this model, let

$$\boldsymbol{\mu} = \frac{\exp(\mathbf{xb})}{1 + \exp(\mathbf{xb})} \tag{10}$$

be the $n \times 1$ vector of expected values μ_i for each species i, where $\mathbf{x}$ is the $n \times (p + 1)$ matrix whose first column contains ones and remaining p columns contain values of X^j ($j = 1, \ldots, p$) and $\mathbf{b}$ is the $1 \times (p + 1)$ vector containing regression coefficients $b_0, \ldots, b_p$. We assume that the first component of the model occurs with an asymptotic probability of being in State 1 equal to $\bar{\mu}$, the mean value of $\boldsymbol{\mu}$; this gives the largest possible value for the maximum phylogenetic correlations that can be produced by the model. For the second component, we assume that if $\mu_i < \bar{\mu}$, then trait Y will evolve toward 0; if it equals 0 at the end of the first component, it will remain 0, whereas if it equals 1 at the end of the first component, it will switch to 0 with probability $1 - \mu_i/\bar{\mu}$. In this construction, the smaller the value of μ_i, the more rapidly the trait for species i evolves toward 0. Conversely, if $\mu_i > \bar{\mu}$, then trait Y will evolve toward 1; if it equals 1 at the end of the first component, it will remain 1, whereas if it equals 0 at the end of the first component, it will switch to 1 with probability $1 - (1 - \mu_i)/(1 - \bar{\mu})$. It is possible to show that the resulting correlation matrix $\tilde{\mathbf{C}}(\alpha)$ is related to the correlation matrix for the univariate case (equation (1)) by

$$\tilde{\mathbf{C}}(\alpha) = \mathbf{MC}(\alpha)\mathbf{M} - \text{diag}(\mathbf{MC}(\alpha)\mathbf{M}) + \mathbf{I}, \tag{11}$$

where $\mathbf{M}$ is the diagonal matrix with elements $m_{ii} = (1 - \bar{\mu})[\mu_i/(1 - \mu_i)]^{1/2}$ for $\mu_i < \bar{\mu}$ and $m_{ii} = \bar{\mu}[(1 - \mu_i)/\mu_i]^{1/2}$ for $\mu_i > \bar{\mu}$ and the diag() function sets the nondiagonal elements of a matrix to 0.

An important property of this model for multivariate regression is that when phylogenetic correlations are all assumed to be zero (i.e., $\tilde{\mathbf{C}}(\alpha)$ is the identity matrix), it reduces to standard logistic regression. Furthermore, we do not make assumptions about the distribution of the independent variables; they may show phylogenetic signal or they may not. It is possible to formulate a model in which the independent variables evolve along a phylogenetic tree and the dependent variable evolves in response (for an example for continuous-valued traits, see Hansen and Orzack 2005) that contrasts our model in which response to independent variables occurs following the establishment of phylogenetic correlations; however, this would introduce considerable statistical difficulties. It would also necessitate assumptions about the evolution and resulting distribution of the independent variables, whereas our method treats the independent variables as having fixed values (i.e., we do not specify a statistical distribution of the independent variables but instead treat their values as known). Although we recognize that, as with any statistical model, the process we used to construct our model is unlikely to hold exactly for all real data sets, it nonetheless incorporates phylogenetic correlations into logistic regression in a simple and reasonable way and will likely approximate other models of evolution well.

To avoid confusion that can arise regarding transforming independent variables, we need to take a short digression and consider the case of a continuous dependent variable analyzed using either PIC or PGLS. In PIC, independent contrasts must be computed for both dependent and independent variables, and this makes PIC conform to PGLS (Garland and Ives 2000; Rohlf 2001); thus, even though independent contrasts are computed for the independent variables, the phylogenetic signal is nonetheless confined to the residuals of the regression model. In other words, computing independent contrasts for the independent variables is not equivalent to making the assumption that the independent variables themselves show phylogenetic signal. Instead, computing independent contrasts for the independent variables X is better viewed as a transform of variables; because the values of Y are transformed using independent contrasts, so too must the values of X be transformed to match Y. (The same logic applies when computing correlations with PIC.) Our phylogenetic logistic regression is comparable to PGLS, not PIC, in that no transformation of the independent variables is needed.

Parameter Estimation and Statistical Inference

All estimation and inference approaches used for the univariate case can be applied directly to the multivariate case by substituting $\tilde{\mathbf{C}}(\alpha)$ for $\mathbf{C}(\alpha)$ in equations (2)–(9).

Example

We return to the data set of Brashares et al. (2000) to illustrate the multivariate analyses. We test the hypothesis proposed by Jarman (1974) that species living in larger groups are more likely to flee/fight predators, whereas solitary or pair-living species are more likely to hide. Group size ranges between 1 and 70, and we treat log-transformed group size as a continuous variable. Because body size is likely also to affect antipredator behavior, with larger bodied species more likely to flee/fight than hide, we follow Brashares et al. (2000) and also include log body mass as a second independent, continuous-valued variable. Both independent variables were standardized to have mean equal to 0 and standard deviation equal to 1; this makes the regression coefficients represent effect sizes of the independent variables whose magnitudes reflect the size of effect of the variable (as is done, e.g., by convention in path analysis).

In the full phylogenetic analysis with both log group size and log body mass, the effect of group size, b_2, is statistically different from zero, as determined from both the asymptotic approximation and the parametric bootstrapping (Table 2). The bootstrap analysis indicates that there is an upward bias in the estimate of b_2; the bootstrap simulations were performed with an input value of $b_2 = 1.36$ (the value estimated from the data), yet the mean of the bootstrap estimates is 1.57, a bias of $1.57 - 1.36 = 0.21(15\%)$. The lower bound of the bootstrapped confidence interval for b_2, (0.47, 3.20), is higher than that obtained from the GEE approximation, (0.39, 2.33); this can be explained in part by the bias because the difference in lower bounds, $0.47 - 0.39 = 0.08$, is in the same direction as the bias in the mean of 0.21. Although the estimates of b_2 are biased, they are far less biased than the case without the Firth correction (equation (6)), in which the value used to simulate bootstrapped data sets is $b_2 = 1.44$ and the bootstrapped mean estimate of b_2 is 2.07, a bias of 44%.

When standard logistic regression is applied to the data, the estimates of b_2 are much higher, 2.27 with the Firth correction and 2.46 without (Table 2). The Firth correction leads to an unbiased estimate of b_2, as demonstrated by the mean of the bootstrapped estimates of b_2 equal to 2.29. Without the Firth correction, the estimates are strongly upward biased, with the bootstrap mean of 2.73. In both forms of standard regression, the lower bound of the confidence limit is much higher than those obtained in the phylogenetic analyses (Table 2). Even though the standard logistic regression estimate of b_2 with the Firth correction is unbiased, this does not mean that it is correct. In fact, the strong phylogenetic signal estimated by the phylogenetic logistic regression ($a = 0.50$) shows that the assumption of independence among species made by standard logistic regression is not satisfied.

