R^2s for Correlated Data: Phylogenetic Models, LMMs, and GLMMs

Anthony R. Ives, Department of Zoology, UW-Madison, Madison, WI 53706 arives@wisc.edu

Abstract

Many researchers want to report an R^2 to measure the variance explained by a model. When the model includes correlation among data, such as mixed models and phylogenetic models, defining an R^2 faces two conceptual problems. (i) It is unclear how to measure the variance explained by predictor (independent) variables when the model contains covariances. (ii) Researchers may want the R^2 to include the variance explained by the covariances by asking questions such as "What is the partial R^2 for random effects in a linear mixed model?" or "How much of the variance is explained by phylogeny?".

I propose using three R^2s for mixed and phylogenetic models. A least-squares R^2_{ls} is an extension of the ordinary least-squares R^2 that weights residuals by variances and covariances estimated by the model. The likelihood ratio R^2_{lr} was first used by Cragg and Uhler (1970) for logistic regression, and here is used with the standardization proposed by Nagelkerke (1991). The conditional expectation R^2_{ce} is based on "predicting" each residual from the remaining residuals of the fitted model. These three R^2s can be formulated as partial R^2s to compare the contributions of mean components (fixed effects in mixed models and regression coefficients in phylogenetic models) and variance components (random effects and phylogenetic signal) to the fit of models.

I investigated the properties of the R^2s for linear and generalized linear mixed models (LMMs and GLMMs), and phylogenetic models for continuous and binary response data (PGLS

and phylogenetic logistic regression). For LMMs and GLMMs, I compared the R^2s to R^2_{glmm} from Nakagawa and Schielzeth (2013), and for LMMs also the ordinary least-squares R^2 treating random effects as fixed effects.

R^2_{ls}, R^2_{lr}, and R^2_{ce} have reasonable performance, and each has advantages and disadvantages for different applications. Overall, R^2_{lr} showed less variation among repeated simulations of the same model than R^2_{ls} and R^2_{ce} (and also R^2_{glmm}), making it the most precise estimate of goodness-of-fit. Nonetheless, all three can be used with a wide range of models for correlated data.

key-words: binomial regression, coefficient of determination, non-independent residuals, phylogenetic model, pseudo-likelihood

Introduction

Researchers often want to calculate a coefficient of determination, an R^2, to give a measure of the amount of variance in their data explained by a statistical model. For ordinary least-squares models (OLS), such as regression and ANOVA, the R^2 is simple to calculate and interpret. Many types of models, however, assume that the errors among response variables are correlated. Linear mixed models (LMMs) include random effects that generate correlation in the residual variation; for example, LMMs can account for correlation between residuals of experimental replicates within the same block. Similarly, phylogenetic generalized least squares models (PGLS) allow the possibility of phylogenetically related species being more similar to each other, leading to phylogenetic correlations in the errors. The situation is more complex for generalized linear mixed models (GLMMs) and phylogenetic logistic regression models (PLOG) in which the response variable is discrete. For models of discrete distributions, even perfectly

48 fitting models have residual variation due to the discreteness of the data, and this complicates the
49 interpretation of an R^2.

50 Correlated errors in statistical models cause two issues for defining an R^2. The first
51 involves assessing the goodness-of-fit of predictor variables (fixed effects) in terms of the
52 explained variance. For OLS models, the errors are assumed to be identical and independently
53 distributed, and therefore the variance in the residuals can be calculated directly to give the total
54 variance that is not explained by the model. In models for correlated data, however, the errors are
55 not independently distributed. To properly calculate an "explained variance", it is necessary to
56 incorporate the estimated covariances (Judge *et al.* 1985 p. 32).

57 The second issue for defining an R^2 involves assessing the goodness-of-fit of the
58 covariances (random effects) estimated in the model. For phylogenetic models, this is embodied
59 by the question "How much of the data is explained by phylogeny?" The difficulty is that a
60 phylogenetic model can be used to estimate the strength of phylogenetic signal (covariances) in
61 the errors, but the phylogenetic signal does not lead to predictions of the fitted data. Therefore, it
62 is not immediately clear what it means for a phylogeny to “explain” the data. This conceptual
63 issue also arises in mixed models, although it is more subtle. In some algorithms used to fit
64 LMMs, coefficients are estimated for each level of the random effect during the fitting, akin to
65 what would be done in OLS if the random effect was treated as a fixed effect. In LMMs, it is
66 possible to use these coefficients to estimate residual variances that are not captured by the
67 variances in the random effects (Nakagawa & Schielzeth 2013). Nonetheless, the random effects
68 of LMMs are still mathematically given by covariances in the model. This contrasts an OLS
69 model in which R^2 is calculated by minimizing the unexplained variance in the data. Thus, the
70 R^2s from LMMs and OLS models measure subtly different things. The conceptual issue facing

LMM and PGLS of explaining the data from variances estimated in the model is more complicated for GLMM and PLOG in which even perfectly fitting models have residual variation.

Here, I address the problem of defining and calculating R^2s for models with fitted parameters governing covariances. The general form of the models is

$$Y_i \sim \mathcal{F}(\mu_i)$$

$$g(\mu_i) = \beta_0 + \beta_1 x_i + e_i$$

$$e \sim \text{Gaussian}(\mathbf{0}, \sigma^2\mathbf{\Sigma}(\theta)) \qquad \text{eqn 1}$$

where data Y_i ($i = 1, ..., n$) are distributed by a member $\mathcal{F}$ of the exponential family of distributions (McCullagh & Nelder 1989). The parameter μ_i of distribution $\mathcal{F}$ is itself a random variable, and applying the link function $g()$ to μ_i gives a linear equation in terms of the predictor variable x_i and an error term e_i. The error term e_i has a multivariate Gaussian distribution with means 0 and covariance matrix $\sigma^2\mathbf{\Sigma}(\theta)$ that depends on a vector of parameters θ. This general model form produces GLMMs (and LMMs as a special case) when the random effects are contained as block-diagonal elements in the covariance matrix $\sigma^2\mathbf{\Sigma}(\theta)$ (Gelman & Hill 2007); for GLMMs, the parameters θ governing the covariances are the variances of the random effects. In phylogenetic models, $\sigma^2\mathbf{\Sigma}(\theta)$ contains the phylogenetic covariance among species given by their evolutionary relatedness (Lavin *et al.* 2008); the parameters θ govern the strength of phylogenetic signal. For simplicity, equation 1 only includes a single predictor variable x and

92 single variance parameter θ. Nonetheless, multiple regression and multiple parameters θ can be
93 included in the obvious way and are allowed in the accompanying R code.

94 I derive three different R^2s for models given by equation 1. Because the models can
95 contain multiple parameters, these R^2s are derived to compare a full model with a reduced model
96 in which one or more of the parameters are removed; thus, they are partial R^2s that give the
97 explained variance by the components that differ between full and reduced models. The total R^2s
98 are obtained by selecting the reduced model in which there is only an intercept and residuals are
99 independent.

