Chapter 6
Statistical Issues and Assumptions of Phylogenetic Generalized Least Squares

Roger Mundry

Abstract Using phylogenetic generalized least squares (PGLS) means to fit a linear regression aiming to investigate the impact of one or several predictor variables on a single response variable while controlling for potential phylogenetic signal in the response (and, hence, non-independence of the residuals). The key difference between PGLS and standard (multiple) regression is that PGLS allows us to control for residuals being potentially non-independent due to the phylogenetic history of the taxa investigated. While the assumptions of PGLS regarding the underlying processes of evolution and the correlation of the predictor and response variables with the phylogeny have received considerable attention, much less focus has been put on the checks of model reliability and stability commonly used in case of standard general linear models. However, several of these checks could be similarly applied in the context of PGLS as well. Here, I describe how such checks of model stability and reliability could be applied in the context of a PGLS and what could be done in case they reveal potential problems. Besides treating general questions regarding the conceptual and technical validity of the model, I consider issues regarding the sample size, collinearity among the predictors, the distribution of the predictors and the residuals, model stability, and drawing inference based on *P*-values. Finally, I emphasize the need for reporting checks of assumptions (and their results) in publications.

6.1 Introduction

The method of phylogenetic generalized least squares (PGLS) is an extension of the general linear model. The general linear model, in turn, is a unified framework allowing us to analyze the impact of one or several predictor variables on a single

R. Mundry (✉)
Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
e-mail: roger_mundry@eva.mpg.de

L. Z. Garamszegi (ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, DOI: 10.1007/978-3-662-43550-2_6,

quantitative and continuous response (e.g., Quinn and Keough 2002). In fact, it is nothing else than the classical multiple regression. Categorical predictors ('factors') can be easily included into multiple regression using (usually dummy) coding, interactions can be modeled by including products of predictors into the model, and nonlinear effects are usually included by including transformed predictors (in addition to or instead of the untransformed ones) into the model (for more details about how the effects of factors, interactions, and nonlinear effects can be modeled, see below and, e.g., Cohen and Cohen 1983; Aiken and West 1991). Hence, the general linear model encompasses multiple regression, ANOVA, ANCOVA, and the t-tests.

A crucial assumption of the general linear model is independence of the residuals. This assumption is likely to be violated when the cases in the data set (see Glossary for a definition of some of the terms used here) represent different taxa (e.g., species) which share larger or smaller fractions of their evolutionary history (Felsenstein 1985). Obviously, taxa sharing a larger fraction of their evolutionary history (i.e., having a more recent common ancestor) are likely to be more similar to one another (even after considering the effects of various potential predictors) leading to non-independent residuals.

The method of PGLS (Grafen 1989) has been developed to cope with such phylogenetically driven non-independent residuals. PGLS is an extension of the general linear model, allowing us to account for the phylogenetic history and, by this means, controlling for potential non-independence of the data and leading to independent residuals. The properties and assumptions of PGLS with regard to the assumed evolutionary process and its consequences for the character states of the taxa investigated have received quite some attention. For instance, different models of character evolution (e.g., Brownian motion, Ornstein–Uhlenbeck; Felsenstein 1985, 1988; Hansen 1997; Pagel 1999; Chaps. 5 and 15) can be assumed, or different branch-length scaling parameters (e.g., lambda, kappa, delta; Pagel 1999; Chap. 5) can be chosen, and the particular choices obviously can have crucial implications for the results of the analysis (e.g., Díaz-Uriarte and Garland 1996, 1998; Martins et al. 2002). Similarly, the details of the estimation process can have clear impacts on the reliability of the results (e.g., Revell 2010), as could heterogeneities of the underlying process across the clade investigated (Garland and Ives 2000), heterogeneous sampling (Freckleton and Jetz 2009; Cooper et al. 2010), and errors in the phylogenetic tree (e.g., Díaz-Uriarte and Garland 1998), and these issues (to mention just a few) need to be carefully addressed when using PGLS. Here, I focus on issues and assumptions of PGLS that arise from its similarity with multiple regression. First of all, a PGLS makes assumptions about the distribution of the residuals that are largely identical to those of multiple regression. Furthermore, a model's reliability depends on the distribution(s) of the predictor(s), their number in relation to the sample size, as well as absence of strong collinearity and influential cases. A violation of the assumptions about the residuals or model instability can severely affect the conclusions drawn, and hence, it is of crucial importance that these are thoroughly checked and an assessment is made about how much the models can be trusted.

In the following, I shall treat the statistical assumptions of PGLS and also questions regarding model validity, stability, and reliability. I shall begin with questions regarding the conceptual and technical validity of the model and subsequently consider issues related to the number and distribution of the predictors and interrelations among them (i.e., issues that could be dealt with before the model is fitted). Following that, I shall consider assumptions about the residuals and questions related to model stability (i.e., issues that can be dealt with only after the model was fitted). Finally, I shall briefly touch on questions related to drawing inference based on significance testing and also give recommendations regarding the reporting of the analysis. Most sections have complementary parts in the Online Practical Material (http://www.mpcm-evolution.org) where I show how the respective checks can be conducted in R (Core Team 2013).

It must be stated here that the majority of the issues I consider are not specifically linked to a particular statistical approach (i.e., whether inference is drawn based on information theory, null-hypothesis significance testing or in a Bayesian framework; the exception is the section about drawing inference using null-hypothesis significance testing). Instead, they are generic in the sense that regardless of which particular statistical philosophy one follows, one should consider them. In this context, it might be worth noting that the issues I consider are also not specific to phylogenetic analyses but generic to general linear models. In fact, what I present here are the issues I regularly consider when fitting linear models (such as standard multiple regressions, generalized linear models, or linear mixed models). I also want to emphasize that most of the issues I consider here are not really 'assumptions' of PGLS (or linear models in general) in the sense that a model requires them to be fulfilled (as is the case with assumptions about the residuals). In fact, linear models do not rely on assumptions such as absence of influential cases or collinearity or a certain sample size in relation to the number of cases. However, the confidence one can have in the conclusions drawn from a model might crucially depend on what investigations about, for instance, model stability reveal.

A crucial assumption of all statistical analyses is that the data are correct and complete and for predictors that they are measured without error. A special issue in the context of phylogenetic data is that data availability could vary systematically with species traits (see Garamszegi and Møller 2012). I take it for granted here that the data are correct and complete and that missing data occur at random. Furthermore, I want to emphasize that I am solely focusing on the *statistical* issues related to the use of PGLS and not assumptions regarding the validity of the entire approach, for instance, assumptions about the particular model of evolution (e.g., Brownian motion or Ornstein–Uhlenbeck; Chaps. 5 and 15), the parameter used to model phylogenetic signal in the residuals (e.g., lambda or kappa; see Revell 2010; Chap. 5), and the correctness of the phylogeny used. These 'phylogenetic' assumptions have been treated in quite some detail elsewhere (see above and also Harvey and Pagel 1991 or Nunn 2011, and for possible approaches to deal with uncertainty in the phylogeny or the model of evolution, see, e.g., Chap. 12). However, to my knowledge, with the exception of the distribution of the residuals

(Freckleton 2009), statistical issues like model stability which have relevance for a PGLS as in any other general linear model have received less attention so far.