In the phylogenetic logistic regression with and without the Firth correction, there is strong phylogenetic signal with the estimates of a equaling 0.50 and 0.46, respectively. However, in neither case is the phylogenetic signal statistically significant. The apparently low power for detecting phylogenetic signal occurs because the effects of group size are large and therefore some species (those with large group sizes) are expected to have a high probability of fleeing/fighting, whereas others are expected to have a high probability of hiding. This strong effect of an independent variable limits the residual strength of correlation that is possible in the

TABLE 2. Phylogenetic and standard logistic regression parameter estimates for the effects of $\log_{10}$ group size and $\log_{10}$ body mass on the antipredator behavior (0 = hide, 1 = flee or fight) of 75 antelope species

Parameter[a]	Estimate	SE[b]	t score	P value	Approximate confidence interval	Bootstrap mean[c]	Bootstrap confidence interval[c]	Bootstrap P value[c]
Phylogenetic logistic regression with Firth correction								
a	0.50					0.15	(−4, +4)	0.09
b_0	−0.82	0.87	−0.96	0.34	(−2.54, 0.90)	−0.69	(−2.78, 1.36)	0.51
b_1 (body mass)	0.096	0.45	0.21	0.84	(−0.80, 0.99)	0.15	(−0.96, 1.32)	0.39
b_2 (group size)	1.36	0.49	2.78	0.007	(0.39, 2.33)	1.57	(0.47, 3.20)	<0.01
Phylogenetic logistic regression without Firth correction								
a	0.46					−0.72	(−4, +4)	0.15
b_0	−1.06	0.90	−1.18	0.24	(−2.85, 0.72)	−1.30	(−5.00, 1.44)	0.42
b_1 (body mass)	0.11	0.47	0.23	0.82	(−0.82, 1.04)	0.28	(−1.14, 1.95)	0.76
b_2 (group size)	1.44	0.51	2.82	0.006	(0.42, 2.46)	2.07	(0.61, 5.26)	<0.01
Standard logistic regression with Firth correction								
b_0	−0.24	0.30	−0.79	0.43	(−0.83, 0.36)	−0.25	(−0.91, 0.32)	0.41
b_1 (body mass)	−0.65	0.44	−1.53	0.23	(−1.53, 0.23)	−0.66	(−1.77, 0.29)	0.17
b_2 (group size)	2.27	0.57	3.99	0.0002	(1.14, 3.41)	2.29	(1.23, 3.91)	<0.01
Standard logistic regression								
b_0	−0.26	0.31	−0.84	0.40	(−0.88, 0.36)	−0.29	(−0.99, 0.37)	0.38
b_1 (body mass)	−0.72	0.50	−1.43	0.15	(−1.72, 0.28)	−0.77	(−2.04, 0.20)	0.12
b_2 (group size)	2.46	0.67	3.69	0.0002	(1.13, 3.79)	2.73	(1.52, 4.70)	<0.01

[a] All independent variables were standardized to have mean 0 and variance 1 prior to analysis.

[b] Standard errors (SE) of the estimates and confidence intervals were obtained using the GEE approximation (equation (9)).

[c] Parametric bootstrapping was performed by simulating 2000 data sets to obtain means and confidence intervals. Parametric bootstrapping was also used to test the null hypotheses that there is no phylogenetic signal in the residuals (H_0: $a = -4$, 1-tailed test) and that the regression coefficients equal 0 (H_0: $b_i = 0$, 2-tailed tests).

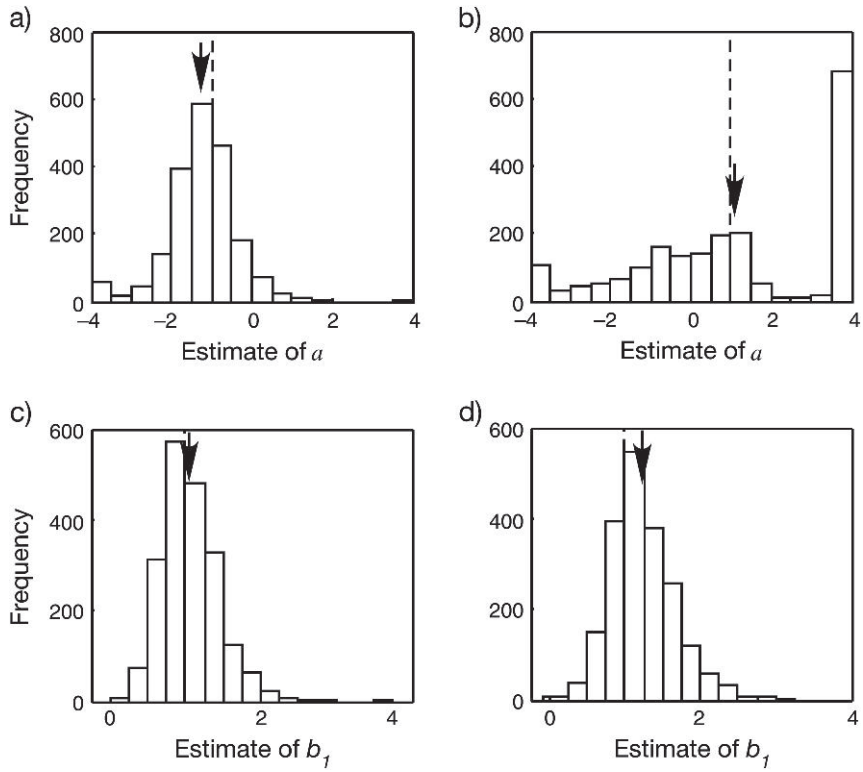

FIGURE 3. The performance of the estimators of phylogenetic signal a and the regression coefficient b_1 in 2000 simulated data sets where the true value of b_1 is 1 and phylogenetic signal is either weak ($a = -1$) (a and c) or strong ($a = 1$) (b and d). The data sets were simulated assuming there is one independent variable X that undergoes Brownian motion evolution up the phylogenetic tree of the 75 antelope species analyzed by Brashares et al. (2000). The true values of a (a and b) and b_1 (c and d) are shown in each panel by the vertical dashed line, and the mean of the estimates from the simulated data are shown by the arrows. a) and b) show that the estimates of a are unbiased, but when phylogenetic signal is strong ($a = 1$, b), the estimator is less precise (has higher variance). In (b) many estimates of a are -4 and 4, the minimum and maximum allowed values of a in the statistical estimation that indicate either no or very high phylogenetic signal in the residual variation. c) and d) show that, whereas the estimator of b_1 is unbiased for weaker phylogenetic signal (c), stronger phylogenetic signal causes the estimator of b_1 to be upward biased; the mean of the 2000 estimates of b_1 is 1.22, even though the value of $b_1 = 1$ to simulate the data.

dependent variable; this is discussed in more detail below. The nonsignificant effect of phylogenetic signal a might incorrectly be used as an argument to perform only conventional logistic regression. The differences between conventional and phylogenetic analyses in the estimates of the regression coefficients in Table 2 give evidence against this.

Properties of the Estimators

To investigate the properties of the estimators, we first perform simulations to analyze bias in the estimates and then assess the accuracy of the confidence intervals and the ability of the analyses to identify regression coefficients that are statistically significantly different from zero. For the phylogenetic relationships among species, we use the tree for antelope given by Brashares et al. (2000). We assume a single continuous independent variable X evolved by Brownian motion evolution, with its covariance matrix $\mathbf{W}$ given by the same phylogenetic tree we use in the model for evolution of the binary dependent variable. Note, however, that the phylogenetic logistic regression model makes no assumption about the distribution of independent variables, and we could equally have assumed that X contained no phylogenetic signal. As in the analyses of the real data, we standardize X to have mean 0 and standard deviation 1, so that the coefficient b_1 is a measure of effect size.