100 R^2s can be assessed on multiple grounds (Kvalseth 1985), and here I use three. First, does
101 the R^2 give a good measure of fit of a model to data? To serve as a basis for assessment, I use the
102 log-likelihood ratio (LLR) of the full and reduced models. The LLR approaches a χ^2 distribution
103 for large samples and is therefore used for hypothesis tests of full vs. reduce models (Judge *et al.*
104 1985). Also, the LLR is linearly related to the AIC and other measures used for model selection
105 (Burnham & Anderson 2002). Therefore, the LLR is a natural choice to assess R^2s: a good R^2
106 should be monotonically related to the LLR. Second, can the R^2 separate the contribution of
107 different components of the model to the overall model fit? For the simple case of equation 1 in
108 which there is only a single regression coefficient (β_1) and a single variance parameter (θ), I ask
109 whether the R^2s can distinguish between the two in their contributions to the fit of the model.
110 Although not done here, the R^2s could also be used to sort among multiple regression coefficients
111 or variance parameters. Third, does the R^2 give similar values when applied to data generated by
112 the same statistical process? If the values of R^2 when applied to data generated from the same
113 statistical process are all similar, then the R^2 gives a precise measure of goodness-of-fit.

I assess the R^2s using four special cases of equation 1: LMM, PGLS, GLMMs for binary (binomial) data, and PLOG. Functions in the statistical computing language R code are provided for the three R^2s that can be applied to fitted models of classes lmerMod and glmerMod {lme4}, phylolm and phyloglm {phyloglm}, and binaryPGLMM {ape}.

Materials and Methods

R^2_{ls} is derived from generalized least-squares (GLS) and therefore has a close conceptual tie to the standard R^2 from OLS. R^2_{lr} is the application of an R^2 proposed for logistic regression (Cragg & Uhler 1970; Maddala 1983; Cox & Snell 1989) and generalized by Magee (1990) and Nagelkerke (1991). It is based on the likelihood ratio between the full and reduced models. R^2_{lr} is closely related to R^2_{ls}, because for linear models they differ only by the way in which they are scaled. R^2_{ce} is based on the conditional expectations of new data points given the full versus reduced models. For comparison with these models in application to LMMs and GLMMs, I also consider R^2_{glmm} proposed by Nakagawa and Schielzeth (2013), and for LMMs the standard R^2_{ols} in which random effects are treated as fixed effects. I know of no existing R^2s that have been applied for phylogenetic models that can be used to compare with the three proposed R^2s.

R^2_{ls}

For linear models with correlated errors, the R^2 can be calculated from GLS as

$$R^2_{ls} = 1 - \frac{\mathrm{MSE}_f\left(\hat{\theta}_f\right)}{\mathrm{MSE}_r\left(\hat{\theta}_r\right)} \qquad \text{eqn 2}$$

where MSE_f is the mean squared errors for the full model, and MSE_r is for the reduced model. Both full and reduced models may contain parameters in vectors θ_f and θ_r that involve the variances and covariances among samples. For a GLS model

$$\mathrm{MSE}\left(\hat{\theta}\right) = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)' \mathbf{V}\left(\hat{\theta}\right)^{-1} \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) \qquad \text{eqn 3}$$

where $\mathbf{Y}$ is the $n \times 1$ vector of response values Y_i, $\mathbf{X}$ is the $n \times p$ matrix for p predictor variables (including the intercept), $\hat{\beta}$ is the $1 \times p$ vector of estimated regression coefficients (fixed effects) that may depend on θ (for discrete distributions), and $\mathbf{V}\left(\hat{\theta}\right)^{-1}$ is the inverse of the $n \times n$ matrix $\mathbf{V}(\theta)$ that contains the variances and covariances of the errors. For many models, $\mathbf{V}(\theta)$ will depend on estimated parameters $\hat{\theta}$, and therefore equation 3 technically gives the MSE of an estimated generalized linear model (EGLS Judge *et al.* 1985). The MSE for OLS models is the special case in which $\mathbf{V}(\theta) = \mathbf{I}$, the $n \times n$ identity matrix, which gives the standard R^2.

Setting $\mathbf{V}(\theta) = \mathbf{\Sigma}(\theta)$, the MSE gives an estimate of the variance term σ^2 from equation 1. However, $\mathbf{V}(\theta)$ can be scaled by a constant without changing the fit of the statistical model; the only effect of scaling $\mathbf{V}(\theta)$ by a constant is to change the value of σ^2 by 1/constant. When comparing full and reduced models, it will generally be the case that $\mathbf{V}(\hat{\theta}_f) \neq \mathbf{V}(\hat{\theta}_r)$; for example, even for LMMs including the same random effects, $\mathbf{V}(\hat{\theta}_f) \neq \mathbf{V}(\hat{\theta}_r)$ if removing fixed effects from the full model changes the estimated variances of the random effects in the reduce model. Therefore, the calculation of the GLS R^2 depends on how $\mathbf{V}(\theta)$ is scaled.

For LMMs, a natural scaling of $\mathbf{V}(\theta)$ is to let $\mathbf{V}(\theta) = \mathbf{I} + \mathbf{G}(\theta)$ where $\mathbf{G}(\theta)$ is the block-diagonal matrix containing the variances of the random effects divided by the residual variance. In this case, the MSE in equation 2 is the estimate of the residual variance in a LMM under maximum likelihood estimation. For phylogenetic models, $\mathbf{V}(\theta)$ can be scaled by dividing all elements in the matrix by the sum of the branch lengths of the phylogenetic tree used to derive $\mathbf{V}(\theta)$. This standardization means that $\mathbf{V}(\hat{\theta}_f)$ and $\mathbf{V}(\hat{\theta}_r)$ represent the same total amount of independent phylogenetic divergence, since the rescaled phylogenies have the same total branch lengths. Standardizing by summed branch lengths is a reasonable convention, and it produces sensible values of R^2.

For non-Gaussian models, it is necessary to account for the variation introduced by discrete data. This can be done by defining

$$\text{MSE} = \frac{1}{n}\left(\mathbf{Y} - \hat{\mu}\right)' \mathbf{A}\left(\hat{\mu}\right)^{-1/2} \mathbf{V}\left(\hat{\theta}\right)^{-1} \mathbf{A}\left(\hat{\mu}\right)^{-1/2} \left(\mathbf{Y} - \hat{\mu}\right) \qquad \text{eqn 4}$$

where $\hat{\mu} = g^{-1}\left(\mathbf{X}\hat{\beta}\right)$ are the fitted values of μ, and $\mathbf{V}(\hat{\theta}) = \hat{\sigma}^2\mathbf{\Sigma}(\hat{\theta}) + \mathbf{I}$ is estimated when fitting the model (Schall 1991; Breslow & Clayton 1993; Ives & Helmus 2011). The matrix $\mathbf{A}$ has diagonal elements given by the variance function $v(\mu)$ for the link function $g()$; for example, a binomial model with a logit link function will have $v(\mu) = \mu(1-\mu)$. Equation 4 can be interpreted as the MSE for the distribution of e in equation 1, with the residuals transformed to have a variance of 1. When $\hat{\sigma}^2 = 0$, the model reduces to a GLM, and equation 4 becomes the quasi-likelihood score function that can be used to estimate GLM parameters (McCullagh & Nelder