6.2 Model Development

This section deals with the general outline of the model. In fact, before any model can be fitted, it is required to think about which particular terms should be included and how these should enter the model. In fact, fitting a model crucially requires considering whether the model makes any sense at all or, in other words, whether it is appropriate for the question considered. This is what this section is about. Specifically, it briefly touches the question which predictors at all and also whether interactions and/or nonlinear terms (or interactions involving nonlinear terms) should be considered. I also briefly consider some technical questions regarding model validity that particularly come into play when interactions and/or nonlinear terms are included.

6.2.1 Conceptual Validity of the Model

The first issue to be considered is which predictors should be included in the model. I sometimes have the impression that many researchers believe that this question is not really an issue anymore since ‘model selection’ provides a simple and automated approach telling which predictors are important or not. However, it must be clearly stated that model selection and significance testing are two approaches to statistical inference that are not conformable (e.g., Burnham and Anderson 2002; Mundry 2011), and once model selection has been applied, significance tests are meaningless. Hence, whenever inference should be based on null-hypothesis significance testing (i.e., *P*-values), decisions about which predictors are to be included in the model have to be based on scientific reasoning and cannot be substituted by an automated model selection approach. Moreover, also the proponents of model selection clearly emphasize the need of a careful development of the models to be fitted (e.g., Burnham and Anderson 2002). Of course, decisions about which particular predictors to be considered or controlled for are highly specific to each individual investigation. Nevertheless, I am convinced that many considerations are straightforward in this context. For instance, it seems obvious that any investigation of the impact of whichever predictors on longevity or brain size *must* control for body size. I am convinced that every researcher can easily come up with other such examples in her or his own research area.

Pretty much the same applies to decisions about which interactions to be considered in a model.[1] Again, such decisions *must* be made based on reasoning and cannot be solved using technical solutions such as 'exploratory data analysis' at least when inference should be based on *P*-values. The reason for this is the exact same as for individual predictors. Hence, decisions about which interactions are to be included into a model must be made prior to any analysis and based on reasoning, and they should completely disregard the actual data at hand. As for individual predictors, such decisions are highly specific to the individual study, but again, common sense seems to potentially provide a lot in this context, too. For instance, when one wants to investigate the impact of environmental complexity (e.g., biodiversity) on (relative) brain size, one might need to consider the interaction between environmental complexity and diet as a predictor, simply because folivorous species (that potentially could eat pretty much everything that is 'green') might face much less difficulties in finding food than frugivorous or carnivorous species (which presumably are more specialized in their dietary needs), and this difference might be particularly pronounced in complex environments with high species richness.

The same line of reasoning applies to nonlinear terms. Most commonly nonlinear terms are included as squared terms allowing for an optimum regarding the impact of a covariate on the response (in the sense that one allows for the response to show particularly large (or small) values at intermediate values of the covariate). The details of reasoning about which covariates should be considered to potentially have nonlinear effects are, of course, again very specific to the particular investigation. However, as before, common sense and reasoning can presumably reveal clear hints about which covariates should be considered to be included as nonlinear terms. For instance, in a study of the impact of group size (the predictor) on brain size (the response) within a certain clade (e.g., genus or family), one might hypothesize that intermediate group sizes lead to particularly large relative brain sizes, because small groups are socially not very complex by definition and large groups could be socially not very complex because of being anonymous.

A final issue to be considered here is that adding terms to the model potentially conflicts with the size of the data set which might impose limitations on model complexity (see below). However, my personal preference when being in such a conflict is to give priority to the 'right' model, which means to potentially have more terms than desired in the model. The simple reason is that from a model which is known to be wrong, for instance, because a potentially important confounder or a likely interaction is neglected, potentially not much can be learned (actually, such a model would violate the assumption of independent residuals; see below). I also quite frequently had the impression that a neglected confounder, interaction, etc., can lead to an inflated error variance making tests actually

[1] Having an interaction between two predictors in a model means to allow for a situation where the impact of one of the two on the response is dependent on the value or state of the other and vice versa. Interactions can involve two or more covariates, two or more factors, and any mixture of covariates and factors.

conservative, and this effect can be stronger than the reduction in power and model stability coming along with including an additional term to the model.

6.2.2 Technical Validity of the Model

Whenever an interaction and/or nonlinear (e.g., squared) term is in a model, this necessitates the terms encompassed by them to be in the model as well. Practically, this means that when a two-way interaction is in a model, then the two terms interacting (the 'main effects') *must* be in the model as well (e.g., Aiken and West 1991). Correspondingly, when a model comprises a three-way interaction, the model must also comprise all three two-way interactions encompassed by the three-way interaction and also the respective three main effects. Similarly, when a squared covariate is in a model, this *must* comprise also the respective unsquared covariate. If these requirements are not fulfilled, the model is actually meaningless, and the results revealed for the respective interaction or squared term have no interpretation. It might seem trivial to state this, but among published papers, one finds a surprisingly large proportion being unclear about this point.

6.2.3 Scaling and Centering of the Predictors

A question frequently arising is whether covariates should be *z*-transformed to a mean of zero and a standard deviation of one (note that a *z*-transformation is always done *after* a potential other transformation of a covariate; see below). Strictly spoken, a *z*-transformation is never really required, but it might make interpretation easier quite frequently. First of all, the coefficients obtained for covariates being z-transformed represent the average change in the response per standard deviation of the covariate, and hence, they are directly comparable between all covariates, regardless of what they present and on which scale they were measured (Aiken and West 1991). Hence, getting comparable coefficients is one of the main reasons for z-transforming covariates. The other main reason for z-transforming covariates is to enhance model interpretability in case of models including interactions and/or nonlinear (usually squared) terms (Schielzeth 2010). The reason is that when an interaction between two covariates is in a model, then the coefficients derived for the respective main effects indicate their effect at the respective *other* covariate having a value of zero. Many covariates, however, never have a value of zero in nature (e.g., brain size, body size, life span). As an example, if an interaction between the effects of brain size and life span on a response is in a model, the estimated coefficients would be response $= c_0 + c_1 \times$ brain size $+ c_2 \times$ life span $+ c_3 \times$ brain size $\times$ life span. The coefficient c_1 then represents the effect of brain size on the response for life span being zero—not a very meaningful quantity.