To illustrate properties of the estimators of a and b_1, we simulated 2000 data sets with relatively weak phylogenetic signal ($a = -1$) and relatively strong signal ($a = 1$), in both cases using a true value of $b_1 = 1$ (Fig. 3). In both cases, more than 5% of the data sets had estimates of a less than -4 which corresponds to no detectable phylogenetic signal. More data sets showed no phylogenetic signal when phylogenetic signal was in fact stronger ($a = 1$), as was found in the analyses of the univariate case. This indicates that when phylogenetic signal is strong, there is often less information in a given data set, and therefore the analyses have little power to detect phylogenetic signal in the residual variation.

Estimates of b_1 were biased upward (Fig. 3c,d), the severity of bias being greater for stronger phylogenetic signal (estimate of $b_1 = 1.22$ when $a = 1$) than for weaker

phylogenetic signal (1.06 when $a = -1$). To investigate bias in more detail, we simulated data with a value of $b_1 = 1$ but varied the value of a from -4 (no signal) to 2 (strong signal). Rather than assume there are 75 species, we reduced the number of species to 25; we expected bias to be more severe with smaller sample sizes, and therefore this provides a more stringent test for the robustness of our methods. To create a 25-species phylogeny, we selected every third species from the Brashares et al. (2000) tree in order to preserve the general structure of the phylogeny. Finally, we performed these analyses using phylogenetic logistic regression both with and without the Firth correction (equation (6)) and standard logistic regression with and without the Firth correction.

Not surprisingly, phylogenetic logistic regression with the Firth correction outperformed the other 3 methods (Fig. 4). When phylogenetic signal is weak ($a < -1$), both phylogenetic logistic regression and standard logistic regression with the Firth correction are approximately unbiased, yet as a exceeds zero, standard logistic regression becomes increasingly biased upward, with the mean of the estimates reaching 1.89 when $a = 2$. In contrast, phylogenetic logistic regression is only slightly biased, with the maximum of the mean estimate remaining below 1.15. The other 2 methods, phylogenetic logistic regression without the Firth correction and standard logistic regression, show generally high bias, with standard logistic regression in particular showing very large bias and high imprecision (highly variable estimates) when there is phylogenetic signal.

To assess the GEE approximate confidence intervals for **b** (equation (9)), we simulated 2000 data sets with a moderate phylogenetic signal ($a = 0$) at each of 13 points taken over a range of values of b_1 and at each point used the GEE approximation to determine whether we could reject the null hypothesis that $b_1 = 0$. This procedure generates a power curve for the estimates of b_1. We did these simulations for the case of both $n = 75$ species (Fig. 5a) and $n = 25$ species (Fig. 5b) and also estimated parameters using standard logistic regression both with and without the Firth correction. For the case with $n = 75$ species, phylogenetic logistic regression performed well. When the value of $b_1 = 0$ was used to generate the simulated data, the null hypothesis that $b_1 = 0$ was rejected for 6% of the data sets at an alpha $= 0.05$ level; thus, the method gave type I error rates (rejecting the null hypothesis of $b_1 = 0$ when it is true) that were slightly inflated relative to the specified alpha level in the logistic regression test. In contrast, standard logistic regression both with and without the

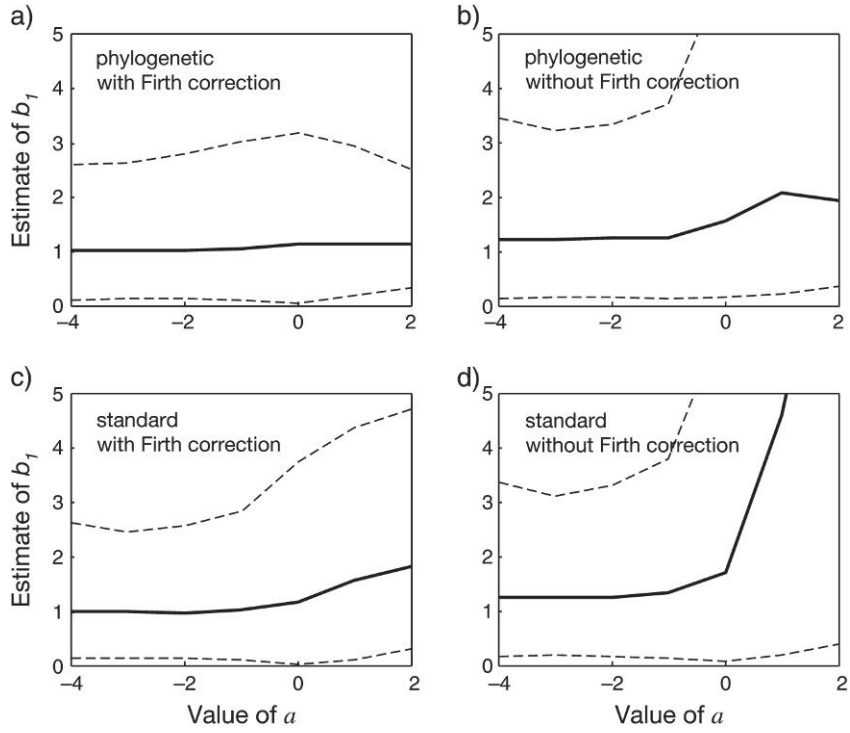

FIGURE 4. Analyses of the bias of the estimator of b_1 using 4 different types of logistic regression: a) phylogenetic logistic regression (equations (1)–(9)), b) phylogenetic logistic regression without the Firth correct (equation (6)), c) standard logistic regression with the Firth correction, and d) standard logistic regression without the Firth correction. Overall, these analyses show that only the phylogenetic logistic regression with the Firth correction (a) has good statistical properties. The data were estimated assuming that there are $n = 25$ species whose phylogenetic tree was created by selecting every third species in the tree for 75 antelope given by Brashares et al. (2000). We assumed that there was a single independent variable evolving under Brownian motion and that the true value of b_1 is 1. Two thousand simulations were performed at each integer value of a from -4 to 2, with -4 corresponding to no phylogenetic signal. The solid line gives the mean of the estimates from the simulated data, and the dashed lines give the 95% inclusion intervals that contain 95% of the estimates.