177 1989). For models for discrete data, **V**(θ) should be standardized in the same way as for models
178 for continuous data.
179
180 R^2_{lr}
181 For Gaussian models, R^2_{lr} differs from R^2_{ls} only in the scaling of **V**(θ). If **V**(θ) is scaled
182 so that the determinant det(**V**(θ)) = 1, then the maximum log likelihood is
183
184 $$logLik\left(\hat{\theta}\right) = -\frac{n}{2}\left(\log\left(2\pi \mathrm{MSE}\left(\hat{\theta}\right)\right)+1\right). \qquad \text{eqn 5}$$
185
186 Substituting into equation 2 then leads to
187
188 $$R^2_{lr} = 1-\exp\left(\frac{-2}{n}\left(logLik\left(\hat{\theta}_f\right)-logLik\left(\hat{\theta}_r\right)\right)\right). \qquad \text{eqn 6}$$
189
190 This definition of R^2_{lr} in terms of likelihoods extends immediately to models for discrete data.
191 However, for discrete data, equation 6 does not have a maximum of 1, because the maximum
192 attainable log-likelihood for discrete data is zero. Therefore, Nagelkerke (1991) and Cameron
193 and Windmeijer (1997) proposed dividing by the maximum attainable value, which is equation 6
194 with $logLik\left(\hat{\theta}_f\right) = 0$; throughout, I have used this Nagelkerke standardization. The R code in the
195 supplement computes R^2_{lr}, which is also computed for a range of models in the MuMIn package
196 of R (Barton 2016). The deviance $2\left(logLik\left(\hat{\theta}_f\right)-logLik\left(\hat{\theta}_r\right)\right)$ is approximately χ^2 distributed
197 with degrees of freedom equal to the number of parameters differing between full and reduced

models, and this establishes a direct link between R^2_{lr} and a formal test of goodness-of-fit. Note that because R^2_{lr} and R^2_{ls} scale $\mathbf{V}(\theta)$ differently, their values will differ.

R^2_{ce}

From a fitted Gaussian model in which $\mathbf{V}(\hat{\theta})$ is estimated, it is possible to compute the expected value of each residual assuming that the residuals for the other data points are known. Specifically, for the general form of equation 3, the expected value of residual $z_i = y_i - \hat{y}_i$ from the remaining residuals $\mathbf{Z}_{[-i]}$ is

$$\hat{z}_i = \bar{z} + \mathbf{V}_{[i,-i]}\mathbf{V}^{-1}_{[-i,-i]}\left(\mathbf{Z}_{[-i]} - \bar{z}\right), \qquad \text{eqn 7}$$

where $\bar{z}$ is the GLS mean of the residuals, $\mathbf{V}_{[i,-i]}$ is row i of $\mathbf{V}$ with column i removed, and $\mathbf{V}_{[-i,-i]}$ is $\mathbf{V}$ with row i and column i removed (Petersen & Pedersen 2012). The MSE is then the variance of the residuals after updating the estimates $\hat{y}$: $\text{var}\left(\mathbf{Y} - \left(\hat{y} + \hat{\mathbf{Z}}\right)\right)$. For discrete data, it would seem natural to use equation 7 with $\mathbf{V}$ replaced by $\mathbf{A}^{-1/2}\mathbf{V}\mathbf{A}^{-1/2}$ as in equation 4; however, in the simulations performed here, this approach led to occasional values of R^2_{ce} far below zero, and therefore I used $\mathbf{V}$ even for non-Gaussian models.

Alternative methods

For LMMs and GLMMs, R^2_{ls}, R^2_{lr} and R^2_{ce} can be compared to R^2_{glmm} given by Nakagawa and Schielzeth (2013), and for LMMs comparison can also be made to R^2_{ols} that treats the random effects in the LMM as fixed effects.

From Nakagawa and Schielzeth (2013), the conditional R^2 for the model in equation 1 is

$$R^2_{glmm(c)} = \frac{\hat{\sigma}^2_f + \hat{\sigma}^2_l}{\hat{\sigma}^2_f + \hat{\sigma}^2_l + \hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_d} \qquad \text{eqn 8}$$

where $\hat{\sigma}^2_f$ is calculated from the fixed effects, $\hat{\sigma}^2_l$ is calculated from the random effects, $\hat{\sigma}^2_\varepsilon$ is the residual variance, and $\hat{\sigma}^2_d$ is the distribution-specific variance. This corresponds to the total R^2 that gives the proportion of residual variance explained by the fixed and random effects. For discrete GLMMs, $R^2_{glmm(c)}$ with this formulation never reaches 1, because $\hat{\sigma}^2_d$ is never zero. The marginal $R^2_{glmm(m)}$ gives the proportion of the variance explained by only the fixed effects and is given by equation 8 after removing $\hat{\sigma}^2_l$ from the numerator. Note that the marginal $R^2_{glmm(m)}$ is not equivalent to the partial R^2 for the fixed effects; a partial R^2 would refit the GLMM without the fixed effects as a reduced model, giving new variances $\hat{\sigma}^2_l$ for the random effects. To give a comparable measure to $R^2_{glmm(m)}$ for the proportion of the variance explained by the random effects, I will define $R^2_{glmm(v)}$ as equation 8 after removing $\hat{\sigma}^2_f$ from the numerator.

To calculate OLS R^2s, LMMs can be converted to LMs by treating the random effects as fixed effects; I then applied adjusted partial R^2s from OLS to give R^2_{ols}.

Simulations

The simulations to explore LMM, PGLS, GLMM, and PLOG from equation 1 all follow the same strategy. For each, data were simulated when there is only a fixed effect ($\beta_1 > 0$, $\theta = 0$), only a random effect ($\beta_1 = 0$, $\theta > 0$), and when there is both ($\beta_1 > 0$, $\theta > 0$). For each case, the

model parameters were the same for all simulations, so that variation in values of a given R^2 among datasets is caused by random sampling from the same statistical process. For example, for the LMM with $\beta_1 > 0$ and $\theta = 0$, the model $y_i = \beta_0 + \beta_1 x_i + e_i$ with $e \sim \text{Gaussian}(\mathbf{0}, \sigma^2\mathbf{I})$ was simulated repeatedly for the same values of β_0, β_1, and σ^2.

For LMM, data were simulated with the model

$$y_i = \beta_0 + \beta_1 x_i + b u_i + \phi_i \qquad \text{eqn 9}$$

where x_i follows a Gaussian distribution with mean 0 and variance 1, and the random effect u_i has 10 levels, with b following a normal distribution with mean 0 and variance θ. I selected parameter values to generate moderate R^2 values. When there is a fixed effect, $\beta_1 = 1$, and when there is a random effect, $\theta = 1.5$. The variance of the residual term ϕ_i is 1. For GLMM data, I used a binomial (binary) model with logit link function $g()$ having the same structure as the LMM. Values for the fixed and random effects were $\beta_1 = 1.8$ and $\theta = 1.8$, and there was no residual variation, $\phi_i = 0$. Models were fit using lmer and glmer in the lme4 package of R (Bates *et al.* 2014).

For the PGLS model, to obtain the covariance matrix $\mathbf{\Sigma}(\theta)$ in equation 1, I first simulated random phylogenetic trees using the rtree function of the ape package of R (Paradis, Claude & Strimmer 2004), standardizing the base-to-tip lengths to be 1. Thus, a different tree was simulated for each dataset. Under the assumption of Brownian motion (BM) evolution, the expected covariance in trait values between two tips is given by the height of the most recent common node (ancestor), and from this it is possible to construct the covariance matrix $\mathbf{\Sigma}_{\text{BM}}$ (Martins & Hansen 1997; Blomberg, Garland & Ives 2003). For PGLS simulations, the strength

of phylogenetic signal was varied using Pagel's lambda transform, $\mathbf{\Sigma}(\lambda) = (1 - \lambda)\mathbf{I} + \lambda\mathbf{\Sigma}_{BM}$, in which values of $\lambda = 0$ imply no phylogenetic correlations and $\lambda = 1$ recovers $\mathbf{\Sigma}_{BM}$. Values of x_i were simulated under the BM assumption using the rTraitCont function (Paradis, Claude & Strimmer 2004). Values of the regression coefficient (fixed effect) and phylogenetic signal (random effect) were $\beta_1 = 1.5$ and $\theta = \lambda = 0.7$. The simulated data were fit using penalized maximum likelihood with the function phylolm assuming a Pagel's lambda transformation in the package phyloglm in R (Ho & Ane 2014).