If brain size and life span are z-transformed (meaning that their average is zero), the coefficients c_1 and c_2 have a much more reasonable interpretation; that is, they indicate the effect of the two covariates at the *average* of the respective other covariate. Pretty much the same logic applies whenever nonlinear terms are in a model. For the same reasons, one could also consider centering manually dummy coded factors to a mean of zero (see Schielzeth 2010 for a more in-depth account on these considerations). A final reason to *z*-transform covariates and scale dummy coded factors is to easier create plots of the modeled effects of individual predictors on the response (because ignoring all other terms in the model when plotting the particular effect implies assuming them to be at their average).

However, besides these many advantages of *z*-transforming all covariates (and potentially also to center all dummy coded factors), it has the disadvantage that coefficients reported for different data sets are not comparable anymore since the standard deviation of any particular covariate will vary between data sets and studies (and a coefficient obtained for a *z*-transformed covariate indicates the change of the response per *standard deviation* of the covariate). As a consequence, one should routinely report the original standard deviations of the covariates being *z*-transformed (and also their means; Schielzeth 2010).

6.3 Statistical Reliability of the Model

This section deals with preparatory steps potentially taken to avoid certain problems as well as assumptions about the residuals and questions regarding model stability. If these reveal problems, the validity of the conclusions might be questionable for solely statistical reasons.

6.3.1 Things to be Checked Before the Model Is Fitted

A couple of issues can (and should) be dealt with prior to fitting any model. These refer to the number of predictors in relation to the number of cases, the distribution of the predictors, and absence of strong collinearity.

6.3.1.1 Number of Predictors and Sample Size

Maybe the first and simplest to check is the number of cases (i.e., sample size, *N*) in relation to the number of predictors (*k*).[2] In fact, for standard linear models other

[2] Note that 'number of predictors' should actually be labeled 'number of estimated terms' (meaning that a factor would be counted as the number of its levels minus 1, interactions and

than PGLS, it is well established that the sample size should considerably exceed the number of predictors. If this is not the case, the power of the analysis decreases (potentially considerably), and the results are likely to suffer from instability (i.e., slight changes in the data may lead to drastic changes in the results). However, no simple universally accepted and applicable rule for what would be an appropriate ratio of sample size to the number of predictors does exist (see Field (2005) for recommendations regarding multiple regression). This makes sense, though, since any consideration of sample size needs to take into account expected (or minimum detectable) effect sizes and the power desired (e.g., Cohen 1988; Gelman and Hill 2007). Since expected and minimum effect sizes to be detected are rarely available (if at all in phylogenetic and many other analyses), I tend to use a very simple rule which is that the number of cases should be roughly and at least 10 times the number of estimated terms (including the intercept and, e.g., lambda in case of a PGLS). Surely, this rule is extremely crude and overly simplistic, and whenever possible, one should replace it by something more appropriate (ideally a power analysis based on simulations of the expected or minimum effect size to be detected, conducted using phylogenetic data as close as possible to the ones eventually to be analyzed). At the same time, though, the ten-cases-per-estimated-term rule is simple, allows for a rapid exclusion of model–data combinations that do not make much sense at all (e.g., 5 predictors and 10 cases), and is better than nothing.

If the number of predictors is too large (identified by whichever rule), one needs to reduce them. At least three options do exist in such a case: (1) exclude predictors based on reasoning about which are the least likely to be of relevance for the response under question; (2) conduct a principal component or factor analysis and use the derived (principal component or factor) scores rather than the original covariates (for more about principal component and factor analysis, see, e.g., Budaev 2010, and for phylogenetic principal components analysis see Revell 2009 and Polly et al. 2013); and (3) exclude predictors based on collinearity (i.e., the variance inflation factors revealed for them; see section about collinearity). In the context of phylogenetic analyses, where the number of available taxa might be limited, one will at occasions be confronted with a situation where the model seems to be too complex for the size of the available data set. It is hard to give general recommendations about what can be done in such a situation. However, as stated above, I tend to give priority to the 'right' model (with regard to the predictors included) over one that is oversimplified only to meet an assumed limit of model complexity. After all, the model used needs to be appropriate with regard to the hypotheses to be addressed and the variables to be controlled for. However, a model being too complex might appear unstable (see below). On the other hand, though, a model with two or three predictors might still reveal reasonable results for surprisingly small data sets (with, e.g., just some 15 cases).

(Footnote 2 continued)
squared terms need to be considered, and in the context of a PGLS a parameter like lambda needs to be counted as well).

6.3.1.2 Distribution of Quantitative Predictors (Covariates)

Presumably, the far majority of phylogenetic data sets comprise at least one 'covariate,' i.e., a quantitative predictor such as average group size, body size, brain size, or longevity. Perhaps surprisingly, a PGLS (like a general linear model) does not make any explicit or direct assumptions about the distribution of such covariates (Quinn and Keough 2002; Field 2005). Nevertheless, it is good practice to generally inspect the distribution of each covariate (using, e.g., a histogram or qq-plot; see below) before fitting the model. Besides the fact that such an inspection can reveal typos, it can give hints to potentially 'problematic' (i.e., influential) cases (see below). Such problematic cases are more likely to arise with skewed covariates (Quinn and Keough 2002; Field 2005). For instance, when a covariate is right-skewed (see Glossary), then the values at the upper (right) end of its distribution are likely to have more influence on the model than those in the middle or at the lower end (simply because there are fewer large than small values; Quinn and Keough 2002). As a consequence, one routinely should check the distributions of covariates and try to transform those that are skewed, trying to achieve a roughly symmetrical distribution (e.g., roughly normal or uniform). Most commonly, a log- or square root transformation can be used for this purpose. The log-transformation is 'stronger' and requires all values to be positive, and the square root transformation requires all values to be nonnegative (see also Fig. 6.1 and Box 6.1). One needs to keep in mind, though, that a PGLS ultimately does not make any particular assumption about the distribution of a covariate, and hence, even if a covariate is quite skewed, it might be that a model with the original, untransformed, covariate is more appropriate. However, most usually, it is a good idea to transform skewed covariates right away.

Box 6.1 log-transform or other model?

I quite frequently encountered the belief that rather than transforming a predictor (and/or a response), one should run a modified (i.e., 'nonlinear') model. While this is frequently required when the assumed function linking the response with the predictors is more complex (e.g., sigmoidal), there are many situations where a simple transformation of the predictor (e.g., the log or square root) seems to be the by far easiest and also a fully appropriate solution (Quinn and Keough 2002). This is frequently the case because of the 'logarithmicity' of life. What I mean by this is that the relevance and perception of variables in biological systems are frequently (perhaps usually) logarithmic. For instance, the exact same evolutionary change in body mass of, say, 10 kg would be probably considered 'huge' in case of a mouse or bird, 'minor' in case of an antelope, and 'negligible' in case of an elephant. Correspondingly, one would expect traits that covary with body mass to be affected by the relative and not the absolute change in body mass. Such an expectation can be easily accounted for by taking the logarithm and not the absolute value of body mass as a predictor into the model. Practically, after log-transforming a variable, two species that differ by the same *relative* value differ by the same quantity in the transformed variable. For instance, the difference in log-transformed weights being 100 and 10 would be the same as that between weights being 1,000 and 100. By the way, the base of the logarithm (e.g., 2, 10 or *e*) does not matter, but probably the natural logarithm (base *e*) is most commonly used.