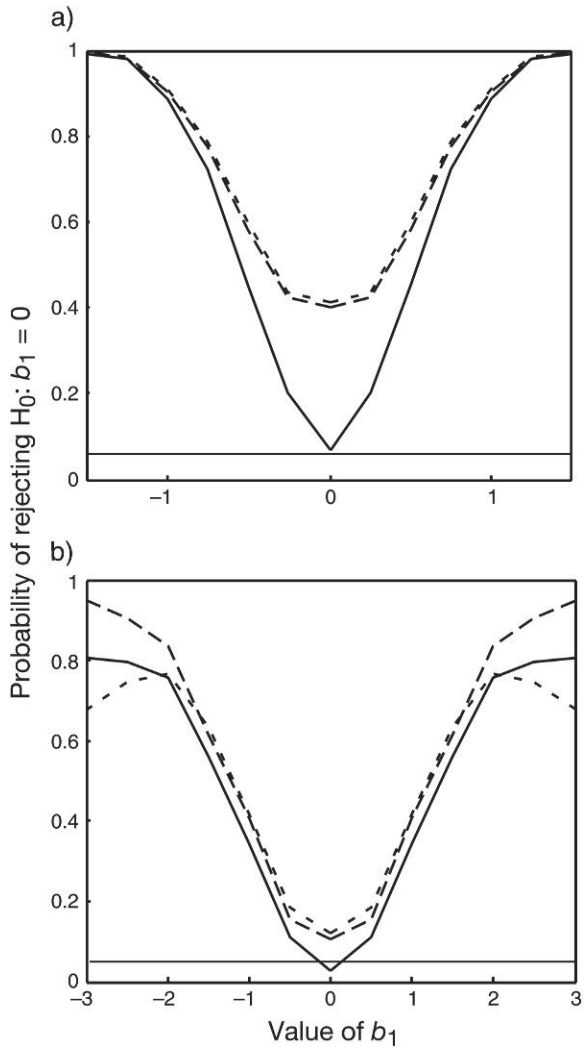

FIGURE 5. Performance of phylogenetic logistic regression with the Firth correction (solid black line) and standard logistic regression with (dashed line) and without (dotted line) the Firth correction, in testing whether the regression coefficient b_1 differs from zero. The vertical axis gives the probability of rejecting the null hypothesis of H_0: $b_1 = 0$ using the approximate GEE confidence intervals (phylogenetic logistic regression, equation (9)) and confidence intervals obtained from conventional logistic regression using estimates of the standard errors. In (a) $n = 75$ species and in (b) $n = 25$ species. At each of 13 values of b_1 ($-1.5, -1.25, \ldots, 1.5$ in a), and $-3, -2.5, \ldots, 3$ in b), 2000 simulations were performed with a moderately strong phylogenetic signal ($a = 0$). For each simulation, the null hypothesis H_0: $b_1 = 0$ was tested using an alpha level of 0.05. If the statistical test for the null hypothesis were performing correctly, then when the true value of b_1 is 0, the null hypothesis should be rejected in 5% of the simulated data sets. The 5% threshold is shown by the thin line at 0.05 in both panels. The results for standard logistic regression with or without the Firth correction show much higher rates of rejecting H_0: $b_1 = 0$, and hence inappropriately high type I error rates. For each simulation, values of X were generated under the assumption of Brownian motion evolution. For (a) where $n = 75$ we used the phylogenetic tree from Brashares et al. (2000), and for (b) where $n = 25$ we created the tree by selecting every third species from the $n = 75$ tree.

Firth correction rejected the null hypothesis that $b_1 = 0$ when it was in fact true for roughly 45% of the simulations when the alpha level was 0.05. Thus, these non-phylogenetic methods frequently and incorrectly identified data sets as having values of b_1 that are statistically significantly different from zero. For the case with $n = 25$ species, the rejection rate of 3% given by phylogenetic logistic regression when $b_1 = 0$ was slightly low; this indicates that the logistic regression test is slightly less likely to reject the null hypothesis when it is true than it should for its specified alpha value. The performance of the standard logistic regressions was better than when $n = 75$, with rejection rates of roughly 10% at $b_1 = 0$. This nonetheless represents a sizable risk (double the nominal type I error rate) that the null hypothesis ($b_1 = 0$) will be rejected even when it is true.

At extreme low and high values of b_1, the probability of rejecting the null hypothesis that $b_1 = 0$ approaches one when $n = 75$, yet for phylogenetic logistic regression this probability plateaus at 0.8 when $n = 25$ (Fig. 5b). This indicates that when there are both small samples sizes and moderate phylogenetic signal ($a = 0$), there is a limit to the statistical power to reject the null hypothesis that $b_1 = 0$.

Overall, these results suggest that using the GEE approximation (equation (9)) to test the statistical significance of logistic regression coefficients is adequate. Nonetheless, we recommend obtaining parametric bootstrap estimates whenever statistical results are marginal (e.g., if the GEE approximation gives a P value of 0.04 and the researcher wants to use an alpha = 0.05 level). Although there is a cost in terms of computing time (our analyses of the antelope data giving Table 2 required several hours on a desktop computer in Matlab), parametric bootstrapping will generally give more accurate confidence intervals for the regression parameters, will identify possible bias in the estimates, and will give the distribution of the estimates of phylogenetic signal a in the residual variation.

Finally, our results illustrate the dangers of using standard logistic regression when there is phylogenetic signal. Standard logistic regression, even with the Firth correction, gave biased estimates of regression coefficients (Fig. 4) and incorrectly high rates of rejecting the null hypothesis that the regression coefficients differed from zero (Fig. 5).

DISCUSSION

Our phylogenetic logistic regression makes it possible to analyze data with binomial dependent variables and continuous or discrete independent variables when the residual variation in the dependent variable is phylogenetically correlated among species. When applied to univariate problems (i.e., when the model contains only the intercept or grand mean, b_0), the method gives a measure of phylogenetic signal, that is, the strength of phylogenetic correlation in trait values among species. Thus, the method is related to methods for continuous-valued traits that estimate the strength of phylogenetic signal (Blomberg and Garland 2002; Freckleton et al. 2002; Blomberg et al. 2003; Housworth et al. 2004; Revell

et al. 2008). It is also related to the randomization test of Maddison and Slatkin (1991) for phylogenetic signal in a single binary character, in which the minimum number of transitions between states up a phylogenetic tree is computed for the data, say N_{obs}, and compared with the distribution of N_{sim} obtained from either randomizing the actual data or generating new data via simulations. In contrast to this type of test, however, our approach is based on parameter estimation and therefore produces a statistical model of the process that can be used, for example, to simulate data. Also, our method makes it possible to incorporate independent variables that could include "nuisance" variables which a researcher wants to factor out of an analysis of phylogenetic signal (e.g., method of calculation for home range area; Perry and Garland, 2002).

When applied with one or more independent variables, our model gives estimates of regression coefficients that account for phylogenetic correlations but does so without making a priori assumptions about the strength of phylogenetic signal in the residual variation; the strength of phylogenetic signal is estimated simultaneously with the regression coefficients. Thus, the method is similar to phylogenetic regression with continuous-valued traits in which phylogenetic signal is simultaneously estimated with the regression coefficients (e.g., Grafen 1989; Huey et al. 2006; Duncan et al. 2007; Lavin et al. 2008; Lajeunesse 2009).

The underlying evolutionary process that gives rise to the phylogenetic structure of our statistical model is identical to that used by Pagel (1994) to derive a correlation test between 2 binary traits; it is a Markov model that assumes a trait has fixed probabilities of changing from State 0 to 1, and from 1 to 0, as it evolves up a phylogenetic tree. Nonetheless, our goal was a model for logistic regression that could accommodate either continuously valued or discrete independent variables and an arbitrary number of them. This is the main difference between our method and not only Pagel (1994) but also other methods for binary traits (Maddison 1990; Ridley and Grafen 1996; Grafen and Ridley 1997; Pagel 1997; Schluter et al. 1997; Cunningham et al. 1998; Lorch and Eadie 1999; Schultz and Churchill 1999; Lindenfors et al. 2003; Pagel and Meade 2006).