The PLOG model was similar to the PGLS model. In contrast to the PGLS, however, the predictor variable x_i was assumed to be independently distributed; including phylogenetic signal in x_i caused challenges for model fitting for some simulated datasets, making the simulation studies difficult. Phylogenetic signal in the residuals e_i was controlled by setting $\mathbf{\Sigma}(\lambda) = \lambda\mathbf{\Sigma}_{BM}$ so that in the absence of phylogenetic signal ($\lambda = 0$) the simulations conformed to a simple logistic regression model. Values of the regression coefficient and phylogenetic signal were $\beta_1 = 1.5$ and $\theta = 2$. To simulate binary data, a logit link function was used in equation 1. To obtain maximum likelihood values, the simulations were fit using a modified version of the function phyloglm (Ho & Ane 2014) in which Nelder-Mead optimization was used; Nelder-Mead optimization was more likely to find the maximum likelihood than the built-in optimizer. Fitting with the modified phyloglm was performed using Firth penalized maximum likelihood, although the regular maximum likelihoods were used to compute R^2_{lr}. For R^2_{ls} and R^2_{ce}, the simulated data were fit using binaryPGLMM (Ives & Garland 2014) in the ape package (Paradis, Claude & Strimmer 2004).

Results

The R^2s were assessed according to the three properties: (i) their ability to measure goodness-of-fit as benchmarked by the LLR of full model and the model with only an intercept, (ii) whether they can partition sources of variation in the model, and (iii) how precise is their inference about goodness-of-fit. Property (iii) treats the R^2s as if they were estimators of goodness-of-fit and asks how variable are the estimates when applied to repeated simulations from the same model. R^2_{ls}, R^2_{lr} and R^2_{ce} are applied to all simulations, while R^2_{glmm} can only be applied to LMMs and GLMMs, and R^2_{ols} is only applied to LMMs. A more comprehensive treatment is given in the Supplement and figures S1-S12.

Goodness-of-fit

Figure 1 plots the total R^2s against the corresponding LLR. All R^2s were positively related to the LLR, which is a minimum requirement for an R^2. R^2_{lr} shows a monotonic relationship with LLR, which is necessarily the case due to the definition of R^2_{lr} (eqns 5, 6). For the remaining R^2s, values for a given LLR were generally lower for simulations in which variation was produced only by the fixed effect ($\beta_1 > 0$, $\theta = 0$; Fig. 1, blue circles). This implies that, relative to the LLR, these R^2s were attributing less "explained" variance to fixed effects than random effects.

For the LMM, R^2_{ls}, R^2_{glmm} and R^2_{ols} were almost identical (Fig. S2). This correspondence suggests that R^2_{ls} gives an R^2 that is comparable to R^2_{glmm} and R^2_{ols} but generalizes to models that do not have block-diagonal covariance matrices that underlie the random effects in LMMs. Thus, R^2_{ls} for PGLS is comparable to R^2_{glmm} for LMM. This comparison, however, has to be made with the caution that R^2_{ls} applied to PGLS requires an assumption about the scaling of the covariance

matrix $\mathbf{V}(\theta)$ (eqn 3) which will affect its value. Because R^2_{lr} is based on likelihoods, it gives a comparison between LMM and PGLS that is not conditional upon scaling decisions.

All of the R^2s other than R^2_{lr} showed greater scatter in their relationships with LLR for the simulations of binary data (GLMM and PLOG). In part, this is due to the difficulty of estimating variance parameters θ in binomial models. For example, there is more scatter in R^2_{glmm} for GLMM simulations than LMM simulations. The scatter seems particularly large for R^2_{ls} applied to PLOG simulations, although this case requires some technical discussion. For PLOG, the LLR was obtained from phyloglm using penalized maximum likelihood, whereas the variance parameter θ was estimated from binaryPGLMM using the pseudo-likelihood. The penalized ML estimate of phylogenetic signal tended to absorb at zero even when the pseudo-likelihood estimate of θ was positive; therefore, R^2_{ls} could be positive even when the LLR was zero. Comparison between penalized maximum likelihood and pseudo-likelihood estimation for phylogenetic logistic regression shows that they have similar performances but do not necessarily give the same conclusions about the presence of phylogenetic signal for the same dataset (Ives & Garland 2014). This contrast between fitting methods is not the only thing that underlies the scatter in R^2_{ls}, however, because R^2_{ce} uses the same estimate of θ as R^2_{ls} but has less scatter.

Partitioning sources of variation

The partial R^2_{ls}, R^2_{lr}, and R^2_{ce} were generally able to partition sources of variation between components of a model, in particular between regression coefficients (fixed effects) and covariance parameters (random effects). Simulations with $\beta_1 > 0$ and $\theta = 0$ should have partial R^2s for β_1 that are positive and partial R^2s for θ that are zero (blue circles, Fig. 2). Simulations

with $\beta_1 = 0$ and $\theta > 0$ should have partial R^2s for β_1 that are zero and partial R^2s for θ that are positive (red triangles, Fig. 2). Simulations with $\beta_1 > 0$ and $\theta > 0$ should have both partial R^2s positive (black x's, Fig. 2). Because the values of β_1 and θ were the same whether or not the other was zero, the partial R^2s for β_1 should be the same for simulations with $\theta = 0$ (blue circles) as for simulations with $\theta > 0$ (black x's), and the partial R^2s for θ should similarly be the same for $\beta_1 = 0$ (red triangles) and $\beta_1 > 0$. Among the R^2s, the worst performance was R^2_{ls} applied to PGLS, in which the partial R^2 for β_1 differed between the cases of $\theta = 0$ and $\theta > 0$. All three R^2s shows a lot of scatter for GLMM and PLOG, which in large part is due to the statistical challenge of estimating regression coefficients and variance parameters from discrete data. This is seen, for example, in the GLMM and PLOG simulations with $\beta_1 > 0$ and $\theta > 0$ in which the partial R^2_{lr} for θ was zero (black x's); these cases occur when the estimate of θ was zero even though a non-zero value was used in the simulations.

The case of R^2_{glmm} is distinct, because rather than use partial R^2s, I used the marginal $R^2_{glmm(m)}$ provided by Nakagawa & Schielzeth (2013) and the comparable $R^2_{glmm(v)}$ for the random effects. A more appropriate comparison would be with a partial R^2_{glmm} (see Discussion), although this has not been presented previously in the literature. By construction, $R^2_{glmm(m)}$ and $R^2_{glmm(v)}$ add up to $R^2_{glmm(c)}$, and this generates the negative correlation between them when $\beta_1 > 0$ and $\theta > 0$ in the simulations, which is especially visible for the LMM (Fig. 2).