6.3.1.3 Categorical Predictors (Factors)

Categorical predictors (i.e., predictors representing a quality, such as herbivorous, frugivorous, or carnivorous) are usually (and by default) entered into PGLS models by dummy coding them, and after this, factors are modeled as any ‘normal’ (i.e., quantitative) predictor (Cohen and Cohen 1983; Aiken and West 1991). As is obvious, dummy coded categorical predictors do not have the property of having a ‘distribution’ of a particular shape (e.g., normal of uniform). Nevertheless, it is important to inspect the frequency ‘distribution’ of a factor with particular focus on the frequencies of the rarer levels. Specifically, none of the levels should be too rare. From a more statistical perspective, cases of rare levels are likely to be unduly influential as compared to cases of more common levels. But also common sense tells that rare levels are unlikely to reveal much reasonable information. Assume, for instance, a study of male song complexity as a function of male investment in the brood in songbirds. Such a study might need to control for factors like whether the species is a cooperative breeder or whether females and males engage in duets (both partners regularly singing a structured song pattern simultaneously). However, both cooperative breeding and duetting species might be very rare in the data set (say, each is represented by less than five species). As a consequence, it might be a better idea to drop those species from the data rather than including two additional factors into the model. The argument for doing so would be that from such small numbers of species, not much can be learned about the respective factors anyways.

When several factors are relevant for the model, it can also be important to check how many times the combinations of their levels do occur in the data set (particularly when their interactions should be included). Here, pretty much the same logic as for the frequency distribution of the levels of a single factor applies: If a combination of levels of two or more factors is very rare (using the above example: if there were, e.g., only two species being cooperative breeders *and* duetting), then such cases are particularly likely to be relatively influential. This becomes even more of an issue when the interaction between the two factors should be included (and one should keep in mind that including the interaction between two factors is only possible and makes sense when all combinations of all

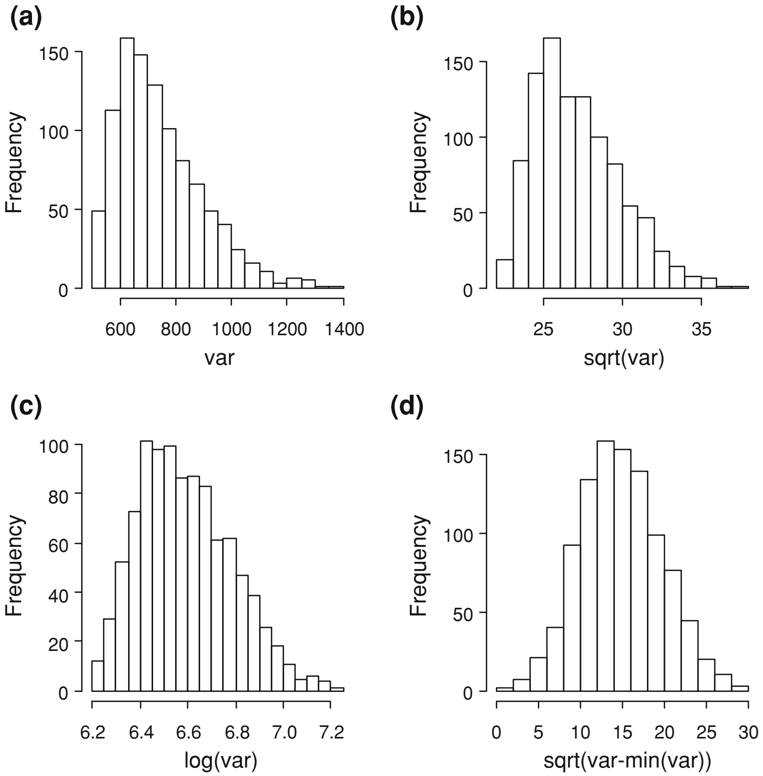

Fig. 6.1 Illustration of the effects of transformations. The original variable **a** is moderately right-skewed with a minimum considerably larger than zero. Neither a square root **b** nor a log-transformation **c** is very effective in removing the skew. However, a standardization of the variable to a minimum of zero and a subsequent square root transformation remove the skew very effectively **d**

their levels occur at least twice in the data[3]). However, such rare levels or combinations of levels of factors might mainly compromise model stability with regard to the factors themselves but not much for the effects of other predictors. An investigation of model stability (see below) will reveal whether this is the case.

6.3.1.4 Collinearity

Absence of strong collinearity (a.k.a., ‘multicollinearity‘) among the predictors is an important requirement for the validity of the results of linear models (Quinn and

[3] For estimating the effect of an interaction reasonably well, more cases per combination of the levels of the factors would be needed.

Keough 2002; Field 2005). 'Absence of collinearity among predictors' basically means that they are not redundant or, in other words, that each of them provides information which is not given by the others (or combinations of them). The simplest case of collinearity is just two predictors being pretty correlated. However, absence of any large (absolute) correlations between the predictors does not rule out collinearity. For instance, one might think of a model which includes the average number of males and females per social group (of the taxa considered) *and* sex ratio or total group size. The latter two predictors are derived from the former two and, hence, do not provide any additional information. As a consequence, any set of predictors comprising three of the variables mentioned would be highly collinear. Finally, collinearity can also arise from the number of predictors being too large.

The consequences of collinearity are simple. First, conclusions about the impact of individual predictors being collinear with others get (potentially very) unreliable and uncertain (Quinn and Keough 2002; Field 2005; Zuur et al. 2010). This manifests in increased standard errors of parameter estimates (and consequently larger confidence intervals), and particularly, non-significance (i.e., $P > 0.05$) can be largely a consequence of collinearity rather than being indicative of absence of an effect. Second, a model suffering from collinearity is likely to be (potentially very) unstable, meaning that small changes in the data can lead to (potentially drastic) changes in the parameter estimates obtained for the collinear predictors (see Freckleton (2011) for an in-depth treatment of the effects of collinearity).