Other conceptually distinct formulations of statistical models for binomial dependent variables are possible. For example, Felsenstein (2005) derives a "threshold" model in which there is an underlying and unobserved continuous-valued "liability" trait evolving up a phylogenetic tree; the observed binary trait is then determined by whether or not the liability has crossed a threshold (e.g., the trait value for a species is 1 if its liability trait $x > 0$). It is also possible to derive a model in which the probability of species i being in State 1, say p_i, evolves up a phylogenetic tree. Specifically, the logit of p_i, $\log[p_i/(1-p_i)]$, could be treated as a continuous, normally distributed variable that evolves according to a Brownian motion or an OU process. The resulting phylogenetic logit-normal compound process could then be analyzed as a generalize linear mixed model (McCulloch et al. 2008, p. 64). In contrast to the approach we take here, however, these other approaches do not recover standard logistic regression as a special case when phylogenetic signal is assumed to be absent.

We have referred to the parameter α that gives the rate of switching among trait values during evolution up the phylogenetic tree as a measure of phylogenetic signal. Mathematical justification for this is provided by the fact that the resulting correlation matrix $\mathbf{C}(\alpha)$ (equation (1)) is identical to that produced for continuous-valued traits under the assumption that evolution follows an OU process and is at stationarity, as assumed by Hansen (1997), Martins and Hansen (1997), and Butler and King (2004) (but not by the formulation of Blomberg et al. 2003). Because the parameter governing the strength of stabilizing selection in an OU process has been associated with phylogenetic signal (Blomberg and Garland 2002; Blomberg et al. 2003), it is appropriate to similarly associate α with phylogenetic signal. Hansen and Orzack (2005) make explicit this association between α from a 2-state Markov process and phylogenetic signal (they use the term "phylogenetic inertia") in an OU process. Heuristically, as α approaches infinity (or $a = -\log \alpha$ approaches $-\infty$), the transition rate between states occurs so rapidly that phylogenetic information is wiped out; thus, smaller values of α (or larger values of a) correspond to the emergence of phylogenetic signal. At the opposite extreme, however, the interpretation of α in terms of phylogenetic signal becomes less clear. As α approaches zero (a approaches $+\infty$), all species will have the trait value of their common ancestor. Although this does give the case in which a phylogenetic effect is strongest, it leaves no variation in which to see the phylogenetic resemblances among species (see Univariate Case: Properties of the Estimator). In other words, even though in the limit as α approaches zero there may be phylogenetic signal that generates correlations in trait values among species, as the variances among species go to zero, these correlations become invisible to our statistical methods. Therefore, although the phylogenetic signal is strong, in that species share trait values with their ancestral species, the evidence of this signal is absent, and hence phylogenetic signal becomes undefined. The statistical ramifications of this are seen in simulations demonstrating the increasing difficulty of statistically detecting phylogenetic signal as a becomes much greater than 1 (Figs. 1–3). Although we think it is still reasonable to refer to α (or a) as a measure of phylogenetic signal, just as the parameter for stabilizing selection in an OU process, it is necessary to understand what these parameters are (and are not) in fact revealing.

Statistical Properties

The simulations we used to test the performance of the methods demonstrate that they have generally good statistical properties. The estimates of regression coefficients b_i tended to be slightly biased away from zero

when there was strong phylogenetic signal in the residual variation. This bias can be detected during parametric bootstrapping, and hence we recommend that bootstrapping be done routinely. When interest is only in whether the regression coefficients are statistically significant, however, the GEE methods to test whether $b_i = 0$ (Fig. 5) performed well and are numerically much less intensive than parametric bootstrapping. An alternative approach that we did not investigate is using odds ratio models (Liang et al. 1992; Carey et al. 1993). In odds ratio models, the analyses are not based on probabilities of having values of 0 or 1 but are instead based on odds ratios, that is, the probability of having value 1 divided by the probability of having value 0. Because the odds ratio is not bounded between 0 and 1, as is the probability of having a value of 1, estimates of odds ratios might be less biased than estimates of probabilities. Nonetheless, most analyses performed to test evolutionary and ecological hypotheses are based on regression, and we have focused on logistic regression to provide a counterpart for phylogenetic regression with continuous-valued traits.

Our simulations show that even though the statistical methods perform well, the statistical power to identify phylogenetic signal and effects of independent variables are often low with a binary dependent variable. This indicates that binary data often do not contain a large amount of information. In simulation studies for continuous-valued traits, Blomberg et al. (2003) found that sample sizes of 20 species were often required to detect phylogenetic signal with a statistical power of $\sim$0.8 (and these simulations assumed no measurement error in the tip data [e.g., Ives et al. 2007] and no error in the phylogenetic topology and branch lengths used for analyses). For the case of binary traits, more data will often be needed. A particularly disconcerting result from our simulations is that it may become harder to statistically detect the existence of phylogenetic signal as it becomes stronger (a increases). This is because, as a becomes large, many species will likely have either all 1 values or all 0 values; when this is the case, there is little variation among species through which phylogenetic correlations can be detected. This problem is compounded when there are independent variables that strongly affect the value of the binary trait. For example, suppose one selects 2 species from a phylogenetic tree with different corresponding values of an independent variable X so that the mean values of their traits are $\mu_1 = 0.2$ and $\mu_2 = 0.8$. For 2 species with mean trait values $\mu_i < \mu_j$, the maximum correlation between trait values is

$$r_{\max} = \left(\frac{\mu_i(1-\mu_j)}{(1-\mu_i)\mu_j} \right)^{1/2}. \quad (12)$$

From this, the maximum value of the correlation for 2 species with $\mu_1 = 0.2$ and $\mu_2 = 0.8$ would be 0.25. In general, for any 2 species that have different means μ_i and μ_j, the maximum correlation in the trait is less than 1. Therefore, any effects of independent variables drive down the maximum possible phylogenetic signal in the residual of the dependent variable between species. The reduced phylogenetic signal is thus more difficult to detect.

Despite the possible difficulties in detecting phylogenetic signal, particularly in the presence of strong effects from independent variables, our simulations show that phylogenetic logistic regression should be used whenever there is the possibility of phylogenetic signal. Applying standard logistic regression when there is phylogenetic signal leads to highly biased estimates of regression coefficients (Figs. 3 and 4) and false tests that the regression coefficients differ from zero (Fig. 5). Furthermore, the problems with applying standard logistic regression increase with increasing sample sizes. For example, in our simulations with $n = 75$ species and moderate phylogenetic signal (Fig. 5a), standard logistic regression with an alpha level of 0.05 rejected the null hypothesis that $b_1 = 0$ in 45% of the data sets that were simulated with a true value of $b_1 = 0$. Therefore, standard logistic regression runs the risk of unacceptable type I errors in which the null hypothesis is falsely rejected. As has been emphasized numerous times, this is also a major reason for using phylogenetic methods when analyzing continuous-valued traits (e.g., Grafen 1989; Harvey and Pagel 1991; Martins and Garland 1991; Garland et al. 1992; Diaz-Uriarte and Garland 1996; Garland et al. 2005; Rohlf 2006).