Inference about underlying process

To summarize the ability of R^2s to infer the fit of the statistical process to the model, I plotted the mean value with 66% and 95% inclusion intervals for simulated datasets with sample sizes 40, 60, …, 160 (Fig. 3). For LMM and GLMM, there were 10 levels of the random effect;

355 datasets were produced by first simulating 160 samples (16 replicates at each level) and then
356 randomly removing two replicates at each level to reduce the sample size in steps of 20. For
357 PGLS and PLOG, each dataset at each sample size was simulated independently.
358 For LMM simulations, R^2_{ls}, R^2_{glmm} and R^2_{ols} showed similar patterns (Fig. 3), reflecting
359 the fact that they give very similar values (Fig. 1, S2). Mean values did not change with sample
360 size, and there was only moderate increase in variability among simulations with decreasing
361 sample size. In contrast, mean values of R^2_{lr} and R^2_{ce} decreased with decreasing sample size. For
362 R^2_{lr} this probably reflects the information that is lost when estimating the model parameters, in
363 the same way that information (degrees of freedom) is lost in OLS causing the non-adjusted R^2_{ols}
364 to decrease with sample size. For R^2_{ce} this occurs because smaller sample size decreases the
365 information available to estimate a residual from the other data points. This can be illustrated
366 with the *reductio ad absurdum* case of a sample size of two, in which the best estimate of one
367 residual is the value of the other residual; this will lead to a negative R^2_{ce}. This happens not only
368 with a sample size of two, but also when there are only two values at each level of a random
369 effect in a LMM. In contrast to LMM simulations, the PGLS simulations showed less change in
370 the means of R^2_{lr} and R^2_{ce} with sample size, presumably because there were more covariances
371 among samples (i.e., the covariance matrix had more non-zero elements) than in the LMM with
372 few replicates per level.
373 For the GLMM, both R^2_{ls} and R^2_{glmm} had higher variances (less precision) than R^2_{lr}. The
374 results for R^2_{ls} were actually worse than shown by figure 3, because I omitted occasional values
375 that were much less than -1. These errant values of R^2_{ls} often occurred when the estimate of the
376 random effect variance $\hat{\sigma}^2_l$ (eqn 8) was very large. These very large estimates of $\hat{\sigma}^2_l$ also caused
377 errant values of $R^2_{glmm} = 1$. The greater variation in values of R^2_{ls} and R^2_{glmm} compared to R^2_{lr}

occurs because R^2_{ls} and R^2_{glmm} depend on estimates of $\hat{\sigma}^2_l$ while R^2_{lr} depends on likelihoods. Thus, R^2_{ls} and R^2_{glmm} are compromised when the estimates of the random effects are poor, as is particularly the case when sample sizes are small. Even though R^2_{ce} is also calculated using $\hat{\sigma}^2_l$, it is not as variable as R^2_{ls} and R^2_{glmm}. This is at least in part because the observation-level variance contained in the matrix **A** (eqn 4) was ignored when calculating R^2_{ce}. For PLOG, values of R^2_{ls} were very rarely negative (2/7000 simulations), and the variation in R^2_{ls} was not much greater than R^2_{lr} and R^2_{ce} (Fig. 3). This is likely because estimates of phylogenetic signal ($\lambda = \theta$) were well-bounded, in contrast to $\hat{\sigma}^2_l$ in the GLMMs.

Discussion

R^2_{ls}, $R^2_{lr,}$ and R^2_{ce} are broadly applicable, easy to implement, and often perform as well or better than previous methods designed for more specialized cases. Below, I first address their specific application to the simulation model considered here, and then give general recommendations.

Applications to LMM, PGLS, GLMM and PLOG

For LMMs, all R^2s had good performance (Table 1). R^2_{ls} gave very similar values to R^2_{ols} computed by treating random effects as fixed effects, and this correspondence to familiar and easily understood OLS argues for using R^2_{ls}. Nonetheless, R^2_{lr} weights the fixed and random effects according to LLRs, and therefore partitioning the contribution of fixed and random effects to the total R^2 is done in a way that can be directly related to hypothesis tests. R^2_{ls}, $R^2_{lr,}$ and R^2_{ce} also had good performance for PGLS. While either R^2_{ls} or R^2_{lr} are reasonable choices,

400 R^2_{ce} has the advantage of addressing how much of the data is "explained by the phylogeny." The
401 disadvantage of R^2_{ce}, however, is that it can be negative for small sample sizes.

402 GLMM and PLOG were more problematic, in large part because of challenges estimating
403 parameters in GLMM and PLOG models. This is not a problem with the R^2s *per se*, but relative
404 insensitivity of R^2s to parameter estimates is an advantage. R^2_{ls} was more sensitive to variation in
405 parameter estimates than R^2_{ce}, leading to greater variation in R^2_{ls} than R^2_{ce} (Fig. 3). R^2_{lr} was the
406 most precise, presumably because it uses likelihoods rather than parameter estimates. All three
407 R^2s, however, were dependent on the model fitting to partition between regression coefficients
408 (fixed effects) and variance parameters (random effects), with considerable scatter produced for
409 all R^2s (Fig. 2). A lesson from these results is that if estimates of variance parameters (random
410 effects) are poor, then R^2s are likely to be of questionable value.

411 For LMMs, the conditional $R^2_{glmm(c)}$ gave almost identical values to the total R^2_{ls} and R^2_{ols}
412 (when the reduced model contained only the intercept). However, instead of partial R^2s to
413 compare with R^2_{ls} and R^2_{ols}, I used the marginal $R^2_{glmm(m)}$ and its counterpart for random effects,
414 $R^2_{glmm(v)}$. By construction, these add up to the conditional $R^2_{glmm(c)}$, and this necessarily generates
415 negative association between $R^2_{glmm(m)}$ and $R^2_{glmm(v)}$ when partitioning components of variation in
416 models (Fig. 2). The conceptual advantage of partial R^2s is that they give the improvement in the
417 fit of the full model relative to the reduced model; they answer "How much better does the model
418 fit when including this parameter?" It is simple to define a partial R^2_{glmm} for either fixed or
419 random effects by comparing full and reduced as

420

421 $$R^2_{glmm.partial} = 1 - \frac{1 - R^2_{glmm(c).full}}{1 - R^2_{glmm(c).reduced}} \qquad \text{eqn 10}$$

422

Using this partial R^2_{glmm} also adds flexibility to compare combinations of fixed and random effects, as well as more-complex random effects such as random slope models (Johnson 2014).

Recommendations

An ideal R^2 would make it possible to compare among different models and among different methods used to fit the same model (Kvalseth 1985 properties of a good R2 #4 and #5). R^2_{ls} and R^2_{ce} can be used for any model and fitting method that estimates the covariance matrix; for example, they could be used to compare LMMs fit with ML vs. REML, or binary phylogenetic models fit with ML or quasi-likelihood (binaryPGLMM). Nonetheless, R^2_{ls} and R^2_{ce} have a disadvantage in terms of generality. For correlated data a decision must be made about how to weight the covariance matrix $\mathbf{V}(\theta)$ (eqn 3). The conventions I used for LMMs and PGLS differed, making in unclear how the R^2s from LMM compare to the R^2s from PGLS. In contrast, R^2_{lr} is restricted to models that are fit with ML estimation; however, if ML is used for fitting, then values of R^2_{lr} can be compared across different types of models. This applies to any type of data and model fit with ML estimation.