A simple mean to detect collinearity is inspection of so-called variance inflation factors (VIF; Quinn and Keough 2002; Field 2005). These are based on the following principle: For each predictor, one model is fitted, taking the particular predictor as the response and the others as the predictors (note that this means that one gets one VIF value for each of the predictors; note also that a response variable is not needed in this context). Then, the R^2 of the respective model is calculated and this, in turn, is used to derive the VIF, as $\mathrm{VIF} = 1/(1-\mathrm{R}^2)$ (see Fox and Monette 1992 for how factors with more than two levels are treated in this context). As is obvious, when the R^2 gets large, the denominator of the equation approaches zero, and hence, the VIF gets large, too. An 'ideal' VIF, i.e., one indicating no collinearity whatsoever at all, has a value of one, and the larger the VIF the worse. For the question of what is a too large VIF, no simple answer does exist. Occasionally, one sees the rule that a value of ten or larger is clearly indicative of a problem (Quinn and Keough 2002; Field 2005) but also much smaller thresholds of three or two have been suggested as being indicative of potential collinearity issues (Zuur et al. 2010).

Unfortunately, collinearity is likely to arise in phylogenetic analyses since many of the commonly used predictors are most likely to scale allometrically (i.e., correlate with body size), leading to potentially conflicting needs of controlling for body size and avoiding collinearity. In case clear collinearity is detected, one could consider the simple omission of one or several of the predictors associated with a large VIF (based on the reasoning that they anyway do not provide much information in addition to that provided by the other predictors, but see Freckleton

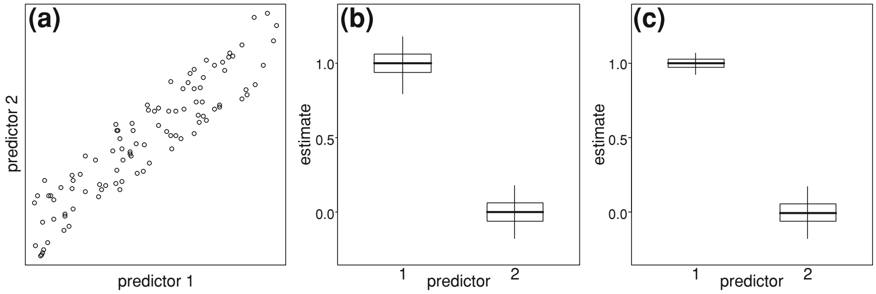

Fig. 6.2 Two collinear predictors that each show reasonable variation at a given value of the respective other predictor (**a**). When sampling predictor 1 from a uniform distribution with a minimum of 0 and a maximum of 10, and predictor 2 as the value of predictor 1 plus a value sampled from a uniform distribution with a minimum of 0 and a maximum of 4 (with a total $N = 100$), the average variance inflation factor for both of them (across 1,000 simulation) was 7.40. When simulating a response by adding a value from a standard normal distribution (mean = 0, sd = 1) to predictor 1, it appeared that the effects of the two predictors could still be assessed quite reliably ((**b**); median, quartiles, and percentiles 2.5 and 97.5 % of 1,000 simulated data sets). However, when predictor 2 was simulated independently of predictor 1, the average variance inflation factor dropped to 1.01 and the variance in the estimate for predictor 1 decreased (**c**). The simulation was based on non-phylogenetic data and a standard linear model

2011). An alternative is to combine predictors (Freckleton 2011) using, for instance, a principal component or factor analysis and use the derived (principal component or factor) scores rather than the original covariates (e.g., Quinn and Keough 2002; for more about principal component or factor analysis, see Budaev 2010). Another option is to try to include more and/or other taxa in the data set, selecting them such that collinearity among the predictors in the phylogeny is minimized, an approach basically being a modification of the method of 'phylogenetic targeting' (Arnold and Nunn 2010) applying another criterion.

However, some level of collinearity will frequently be unavoidable in phylogenetic data sets and the question arises of what to do when this is the case. First of all, it is worth noting that collinearity affects estimation of the effects of predictors being collinear with one another but not that of others. Hence, if predictors merely in the model to control for their effects are collinear to one another but the key 'test predictors' (see below) are not collinear with others, collinearity might be less of a reason to worry. Secondly, even a larger VIF associated with one of the test predictors is not necessarily 'deadly.' This is particularly the case when there is reasonable variation in the predictor at given values of the other(s) it is collinear with, meaning that their independent effects can still be assessed with some certainty. However, estimation would still be more precise when there was no collinearity (Fig. 6.2; see also Freckleton 2011). If one is uncertain about whether collinearity affects conclusions (having VIF values between, say, two and ten), one should assess model stability. This could be done by comparing results from a model with all the predictors included with those obtained from additional models excluding one or several of the collinear predictors and checking whether this has larger consequences for the conclusions drawn. If this is the case, it basically reflects insufficient knowledge and an inability to tease apart the effects of the collinear predictors.

As mentioned earlier, one can use VIF values to exclude predictors in case their number is too large in relation to the number of cases (see above). The procedure is simple: One fits models one after the other and iteratively excludes the predictor with the largest VIF.

6.3.2 Things to be Checked After the Model was Fitted

After the model structure being clarified and having completed the initial checks of the data, one finally can fit the model. Once this is done, one needs to worry about two further issues, namely the distribution of the residuals and absence of influential cases.

6.3.2.1 Distribution of the Residuals (And the Response)

Since the PGLS is, in essence, a general linear model (i.e., a multiple regression) accounting for non-independence of the residuals arising from a phylogenetic history, it has the same assumptions about the distribution of the residuals as a standard general linear model (Freckleton 2009). In particular, these are normality and homogeneity of the residuals.

Normality of the Residuals

Normality of the residuals[4] is a somewhat pretty critical assumption, although, to my knowledge, the consequences of its violation have rarely been systematically investigated in the framework of a PGLS. Taking the general linear model as a reference, it appears that the consequences of violations of this assumption depend on the particular pattern of the violation and also the sample size. For instance, when the sample size is larger and the residuals are somewhat symmetrically distributed, type I and type II error rates are not that strongly affected (see, e.g., Zuur et al. 2010 and references therein), and for PGLS, Grafen and Ridley (1996) showed that it performed reasonably well even when the response was actually binary (however, one should better consider approaches specifically developed for binary responses; see Ives and Garland 2010; Chap. 9). However, skewed

[4] Note that residuals of a PGLS are actually multivariate normal (Freckleton et al. 2011), which has implications for practical checks of their distribution; see the Online Practical Material (http://www.mpcm-evolution.org) for more.

distributions of the residuals and particularly residual distributions comprising one or a few outliers are a reason to worry (and these usually lead to a reduction of power). Hence, one should routinely inspect the distribution of the residuals. Most usually, one will employ visual checks of the assumption, namely a qq-plot (and potentially also a simple histogram). If these reveal residuals to be skewed, one should consider transformations of the predictor(s) and/or the response (see also next section). In case the residuals appear to comprise outliers, one should first check whether these result from coding errors, and if this is not given, one should ask oneself whether there are any predictors missing (e.g., data usually arising from both sexes but occasionally from only females or males). When there appears a missing predictor, one should consider including it into the model or dropping rare levels in case it being a factor (see above). Note that an outlier in the residuals might also give hint to an evolutionary singularity (Nunn 2011; Chap. 21).