Although the methods performed well, there are some limits to the information they provide. In particular, we have not provided formulae for the approximate quasi-likelihood function, and therefore we do not provide a means for likelihood-based tests (such as likelihood ratio tests) and likelihood-based model selection criteria (such as Akaike information criterion). We are hesitant to provide these formulae because application of likelihood-based methods for our models relies upon asymptotic properties of the estimators as sample sizes become "large". Without simulations designed around a specific data set in hand, it is difficult to determine how large is "large", or how poor the approximations perform when the data set is not "large" (e.g., Nelder and Pregibon 1987; Hurvich and Tsai 1995). This uncertainty explains our preference for parametric bootstrapping approaches in which sample size effects are visible. Given the limited statistical information available in binary dependent variables, large numbers of independent variables should generally not be included in a model, reducing the need for model selection approaches. Until the small-sample-size properties of phylogenetic logistic regression are thoroughly investigated through simulations, we recommend the judicious use of parametric bootstrapped confidence intervals. Careful attention should be paid to joint confidence intervals when multiple independent variables are included; these can be calculated from the distributions of bootstrapped parameter values provided by the computer code "PLogReg.m" (see Supplementary Material).

To our knowledge, Paradis and Claude (2002) were the first to propose phylogenetic logistic regression

using GEEs, and Forsyth et al. (2004) were the first to implement GEEs to incorporate phylogenetic correlations in comparative analyses with non-Gaussian dependent variables. A difficulty with this approach, however, is that direct application of GEEs to phylogenetic data requires the specification of a fixed, feasible correlation matrix **C**—fixed in the sense of containing constant correlation coefficients and feasible in the sense that the data could potentially exhibit the fixed correlation coefficients. However, for logistic regression with independent variables, the correlation matrix must depend on the estimated mean value of traits; in our model this is the case, as given by equation (11). To illustrate this problem, consider the situation of a 3-species polytomy for species 1, 2, and 3, and assume that for the independent variable X, species 1 and species 2 have mean values of $\mu_1 = \mu_2 = 0.2$, whereas for species 3 $\mu_3 = 0.8$. Then using equation (12), the maximum residual correlation between species 3 and the other 2 species would be 0.25, even though the residual correlation between species 1 and 2 could possibly be 1. Thus, even though the 3 species are equally phylogenetically related, differences in the independent variable among species constrain the correlation coefficients in residual variation. This constraint on the correlation structure is not incorporated into the methods of Paradis and Claude (2002) and Forsyth et al. (2004), although Martins and Hansen (1997) point out this difficulty. Note that this difficulty would also arise if phylogenetic regression for continuous-valued traits were applied directly to binary data (e.g., Grafen 1996).

To guarantee that our statistical model produces a feasible correlation matrix, we used a model of an explicit evolutionary process. Not only does this overcome the problem of ensuring a feasible correlation matrix, it also results in a statistical model in which the strength of phylogenetic signal in the residual variation is estimated by the parameter a. The resulting phylogenetic correlation matrix $\mathbf{C}(\alpha)$ is never the same as the correlation matrix that would be derived under Brownian motion evolution of continuous-valued traits, **W** (equation (1)), although when $a = 0$ the strength of phylogenetic correlations (off-diagonal elements of $\mathbf{C}(\lambda)$ and **W**) is often of similar magnitude. Our model does introduce greater statistical complexity than the GEE approach of Paradis and Claude (2002) and Forsyth et al. (2004) because a must be estimated. Nonetheless, we provide Matlab programs with a user-friendly interface for estimation and inference using parametric bootstrapping.

We analyzed the data presented in Brashares et al. (2000) on the antipredator behavior of 75 species of antelope. We reached the same conclusion that species having larger group sizes were more likely to flee or fight predators than hide. Their analyses, however, used log group size as the dependent variable and antipredator behavior and log body mass as the independent variables. Therefore, their statistical tests were based on variability in group size among species rather than variability in antipredator behavior. As a consequence, phylogenetic relatedness was incorporated into residual variation in group size, with phylogenetically related species more likely to have similar group sizes. In our analyses, the statistical tests are based on variability in antipredator behavior, and the estimate of $a(= -\log\alpha)$ gives a measure of the probability that phylogenetically related species show the same antipredator behavior. Given the moderately strong phylogenetic signal we observed in the binary antipredator trait after including the effects of group size and body size (Table 2), we suspect that the approach used by Brashares et al. (2000) could be prone to inflated type I errors (falsely rejecting the null hypothesis of no relationship) due to the loss of information caused by phylogenetic signal in binary data. We have not explored this in detail, however, as the phylogenetic logistic regression approach is better suited for this problem in which antipredator behavior is viewed as the dependent variable (Jarman 1974).

Future Directions

The methods we propose can be used for data with strict binary values, such as whether a species has sexual reproduction or wings. They can also be used for continuous-valued traits that have strongly bimodal distributions or distributions with sufficiently many zeros that the assumptions of methods used for continuous-valued traits (e.g., normally distributed residuals) are badly violated. There is necessarily a loss of information when converting a continuous-valued trait into binary outcomes, and the resulting statistical tests will have correspondingly reduced power. Nonetheless, although there are other less drastic statistical approaches, such as using linear or generalized mixed models (McCulloch et al. 2008), these have not yet been thoroughly investigated for phylogenetic data. Although we present only the case of binary outcomes here, a similar approach can be derived for other members of the exponential family of distributions, such as the Poisson and negative binomial distributions. This will require, however, specific models of evolution that generate a particular distribution for the purposes of parameter estimation and hypothesis testing.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Funding

The work was supported in part by the U.S. National Science Foundation grants DEB-041670 and DEB-0816613 to A.R.I. and by NSF DEB-0416085 to D. N. Reznick, M. S. Springer, and T.G.

Acknowledgments

We especially thank Cécile Ané (UW-Madison) for help, stimulating discussions, and insights into models of evolution. The manuscript was greatly improved

with suggestions from Simon Blomberg, Todd Oakley, Jack Sullivan, and an anonymous reviewer.

REFERENCES

Al-kahtani M.A., Zuleta C., Caviedes-Vidal E., Garland T. Jr. 2004. Kidney mass and relative medullary thickness of rodents in relation to habitat, body size, and phylogeny. Physiol. Biochem. Zool. 77:346–365.

Blomberg S.P., Garland T. Jr. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. J. Evol. Biol. 15:899–910.

Blomberg S.P., Garland T. Jr., Ives A.R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution. 57:717–745.

Boyle W.A., Conway C.J. 2007. Why migrate? A test of the evolutionary precursor hypothesis. Am. Nat. 169:344–359.

Brashares J.S., Garland T. Jr., Arcese P. 2000. Phylogenetic analysis of coadaptation in behavior, diet, and body size in the African antelope. Behav. Ecol. 11:452–463.

Butler M.A., King A.A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am. Nat. 164:683–695.

Carey V., Zeger S.L., Diggle P. 1993. Modelling multivariate binary data with alternating logistic regressions. Biometrika. 80: 517–526.

Cunningham C.W., Omland K.E., Oakley T.H. 1998. Reconstructing ancestral character states: a critical reappraisal. Trends Ecol. Evol. 13:361–366.

Diaz-Uriarte R., Garland T. Jr. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: sensitivity to deviations from Brownian motion. Syst. Biol. 45:27–47.

Duncan R.P., Forsyth D.M., Hone J. 2007. Testing the metabolic theory of ecology: allometric scaling exponents in mammals. Ecology. 88:324–333.

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

Felsenstein J. 1988. Phylogenies and quantitative characters. Ann. Rev. Ecol. Syst. 19:445–471.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Felsenstein J. 2005. Using the quantitative genetic threshold model for inferences between and within species. Philos. Trans. R. Soc. B. 360:1427–1434.

Firth D. 1993. Bias reduction of maximum likelihood estimates. Biometrika. 80:27–38.