An ideal R^2 should also be intuitive (Kvalseth 1985 property #1). However, intuitive is in the eye of the beholder. R^2_{ls} is the most similar to R^2_{ols}, which grounds R^2_{ls} in the familiar and intuitive OLS framework. R^2_{lr} is also related to R^2_{ols}: in LMMs and PGLS, R^2_{lr} only differs from R^2_{ls} by the way in which the covariance matrix $\mathbf{V}(\theta)$ (eqn 3) is scaled, and this provides a link between R^2_{lr} and R^2_{ols} through R^2_{ls}. R^2_{ce} "predicts" the data from covariances estimated in the model, and therefore could be viewed as the most intuitive way to relate the variance explained by regression coefficients (fixed effects) to that explained by variance parameters (random

445 effects). This said, however, I suspect that different researchers would rank the intuitiveness of
446 R^2_{ls}, R^2_{lr}, and R^2_{ce} differently.

447 R^2s are often used as "summary statistics" to describe the fit of a model to data in a way
448 that does not involve statistical inference about the underlying stochastic process that generated
449 the data: "How does the model fit these data?" rather than "How much does the model infer
450 about the process that generated the data?" Should R^2s be judged as a summary statistic? I think
451 not. All the R^2s showed high variation among simulations of the same model with the same
452 parameters, especially when sample sizes were small (Fig. 3). This means that how the model fits
453 a specific data set involves a lot of chance, and hence one should not get too excited about a high
454 R^2, or too discouraged about a low one. R^2s are best treated as inferential statistics, that is, as
455 functions of a data-generating process that are themselves random variables. As an inferential
456 statistic, R^2_{lr} outperformed R^2_{ls} and R^2_{ce} – and also R^2_{glmm} – for models with discrete data, since
457 R^2_{lr} was more precise (less variable). For me, this tips the balance to favor R^2_{lr} over the others.
458

459 **Acknowledgments**

460 I thank Ted Garland and Joe Phillips for wonderfully insightful comments that helped to
461 clarify this article. Financial support came from the US National Science Foundation, DEB-
462 LTREB-1052160 and DEB-1240804.
463

464 **References**

465 Barton, K. (2016) MuMIn: Multi-model inference.
466 Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014) lme4: Linear mixed-effects models
467 using Eigen and S4.

468 Blomberg, S.P., Garland, T., Jr. & Ives, A.R. (2003) Testing for phylogenetic signal in
469 comparative data: behavioral traits are more labile. *Evolution,* **57,** 717-745.
470 Breslow, N.E. & Clayton, D.G. (1993) Approximate inference in generalized linear mixed
471 models. *Journal of the American Statistical Association,* **88,** 9-25.
472 Burnham, K.T. & Anderson, D.R. (2002) *Model selection and inference: a practical*
473 *information-theoretic approach,* Second edn. Springer, New York, New York.
474 Cameron, A.C. & Windmeijer, F.A.G. (1997) An R-squared measure of goodness of fit for some
475 common nonlinear regression models. *Journal of Econometrics,* **77,** 329-342.
476 Cox, D.R. & Snell, E.J. (1989) *The analysis of binary data*. Chapman and Hall, London, UK.
477 Cragg, J.G. & Uhler, R.S. (1970) Demand for automobiles. *Canadian Journal of Economics,* **3,**
478 386-406.
479 Gelman, A. & Hill, J. (2007) *Data analysis using regression and multilevel/hierarchical models*.
480 Cambridge University Press, New York, NY.
481 Ho, L.S.T. & Ane, C. (2014) A linear-time algorithm for Gaussian and non-Gaussian trait
482 evolution models. *Systematic Biology,* **63,** 397-408.
483 Ives, A.R. & Garland, T., Jr. (2014) Phylogenetic regression for binary dependent variables.
484 *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary*
485 *Biology* (ed. L.Z. Garamszegi), pp. 231-261. Springer-Verlag, Berlin Heidelberg.
486 Ives, A.R. & Helmus, M.R. (2011) Generalized linear mixed models for phylogenetic analyses of
487 community structure. *Ecological Monographs,* **81,** 511-525.
488 Johnson, P.C.D. (2014) Extension of Nakagawa & Schielzeth's R-GLMM(2) to random slopes
489 models. *Methods in Ecology and Evolution,* **5,** 944-946.

490 Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H. & Lee, T.-C. (1985) *The theory and*
491 *practice of econometrics,* Second edn. John Wiley and Sons, New York.
492 Kvalseth, T.O. (1985) Cautionary note about R2. *American Statistician,* **39,** 279-285.
493 Lavin, S.R., Karasov, W.H., Ives, A.R., Middleton, K.M. & Garland, T., Jr. (2008)
494 Morphometrics of the avian small intestine, compared with non-flying mammals: a
495 phylogenetic approach. *Physiological and Biochemical Zoology,* **81,** 526-550.
496 Maddala, G.S. (1983) *Limited-dependent and qualitative variables in econometrics*. Cambridge
497 University Press, Cambridge, UK.
498 Magee, L. (1990) R2 measures based on wald and likelihood ratio joint significance tests.
499 *American Statistician,* **44,** 250-253.
500 Martins, E.P. & Hansen, T.F. (1997) Phylogenies and the comparative method: A general
501 approach to incorporating phylogenetic information into the analysis of interspecific data.
502 *American Naturalist,* **149,** 646-667.
503 McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models,* 2 edn. Chapman and Hall,
504 London.
505 Nagelkerke, N.J.D. (1991) A note on a general definition of the coefficient of determination.
506 *Biometrika,* **78,** 691-692.
507 Nakagawa, S. & Schielzeth, H. (2013) A general and simple method for obtaining R2 from
508 generalized linear mixed-effects models. *Methods in Ecology and Evolution,* **4,** 133-142.
509 Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R
510 language. *Bioinformatics,* **20,** 289-290.
511 Petersen, K.B. & Pedersen, M.S. (2012) The matrix cookbook. Technical University of Denmark.

512 Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika,* **78,**

513 719-727.

514

515 **Supporting Information**

516 Additional Supporting Information may be found in the online version of this article:

517 Appendix S1. R scripts for computing R^2_{ls}, $R^2_{lr,}$ and R^2_{ce} with examples.

518 Appendix S2. Text and figures S1-S12 giving a comprehensive discussion of the behaviors of

519 R^2_{ls}, $R^2_{lr,}$ and R^2_{ce} in the simulations.

520 **Table 1.** Qualitative comparison among four R^2s with respect to their performance (i) as
521 measures of goodness-of-fit relative to the log-likelihood ratio comparing full to reduced models,
522 (ii) in partitioning explained variance between regression coefficients (fixed effects) and
523 variance parameters (random effects), and (iii) to infer the fit of the model to data generated by
524 the same statistical process (given by the variance in R^2 values among simulations). Three
525 qualitative levels imply good (black), acceptable (dark gray), and poor (light gray) performance.
526 The qualitative comparisons are based only on the simulations in this article and may differ in
527 other contexts.

		LMM	PGLS	GLMM	PLOG
R^2_{ls}	Goodness-of-fit				
	Partitioning variances				
	Inference				
R^2_{lr}	Goodness-of-fit				
	Partitioning variances				
	Inference				
R^2_{ce}	Goodness-of-fit				
	Partitioning variances				
	Inference				
R^2_{glmm}	Goodness-of-fit		-		-
	Partitioning variances	†	-	†	-
	Inference		-		-

528 † Marginal rather than partial R^2_{glmm} was used; partitioning variances will be more effective with
529 partial R^2_{glmm}.