A frequently asked question is whether one should use formal checks of the distribution of residuals (e.g., test for normality). However, a *P*-value is always a function of (at least) two properties of the data: the effect size (here, the deviation from normality) and the sample size. Practically, this means that the *P*-value alone does not provide a simple criterion for rejection (or not) of the assumption that the residuals are normally distributed but can only be interpreted in conjunction with the sample size. Hence, eyeballing a qq-plot (or a histogram) is usually considered the most reliable and valuable tool for assessing whether the residuals are roughly normally distributed (e.g., Quinn and Keough; Zuur et al. 2010).

Homogeneity of the Residuals

The other, and presumably more critical, assumption about the residuals is that they are homogeneous (a.k.a. ‘homoskedastic’). This means that the variation in the residuals should be the same, regardless of the particular constellation of values of the predictors. Not much is known about consequences of violations of this assumption (i.e., ‘heteroskedasticity‘) in the framework of a PGLS. However, the consequences of heterogeneous residuals are probably not such that they simply lead to increased type I or type II error rates. For instance, for the independent-samples t-test (which is a special case of the general linear model), heteroskedasticity can lead to clearly increased type I *and* type II error rates, depending on whether residual variance correlates positively or negatively with sample size (e.g., Ramsey 1980). As a consequence, one should be quite vary regarding violations of this assumption.

A check of this assumption is pretty straightforward, and again, one usually employs a visual check (Quinn and Keough 2002; see also above). What one usually does is plotting the residuals of the model (on the y-axis) against its fitted values (on the x-axis), and what one wants to see here is nothing (i.e., no pattern in the cloud of points). More specifically, what should be discernible from this plot is simply no pattern whatsoever at all; that is, the residuals should show the same pattern of scatter around zero over the entire range of the fitted values. Figure 6.3 shows an example of homogeneous residuals (a) and also two common patterns of

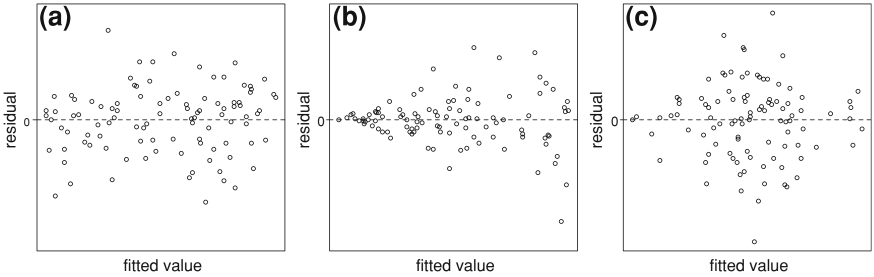

Fig. 6.3 Illustration of the assumption of homogeneous residuals and deviations from it. When the residuals are homogeneous, no relation between residual variance and the fitted value is discernible (**a**). The probably most common violation of the assumption of homoskedasticity reveals a residual variance being positively correlated with the fitted value (**b**). In such a case, a log- or square root transformation of the response might alleviate the issue. Quite frequently, one also sees a pattern where the residual variance is large at intermediate fitted values and small at small and large fitted values (**c**). Such a pattern can occur as a consequence of bottom and ceiling effects in the response and/or when one or several predictors show many intermediate and few small and large values. When the response is bounded between zero and one, the arcsine of the square root-transformed response may (or may not) alleviate the issue; otherwise, a careful selection of the taxa included in the model may help. In case of such a pattern, one should also be particularly worried about influential cases

the assumption being violated (b, c). In certain cases of heteroskedasticity (i.e., when the residual variance is positively or negatively correlated with the fitted value), a transformation of the response might alleviate the issue (Fig. 6.3). The question might arise as to how such plots usually look like when the assumption of homogeneous residuals is not violated. In the Online Practical Material (http://www.mpcm-evolution.org) I show how experience regarding the issue can be rapidly gained.

A potential cause of heterogeneous residuals is a misspecified model. For instance, when a covariate has a nonlinear effect on the response which is not accounted for by the model, then this might become obvious from a plot of residuals against fitted values (Fig. 6.4). Heterogeneous residuals could also arise when the pattern of impact of the predictor on the response is not homogeneous over the entire phylogeny (Rohlf 2006) or when an important main effect or interaction is missing. Obviously, the results of a model with such unmodelled structure in the residuals can be pretty misleading. Hence, one should try to include the missing terms (or revise the taxa investigated in case the effects of a predictor being heterogeneous across the phylogeny). However, such an a posteriori change in the model structure might need to be accounted for when it comes to inference (since a hypothesis being generated based on inspection of some data and then tested using the same data will lead to biased standard errors, confidence intervals, and P-values; Chatfield 1995).

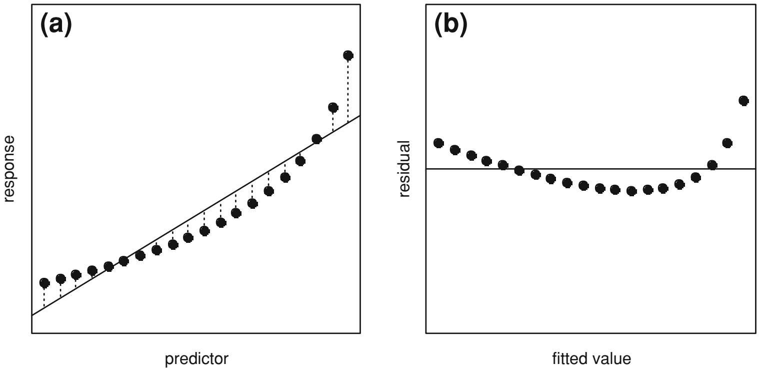

Fig. 6.4 Example of a misspecified model leading to a clear pattern in the residuals. Here, the predictor has a nonlinear effect on the response not being accounted for by the model (*straight line* in **a**). As a consequence, the residuals are large for small and large fitted values and small for intermediate fitted values (**b**)

6.3.2.2 Absence of Influential Cases

Another issue that can only be evaluated once the model was fitted is absence of influential cases. Absence of influential cases means that there are no data points that are particularly influential for the results of the model, meaning that the results do not largely differ depending on whether or not the particular cases are or are not in the data. The influence of a particular data point on the results can be investigated from various perspectives, and which of them is relevant depends on the focus of the analysis. In case of a PGLS where inference about particular effects is usually crucial for the results, the estimates of the effects (and perhaps their standard errors) will usually be in the focus of considerations of model stability.