Forsyth D.M., Duncan R.P., Bomford M., Moore G. 2004. Climatic suitability, life-history traits, introduction effort, and the establishment and spread of introduced mammals in Australia. Conserv. Biol. 18:557–569.

Freckleton R.P., Harvey P.H., Pagel M. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. Am. Nat. 160: 712–726.

Garland T. Jr., Bennett A.F., Rezende E.L. 2005. Phylogenetic approaches in comparative physiology. J. Exp. Biol. 208:3015–3035.

Garland T. Jr., Dickerman A.W., Janis C.M., Jones J.A. 1993. Phylogenetic analysis of covariance by computer-simulation. Syst. Biol. 42:265–292.

Garland T. Jr., Harvey P.H., Ives A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18–32.

Garland T. Jr., Ives A.R. 2000. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. Am. Nat. 155:346–364.

Grafen A. 1989. The phylogenetic regression. Trans. R. Soc. Lond. B, Biol. Sci. 326:119–157.

Grafen A. 1996. Statistical tests for discrete cross-species data. J. Theor. Biol. 183:255–267.

Grafen A., Ridley M. 1997. A new model for discrete character evolution. J. Theor. Biol. 184:7–14.

Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution. 51:1341–1351.

Hansen T.F., Martins E.P. 1996. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution. 50:1404–1417.

Hansen T.F., Orzack S.H. 2005. Assessing current adaptive and phylogenetic inertia explanations of trait evolution: the need for controlled comparisons. Evolution. 59:2063–2072.

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

Heinze G., Schemper M. 2002. A solution to the problem of separation in logistic regression. Stat. Med. 21:2409–2419.

Housworth E.A., Martins E.P., Lynch M. 2004. The phylogenetic mixed model. Am. Nat. 163:84–96.

Huey R.B., Moreteau B., Moreteau J.C., Gibert P., Gilchrist G.W., Ives A.R., Garland T. Jr., David J.R. 2006. Evolution of sexual size dimorphism in a Drosophila clade, the D. obscura group. Zoology. 109:497–505.

Hurvich C.M., Tsai C.L. 1995. Model selection for extended quasi-likelihood models in small samples. Biometrics. 51: 1077–1084.

Ives A.R., Midford P.E., Garland T. Jr. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Syst. Biol. 56:252–270.

Jarman P.J. 1974. The social organisation of antelope in relation to their ecology. Behaviour. 48:215–267.

Lajeunesse M.J. 2009. Meta-analysis and the comparative phylogenetic method. Am. Nat. 174:369–381.

Lapointe F.J., Garland T. Jr 2001. A generalized permutation model for the analysis of cross-species data. J. Classif. 18:109–127.

Lavin S.R., Karasov W.H., Ives A.R., Middleton K.M., Garland T. Jr. 2008. Morphometrics of the avian small intestine, compared with non-flying mammals: a phylogenetic approach. Physiol. Biochem. Zool. 81:526–550.

Liang K.Y., Zeger S.L. 1986. Longitudinal data analysis using generalized linear models. Biometrika. 73:13–22.

Liang K.Y., Zeger S.L., Qaqish B. 1992. Multivariate regression analyses for categorical data. J. R. Stat. Soc. B Methodol. 54:3–40.

Lindenfors P., Dalen L., Angerbjorn A. 2003. The monophyletic origin of delayed implantation in carnivores and its implications. Evolution. 57:1952–1956.

Lorch P.D., Eadie J.M. 1999. Power of the concentrated changes test for correlated evolution. Syst. Biol. 48:170–191.

Maddison W.P. 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution, 44:539–557.

Maddison W.P., Slatkin M. 1991. Null models for the number of evolutionary steps in a character on a phylogenetic tree. Evolution. 45:1184–1197.

Martins E.P., Garland T. Jr. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: a simulation study. Evolution. 45:534–557.

Martins E.P., Hansen T.F. 1996. The statistical analysis of interspecific data: a review and evaluation of comparative methods. In: Martins E.P., editor. Phylogenies and the comparative method in animal behavior. Oxford: Oxford University Press.

Martins E.P., Hansen T.F. 1997. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149:646–667. Erratum 153:448.

MathWorks I. 1996. MATLAB, version 5.0. Natick (MA): The MathWorks, Inc.

McCullagh P., Nelder J.A. 1989. Generalized linear models. 2nd ed. London: Chapman and Hall.

McCulloch C.E., Searle S.R., Neuhaus J.M. 2008. Generalized, linear, and mixed models. Hoboken (NJ): John Wiley & Sons.

Munoz-Garcia A., Williams J.B. 2005. Basal metabolic rate in carnivores is associated with diet after controlling for phylogeny. Physiol. Biochem. Zool. 78:1039–1056.

Nelder J.A., Pregibon D. 1987. An extended quasi-likelihood function. Biometrika. 74:221–232.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. Biol. Sci. 255:37–45.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scr. 26:331–348.

Pagel M., Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am. Nat. 167:808–825.

Paradis E., Claude J. 2002. Analysis of comparative data using generalized estimating equations. J. Theor. Biol. 218:175–185.

Perez-Barberia F.J., Gordon I.J., Pagel M. 2002. The origins of sexual dimorphism in body size in ungulates. Evolution. 56:1276–1285.

Prentice R.L. 1988. Correlated binary regression with covariates specific to each binary observation. Biometrics. 44:1033–1048.

Revell L.J., Harmon L.J., Collar D.C. 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57:591–601.

Reznick D.N., Mateos M., Springer M.S. 2002. Independent origins and rapid evolution of the placenta in the fish genus Poeciliopsis. Science. 298:1018–1020.

Ridley M., Grafen A. 1996. How to study discrete comparative methods. In: Martins E.P., editor. Phylogenies and the comparative method in animal behavior. Oxford: Oxford University Press.

Rohlf F.J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution. 55:2143–2160.

Rohlf F.J. 2006. A comment on phylogenetic correction. Evolution. 60:1509–1515.

Schluter D., Price T., Mooers A.O., Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. Evolution. 51:1699–1711.

Schultz T.R., Churchill G.A. 1999. The role of subjectivity in reconstructing ancestral character states: a Bayesian approach to unknown rates, states, and transformation asymmetries. Syst. Biol. 48:651–664.

Thom M.D., Johnson D.D.P., Macdonald D.W. 2004. The evolution and maintenance of delayed implantation in the Mustelidae (Mammallia: Carnivora). Evolution. 58:175–183.

Zeger S.L., Liang K.Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 42:121–130.

Zeger S.L., Liang K.Y., Albert P.S. 1988. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 44: 1049–1060.

Zhao L.P., Prentice R.L. 1990. Correlated binary regression using a quadratic exponential model. Biometrika. 77:642–648.

File = Phylogenetic_Logistic_Regression_Documentation_v_1.doc

16 September 2009

Phylogenetic Logistic Regression Documentation for Matlab

PLogReg.m

Anthony R. Ives and Theodore Garland, Jr.

emails: arives@wisc.edu, tgarland@ucr.edu

Accompanies:

Ives, A. R., and T. Garland, Jr. 2010. Phylogenetic logistic regression for binary dependent variables. Systematic Biology 59:9-26.