530 - R^2_{glmm} cannot be applied to PGLS and PLOG models

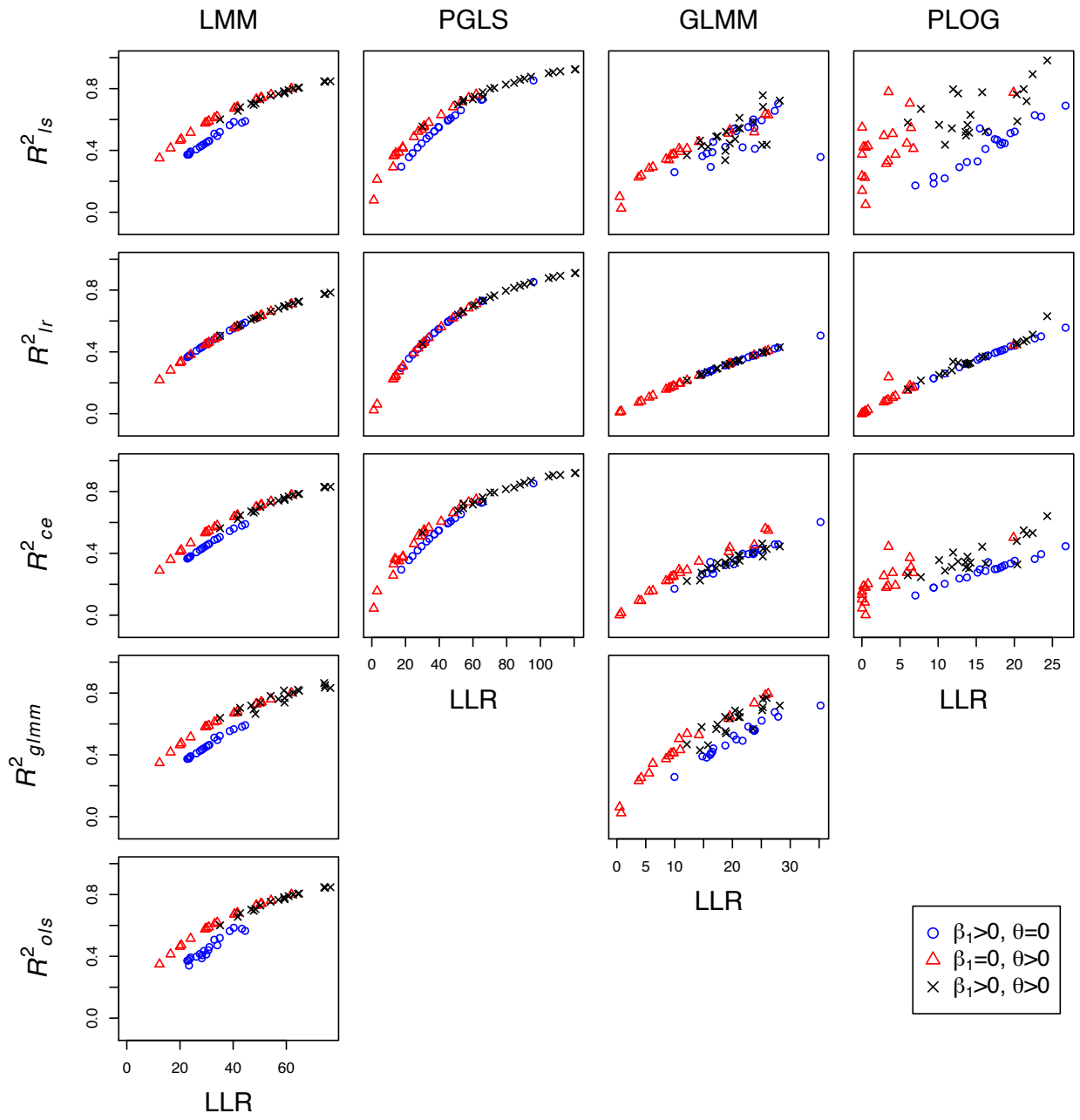


Fig. 1. Results for LMM, PGLS, GLMM, and PLOG simulations giving R^2_{ls} (eqn 2), R^2_{lr} (eqn 6), R^2_{ce} (eqn 7), R^2_{glmm} (eqn 8), and R^2_{ols} versus the log likelihood ratio (LLR) between full model and reduced model containing only an intercept. All simulated data had 100 samples. For LMM, the simulation model (eqn 9) contained a fixed effect with $\beta_1 = 0$ or 1, and a random effect u_i

536 with 10 levels and variance $\theta = 0$ or 1.5. The binomial (binary) GLMM was similar but with $\beta_1 =$

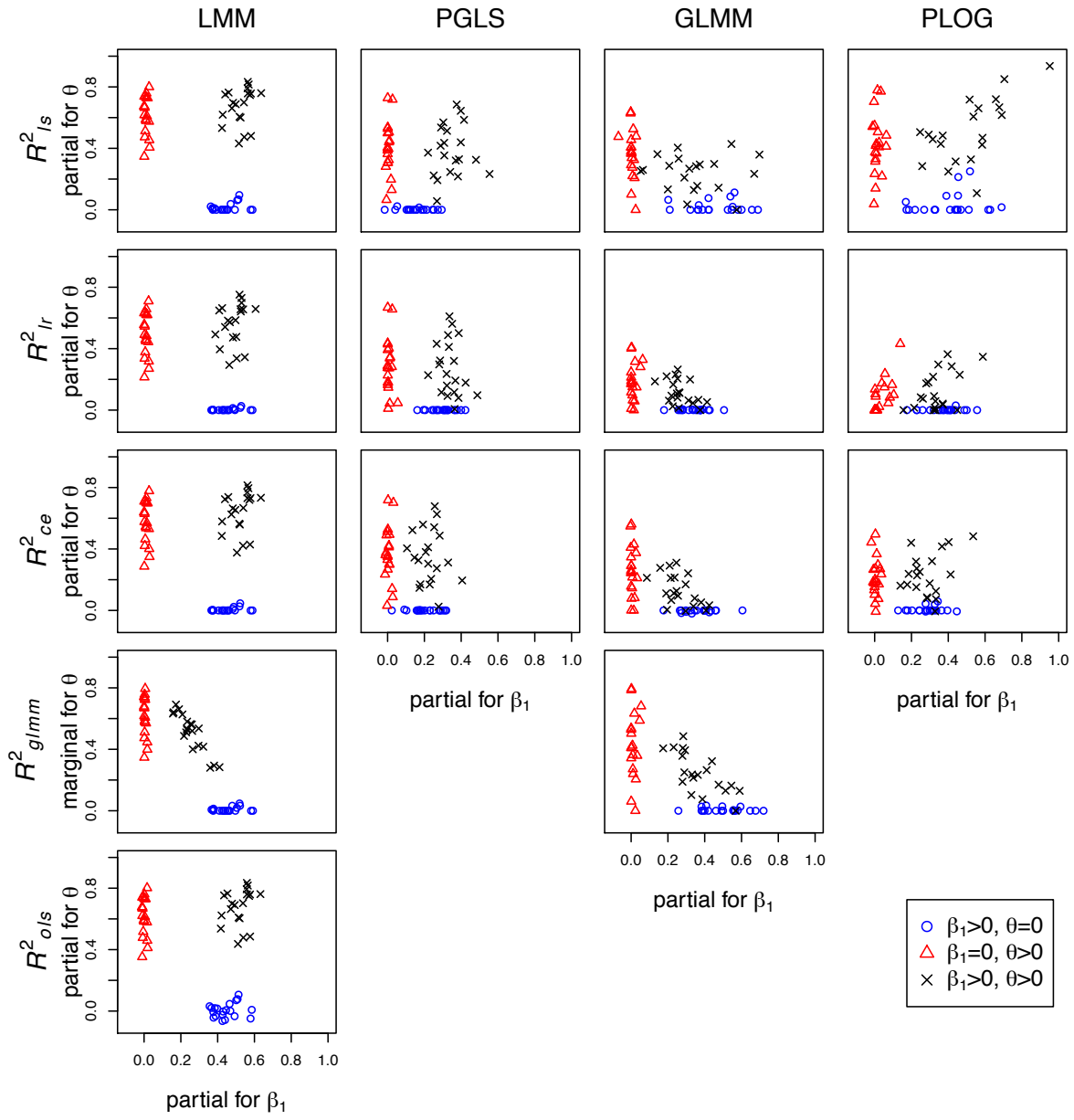
537 0 or 1.8, and $\theta = 0$ or 1.8. For PGLS, $\beta_1 = 0$ or 1.5, and the strength of phylogenetic signal $\theta = \lambda$