In the framework of a standard general linear model or generalized linear model, model stability is usually investigated by simply removing cases one at a time and checking how much the model changes, whereby ‘model change’ is usually investigated in terms of estimated coefficients and fitted values (‘dfbetas’ and ‘dffits,’ respectively; Field 2005). For the assessment of whether any particular case is considered too influential (in the sense of questioning the results of the model), no simple cutoffs do exist, and even if it were possible to unambiguously identify influential cases, the question of what to do with such piece of information is not easy to be answered. Assume, for instance, a situation where an analysis of a data set with 100 taxa revealed an estimate for the effect of a certain covariate being 1. Dropping taxa one by one from the data set revealed that for 99 of the taxa removed, the estimate under consideration had values between 0.9 and 1.1, but when removing the remainder taxon, the estimate obtained was –0.2. What would one conclude from such a result? It is tempting to conclude that 99 out of 100 data sets revealed an estimate within a small range (0.9–1.1) indicating a stable result. However, it is more the opposite which is true here. In fact, this particular outcome also implies that the actual estimate revealed crucially depends on a single taxon being in the data set or not (in the sense that if it is not included, the results are totally different). But what would then consider to be ‘truth’? Obviously, there is no simple answer to these questions and the probably most sensitive one can do is to ask oneself why the particular taxon being found to be so influential is so different from the others? A simple (and not the most implausible explanation) is the existence of an error (of coding, a typo, etc.), but potentially, it is also due to a variable missing in the model or an evolutionary singularity (Nunn 2011; Chap. 21). In fact, the taxon might crucially differ from all others in the data, for instance, by being the only cooperative breeder among the species (to use the example from above). In such a case, one would probably conclude that the story might be different for cooperative breeders and exclude them from the analysis because of poor coverage of this factor by the data at hand (but note that such operations should not be required when the appropriate model is properly specified in advance and when factors and covariates are properly investigated in advance; see above).

6.4 Drawing Conclusions

The final step after the model was developed and fitted and decided to be reliable and stable is to draw conclusions from the results. This penultimate section deals with some issues coming along with this step.

6.4.1 Full-Null Model Comparison

A very frequently neglected issue in the context of the interpretation of more complex models (i.e., models with more than a single predictor) is multiple testing. In fact, as shown by Forstmeier and Schielzeth (2011), each individual term in a model has a five percent chance of revealing significance in the absence of any effects. As a consequence, models with more than a single predictor have an increased chance of at least one of them revealing erroneous significance. A simple means to protect from this increased probability of false positives is to conduct a *full-null model comparison*. The rationale behind this is as follows: The full model is simply the model fitted, and the null model is a model comprising only the intercept (but see below). Comparing the two gives an overall *P*-value for the impact of the predictors as a whole (which accounts for the number of predictors), and if this reveals significance, one can conclude that among the set of predictors being in the model, there is at least one having a significant impact on the response.[5] The *P*-values obtained for the individual predictors are then considered

[5] Note that this requires the model to be fitted using maximum likelihood; see the Online Practical Material (http://www.mpcm-evolution.org) for more details.

in a fashion similar to post hoc comparisons; that is, they are considered significant only if the full-null model comparison revealed significance.

In this context, it might be helpful to distinguish between *test predictors* and *control predictors*. This distinction can be helpful when not all predictors being in the full model are of interest in the light of the hypotheses to be investigated, but some are included only to control for their potential effects. Such predictors could and should be kept in the null model to target the full-null model comparison at the test predictors. For instance, assume an investigation of the impact of social and environmental complexity on brain size. In such a case, body size will be only in the model to control for the obviously trivial effect of body size on brain size. Hence, one could and should keep body size in the null model since its effect is trivial and taken for granted and not of any particular relevance for the hypotheses to be investigated (except for that it needs to be controlled for). I want to emphasize, though, that the entire exercise of the full-null model comparison is only of relevance when one takes a frequentist's perspective on inference, that is, when one draws conclusions based on *P*-values.

6.4.2 Inference About Individual Terms

There are two final issues with regard to drawing inference based on *P*-values that deserve attention here. The first concerns inference about factors with more than two levels being represented by more than a single term in the model. As explained above, factors are usually dummy coded, revealing one dummy variable for each level of the factor except the reference level. Hence, a factor having, for instance, three levels will appear in the model output with two terms. As a consequence, one does not get an overall *P*-value for the effect of the factor, but two *P*-values, each testing the difference between the respective dummy coded level and the reference level. However, quite frequently, an overall *P*-value of the effect of the factor as a whole will be desired (particularly since the actual *P*-values shown are pretty much a random selection out of the possible ones since the reference category is frequently just the one which is the alphanumerically first). Such a test can be obtained using the same logic as for the full-null model comparison: A reduced model lacking the factor but comprising all other terms present in the full model is fitted and then compared with the full model using an F-test. If this test reveals significance, the factor under consideration has a significant impact on the response (see, e.g., Cohen and Cohen 1983 for details).

The other concerns inference of terms involved in interactions, for which interpretation must be made in light of the interaction they are involved in. Assume, for instance, a main effect involved in a two-way interaction: Its estimates (and the *P*-value associated with it) are conditional on the particular value of the other main effect it is interacting with. This can be seen when one considers the model equation with regard to the effects of the two predictors ('A' and 'B') interacting with one another, namely response = $c_0 + c_1 \times A + c_2 \times B +$ - $c_3 \times A \times B$. Since the effect of A on the response is represented by two terms in the model, one of which (c_3) modeling how its effect depends on the value of the other, c_1 models the effect of predictor A at predictor B having a value of zero (which would be its average if B were a *z*-transformed covariate or the reference category if B were a dummy coded factor). The same logic applies when it comes to the interpretation of two-way interactions involved in a three-way interaction and so on and also when it comes to the interpretation of linear effects involved in nonlinear effects like squared terms. This limited (precisely conditional) interpretation of terms involved in an interaction (or a nonlinear term) needs to be kept in mind when interpreting the results (see Schielzeth 2010 for a detailed account on the issue).

6.5 Transparency of the Analysis

A proper analysis requires a number of decisions. These include (but are not limited to) decisions about which main effects, interactions, and nonlinear terms are to be included in the model, potential transformations of predictors and/or the response, and subsequent *z*-transformations of the covariates. Furthermore, conducting a proper analysis means to conduct several checks of model validity and stability which might reveal good or bad results or something in between. All these decisions as well as the results of checks of model stability and validity are an integral and important part of the analysis and should be reported in the respective paper. Finally, also, the software used for the analysis should be mentioned (and in case of R being used also the key functions and the packages (including the version number) that provided them). Otherwise, the reader will be unable to judge the reliability of the analysis and cannot know to what extend the results can be trusted (see also Freckleton 2009). From my understanding, this means to (1) thoroughly outline the reasoning that took place when the model was formulated; (2) clearly formulate the model analyzed; (3) clearly describe the preparatory steps taken (e.g., variable transformations); and (4) clearly describe the steps taken to evaluate the model's assumptions and stability and what they revealed. Nowadays, there is usually the option to put extensive information into supplementary materials made available online, and we should use this option! Only by providing fully transparent analyses, we make our projects repeatable, and repeatability is at the core of science.