PLogReg.m is a menu-driven front end for a collection of programs used for phylogenetic logistic regression, as described in detail by Ives and Garland (2010). The documentation here gives an example of PLogReg.m input and output. We are not providing the code as a download on this site, because we want to avoid the code becoming dead. We are continuously updating our code to suit users' needs and fix bugs as they appear. If you want the most recent version of the code, then please contact Ted Garland at tgarland@ucr.edu.

We have not tried the code with Octave, a freeware version of Matlab. However, we would be very interested to have somebody try.

The example we present here is to demonstrate the data requirements and output of PLogReg.m. The two required files (as plain ASCII text files with no formatting of any kind) are:

1. Tip data file. This contains the tip (comparative) data to be analyzed, organized with species in rows, trait values in columns, and optional column headers and/or row labels. The first column can be alphanumeric if it contains tip names (no spaces or funny characters allowed). The first row can also be alphanumeric if it contains variable names (again, no spaces or funny characters)

2. Phylogenetic variance-covariance matrix C. This is a square matrix. Diagonals represent the branch-length distance from root of the tree to each tip (terminal taxon). Off-diagonals represent the branch-length distance from the root to the last common ancestor of each pair of tips.

One way to create this matrix is with the PDDIST.EXE program of our DOS PDAP package as follows:

After you have a tree/data file loaded (typically in the PDI format), then

a. Select option 5 to produce what is named a DSC matrix.
b. Choose M for matrix output.
c. Use of a header it optional (the Matlab code does not use it if it is there).

d. Do not scale all values.
e. Do not write in a compact format without exponents unless all of your branch lengths are in whole numbers or at least have few decimal places.

It is critical that the rows in the tip data file be in *exactly* the same order as the left-to-right (and top-to-bottom) order of the phylogenetic matrix! Otherwise, all results will be nonsense. Note that PDTREE and PDDIST always save PDI and DSC files in this form, so it is convenient to use these two programs in concert to create your tip data and phylogenetic matrices.

For the example presented below, the tip data and variance-covariance matrix files are named **BrashData.txt** and **BrashV.txt**, and they were analyzed for Table 2 in Ives and Garland (2010).

Below, the user input is in red and PLogReg output is in green. The example uses bootstrapping, which is numerically intensive; the example below took 20 hours on an oldish laptop. The code is slow because it goes through extensive evasive action to detect convergence problems. This makes the code robust at the expense of speed. As you will see, convergence is not always obtained, and when it is not, a warning message is printed to the terminal. Also, lack of convergence is flagged in the output file of all parameter estimates called **paralistP.txt**. In extensive simulations, however, the distribution of estimates when convergence is not obtained is indistinguishable from the distribution when convergence is obtained. In other words, the cases of non-convergence are still fine (although they can be excluded from the analyses using the flag in the paralistP.txt file). There are technical reasons for the lack of convergence that we don't want to go into here. Nonetheless, even when the algorithm technically does not converge, the solution is very close to the true ML estimates.

```
PLogReg.m: Phylogenetic Logistic Regression
(c) Anthony R. Ives and Ted Garland Jr. - 31 August, 2009
Based on Ives & Garland (in review)

All estimation (standard and phylogenetic) is performed using the Firth correction

Do you wish to log the session (Turn diary on)? (Y/N) n
The date is: 13-Sep-2009
Time (24 hour clock): 13:02
Data files should contain values corresponding to species (rows)
    and variables (columns) in plain ASCII text.
Missing data can be indicated in the file by "NaN" (in Matlab v. 7) or "-9999".
Hit Return to choose the data file.
You chose data file: /Users/arives/Text Folder/Logistic regression Folder/PLogReg
31Aug09 pgms/BrashData.txt
Input the number of columns in the file, including tip names if present: 9
Does the data file have a header row? (Y/N) n
Does the file include tip names in the first column? (Y/N) n
Your tip data file contains 75 rows and 9 columns
Hit return to choose the covariance matrix file.
You chose matrix file: /Users/arives/Text Folder/Logistic regression Folder/PLogReg
31Aug09 pgms/BrashV.txt
Does the matrix have a header row? (Y/N) n
Which column contains the dependent variable? 6

How many independent variables do you want to analyze? 2
Which column contains independent variable 1? 8
Do you want to log-transform this variable? (Y/N) n
Do you want to standardize this variable to have mean 0 and variance 1? (Y/N) y

Which column contains independent variable 2? 9
Do you want to log-transform this variable? (Y/N) n
Do you want to standardize this variable to have mean 0 and variance 1? (Y/N) y

Methods:
       (O) Ordinary Logistic Regression (assumes a star phylogeny)
       (F) Ordinary Logistic Regression with the Firth correction (assumes a star
phylogeny)
       (P) Phylogenetic Logistic Regression with the Firth correction
Select a method: p
Do you want obtain bootstrap confidence intervals by simulation? (Y/N) y
Input the number of simulations you want to run (default = 2000): 2000
Select an alpha value for confidence intervals (default = .05): .05

OUTPUT FROM PLogReg.m

Data file saved as 'workingDataFile.txt'
Note: this file will overwrite previous files with the same name


Output from Phylogenetic Logistic Regression

Coefficients with +- standard error from GEE approximation
b0 (intercept) = -0.82278 +- 0.86788
b1 = 0.096001 +- 0.44871
b2 = 1.36 +- 0.48516

a = 0.49966

Transformed covariance matrix C saved as C.txt
Note: this file will overwrite previous files with the same name

X, Y, and mu (prediction of Y) saved as Yestimates.txt
Note: this file will overwrite previous files with the same name

Bootstrapping is commensing.  This may take some time.
Every 100 bootstrap simulations, progress is reported on screen and
the accumulated list of parameter estimates is saved in paralistP.txt
in the order: convergeflag (0 or 1), a, b0, b1, ...

PLogReg.m failed to converge
PLogReg.m failed to converge
```

```
phylogenetic logistic regression bootstrap iteration = 100; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 200; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 300; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 400; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 500; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 600; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 700; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 800; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 900; partial parameter list saved
as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1000; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1100; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
```

```
phylogenetic logistic regression bootstrap iteration = 1200; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1300; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1400; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1500; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1600; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1700; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1800; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 1900; partial parameter list
saved as paralistP.txt
PLogReg.m failed to converge
PLogReg.m failed to converge
phylogenetic logistic regression bootstrap iteration = 2000; partial parameter list
saved as paralistP.txt

Output from Bootstrapping
Bootstrapped bounds of (1-alpha) confidence intervals (lb,mean,ub) and p-value for
H0:b=0
b0 (intercept) = (-2.7818, -0.6908, 1.3572)  p = 0.514
b1 = (-0.96363, 0.15097, 1.3234)  p = 0.749
b2 = (0.46844, 1.5696, 3.2027)  p = 0.001

Bootstrapped bounds of (1-alpha) confidence intervals and p-value for H0:a=-4 (1-tailed)
Note: a=-4 corresponds to no phylogenetic signal (see Ives and Garland in review)

a =  (-4, 0.15314, 4)  p = 0.093


Covariance matrix of bootstrap parameter values in order: a, b0, b1,...
Cov_Matrix_boot =
```

```
    6.3439   -0.1541    0.0864   -0.4054
   -0.1541    1.2050   -0.0287   -0.0834
    0.0864   -0.0287    0.3167   -0.1648
   -0.4054   -0.0834   -0.1648    0.4940

Do you want to perform additional analyses? (Y/N)  n

End PLogReg.m
```