538 $= 0$ or 0.7; while for PLOG $\beta_1 = 0$ or 1.5, and $\theta = 0$ or 2. The LMM was fit using lmer (Bates *et*

539 *al.* 2014); the GLMM was fit using glmer (Bates *et al.* 2014); the PGLS was fit using phylolm

540 (Ho & Ane 2014); and for PLOG LLR and R^2_{lr} were fit using a modified version of phyloglm

541 (Ho & Ane 2014), and R^2_{ls} and R^2_{ce} were fit using binaryPGLMM (Ives & Garland 2014).



542

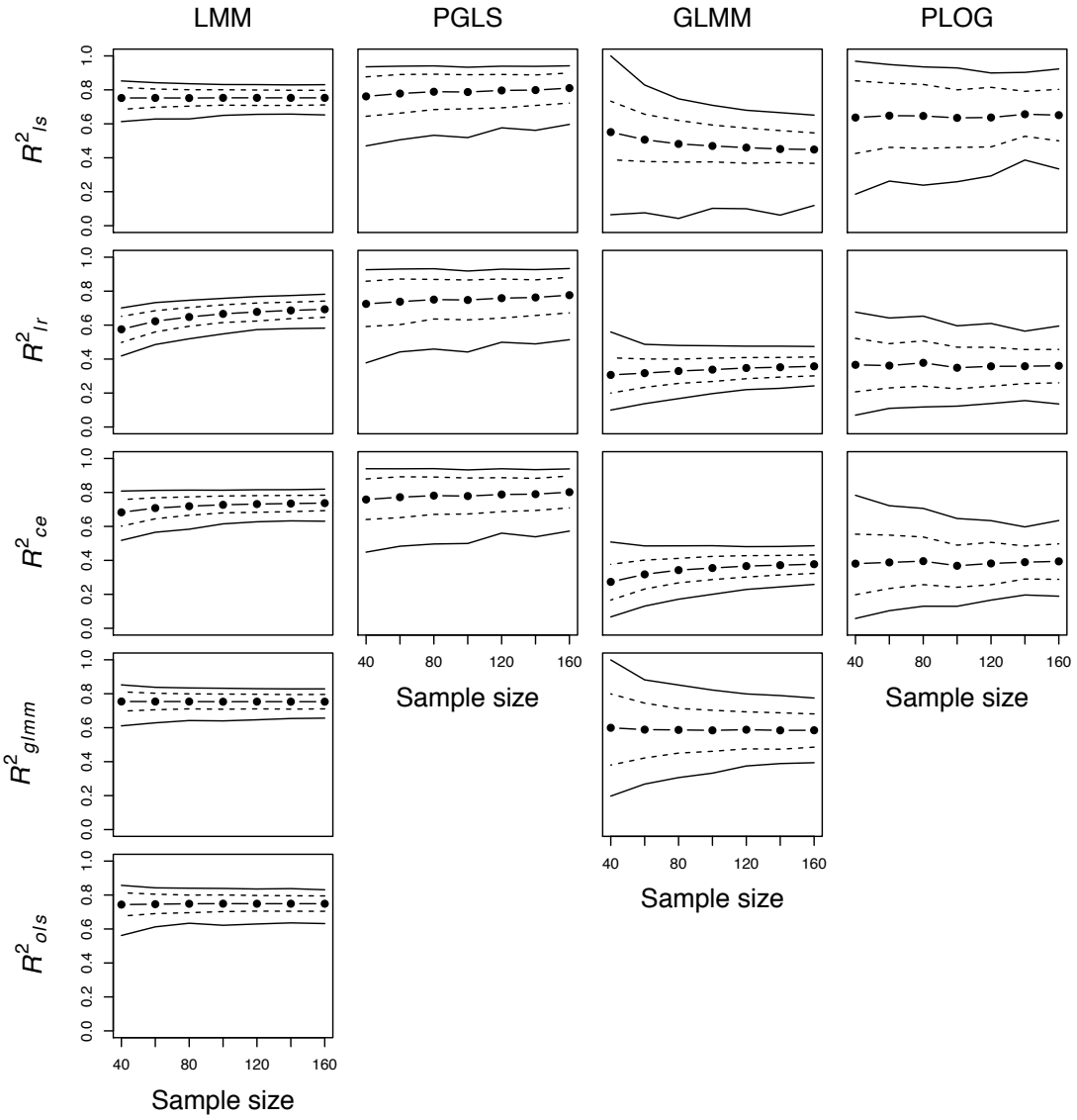
543 **Fig. 2.** Results for LMM, PGLS, GLMM, and PLOG simulations giving partial values of R^2_{ls},

544 R^2_{lr}, R^2_{ce}, R^2_{glmm}, and R^2_{ols}. The partial R^2 for β_1 was calculated using the reduced model in which

545 θ is removed, and for the partial R^2 for θ the reduced model had β_1 removed. The simulated data

546 and fitting methods are the same as in figure 1. For reduced models without variance parameters,

547 fitting was done using lm and glm.



548

549 **Fig. 3.** Results for LMM, PGLS, GLMM, and PLOG simulations showing means, 66% and 95%

550 inclusion intervals for R^2_{ls}, R^2_{lr}, R^2_{ce}, R^2_{glmm}, and R^2_{ols} versus sample size. For GLMM 1000

551 datasets were analyzed at each sample size, while 500 datasets were analyzed for the other

552 models. Parameter values were: LMM, $\beta_1 = 1$, $\theta = 1.5$; PGLS, $\beta_1 = 1.5$, $\theta = 0.7$; GLMM, $\beta_1 = 1.8$,

553 $\theta = 1.8$; and PLOG, $\beta_1 = 1.5$, $\theta = 2$. For GLMM, values of R^2_{ls} less than –1 were excluded; these

554 were 19, 9, 7, 6, 4, 3, and 4 of the 1000 values for sample sizes 40, 60, …, 160.