6.6 Concluding Remarks

What I have presented here are some of the steps that should be taken to verify the statistical validity and reliability of a PGLS model. I focused on issues and assumptions frequently checked for standard general linear models which begin with the formulation of a scientifically and logically adequate model and continue with its technically valid implementation. At the core of the establishment of the statistical validity and reliability of a model are issues regarding the number of estimated terms in relation to the number of cases, the distribution(s) of the predictor(s), absence of collinearity and influential cases, and assumptions about the residuals. Finally, drawing inference requires special care, and all the steps conducted to verify the model need a proper documentation. What I presented here are largely the steps taken to validate the reliability of a general linear model that I more or less simply transferred to PGLS. However, besides the steps I considered here, the validity of a PGLS model also most crucially depends on a variety of issues specific to PGLS (see introduction) and these must not be neglected neither.

With this chapter, I hope to create more attention for evaluations of whether the assumptions of a given model are fulfilled and to what extend a given model is stable and reliable. Only, once the assumptions of a model and its stability have been checked carefully, one can know how confident one can be about its results.

Acknowledgments First of all, I would like to thank László Zsolt Garamszegi for inviting me to write this chapter. I also thank László Zsolt Garamszegi and two anonymous reviewers for very helpful comments on an earlier draft of this chapter. I equally owe thanks to Charles L. Nunn for initially leading my attention to the need for and rationale of phylogenetically corrected statistical analyses. During the three AnthroTree workshops held in Amherst, MA, U.S.A., in 2010–2012 and supported by the NSF (BCS-0923791) and the National Evolutionary Synthesis Center (NSF grant EF-0905606) I learnt a lot about the philosophy and practical implementation of phylogenetic approaches to statistical analyses, and I am very grateful to have had the opportunity to attend them. This article was mainly written during a stay on the wonderful island of Læsø, Denmark, and I owe warm thanks to the staff of the hotel Havnebakken for their hospitality that made my stay very enjoyable and productive at the same time.

Glossary

Case Set of entries in the data referring to the same taxon; represented by one row in the data set and corresponds to one tip in the phylogeny.

Covariate Quantitative predictor variable.

Dummy coding Way of representing a factor in a linear model, by turning it into a set of 'quantitative' variables. One level of the factor is defined the 'reference' level (or reference category), and for each of the other levels a variable is created which is one if the respective case in the data set is of that level and zero otherwise. The estimate derived for a dummy coded variable reveals the degree by which the response in the coded level differs from that of the reference level.

Factor Qualitative (or categorical) predictor variable.

General linear model Unified approach to test the effect(s) of one or several quantitative or categorical predictors on a single quantitative response; makes the assumptions of normally and homogeneously distributed residuals; multiple regression, ANOVA, ANCOVA, and the t-tests are all just special cases of the general linear model.

Level Particular value of a factor (for instance, the factor 'sex' has the levels 'female' and 'male').

Predictor (variable) Variable for which its influence on the response variable should be investigated or controlled for; can be a factor or a covariate.

Response (variable) Variable being in the focus of the study and for which it should be investigated how one or several predictors influence it.

Right (left) skewed distribution Distribution with many small and few large values (a left skewed distribution shows the opposite pattern).

References

Aiken LS, West SG (1991) Multiple regression: testing and interpreting interactions. Sage, Newbury Park

Arnold C, Nunn CL (2010) Phylogenetic targeting of research effort in evolutionary biology. Am Nat 176:601–612

Budaev SV (2010) Using principal components and factor analysis in animal behaviour research: caveats and guidelines. Ethology 116:472–480

Burnham KP, Anderson DR (2002) Model selection and multimodel inference, 2nd edn. Springer, Berlin

Chatfield C (1995) Model uncertainty, data mining and statistical inference. J Roy Stat Soc A 158:419–466

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, New York

Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates Inc., New Jersey

Cooper N, Jetz W, Freckleton RP (2010) Phylogenetic comparative approaches for studying niche conservatism. J Evol Biol 23:2529–2539

Díaz-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

Díaz-Uriarte R, Garland T Jr (1998) Effects of branch length errors on the performance of phylogenetically independent contrasts. Syst Biol 47:654–672

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

Field A (2005) Discovering statistics using SPSS. Sage Publications, London

Forstmeier W, Schielzeth H (2011) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. Behav Ecol Sociobiol 65:47–55

Fox J, Monette G (1992) Generalized collinearity diagnostics. J Am Stat Assoc 87:178–183

Freckleton RP (2009) The seven deadly sins of comparative analysis. J Evol Biol 22:1367–1375

Freckleton RP (2011) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behav Ecol Sociobiol 65:91–101

Freckleton RP, Cooper N, Jetz W (2011) Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. Am Nat 178:E10–E17

Freckleton RP, Jetz W (2009) Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. Proc Roy Soc B—Biol Sci 276:21–30

Garamszegi LZ, Møller AP (2012) Untested assumptions about within-species sample size and missing data in interspecific studies. Behav Ecol Sociobiol 66:1363–1373

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

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Grafen A (1989) The phylogenetic regression. Phil Trans Roy Soc Lond B, Biol Sci 326:119–157

Grafen A, Ridley M (1996) Statistical tests for discrete cross-species data. J Theor Biol 183:225–267

Hansen TF (1997) Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Ives AR, Garland T Jr (2010) Phylogenetic logistic regression for binary dependent variables. Syst Biol 59:9–26

Martins EP, Diniz-Filho JAF, Housworth EA (2002) Adaptive constraints and the phylogenetic comparative method: a computer simulation test. Evolution 56:1–13

Mundry R (2011) Issues in information theory based statistical inference—a commentary from a frequentist's perspective. Behav Ecol Sociobiol 65:57–68

Nunn CL (2011) The comparative approach in evolutionary anthropology and biology. The University of Chicago Press, Chicago

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401:877–884

Polly PD, Lawing AM, Fabre A-C, Goswami A (2013) Phylogenetic principal components analysis and geometric morphometrics. Hystrix, Ital J Mammal 24:33–41

Quinn GP, Keough MJ (2002) Experimental designs and data analysis for biologists. Cambridge University Press, Cambridge

Ramsey PH (1980) Exact type 1 error rates for robustness of student's t test with unequal variances. J Educ Stat 5:337–349

R Core Team (2013) R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria

Revell LJ (2009) Size-correction and principal components for interspecific comparative studies. Evolution 63:3258–3268

Revell LJ (2010) Phylogenetic signal and linear regression on species data. Methods Ecol Evol 1:319–329

Rohlf FJ (2006) A comment on phylogenetic correction. Evolution 60:1509–1515

Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. Meth Ecol Evol 1:103–113

Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Meth Ecol Evol 1:3–14