Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 944–946 doi: 10.1111/2041-210X.12225

FORUM

Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models

Paul C.D. Johnson*

Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Graham Kerr Building, Glasgow G12 8QQ, UK

Summary

1. Nakagawa & Schielzeth extended the widely used goodness-of-fit statistic R^2 to apply to generalized linear mixed models (GLMMs). However, their R^2_{GLMM} method is restricted to models with the simplest random effects structure, known as random intercepts models. It is not applicable to another common random effects structure, random slopes models.

2. I show that R^2_{GLMM} can be extended to random slopes models using a simple formula that is straightforward to implement in statistical software. This extension substantially widens the potential application of R^2_{GLMM}.

Key-words: coefficient of determination, generalized linear mixed model, random slopes model, random regression

Introduction

The coefficient of determination, R^2, is a widely used statistic for assessing the goodness-of-fit, on a scale from 0 to 1, of a linear regression model (LM). It is defined as the proportion of variance in the response variable that is explained by the explanatory variables or, equivalently, the proportional reduction in unexplained variance. Unexplained variance can be viewed as variance in model prediction error, so R^2 can also be defined in terms of reduction in prediction error variance. Insofar as it is justifiable to make the leap from 'prediction' to 'understanding', R^2 can be intuitively interpreted as a measure of how much better we understand a system once we have measured and modelled some of its components.

R^2 has been extended to apply to generalized linear models (GLMs) (Maddala 1983) and linear mixed effects models (LMMs) (Snijders & Bosker 1994) [reviewed by (Nakagawa & Schielzeth 2013)]. Nakagawa & Schielzeth (2013) proposed a further generalization of R^2 to generalized linear mixed effects models (GLMMs), a useful advance given the ubiquity of GLMMs for data analysis in ecology and evolution (Bolker *et al*. 2009). A function to estimate this R^2_{GLMM} statistic, *r.squaredGLMM*, has been included in the *MuMIn* package (Bartoń 2014) for the *R* statistical software (R Core Team 2014). However, Nakagawa and Schielzeth's R^2_{GLMM} formula is applicable to only a subset of GLMMs known as random intercepts models. Random intercepts models are used to model clustered observations, for example, where multiple observations are taken on each of a sample of individuals. Correlations between clustered observations within individuals are accounted for by allowing each subject to have a different intercept representing the deviation of that subject from the global intercept. Random intercepts are typically modelled as being sampled from a normal distribution with mean zero and a variance parameter that is estimated from the data. Although random intercepts are probably the most popular random effects models in ecology and evolution, other random effect specifications are also common, in particular random slopes models, where not only the intercept but also the slope of the regression line is allowed to vary between individuals. Random intercepts and slopes are typically modelled as normally distributed deviations from the global intercept and slope, respectively. For example, random slopes models, under the name of 'random regression' models, are used to investigate individual variation in response to different environments (Nussey, Wilson & Brommer 2007). The aim of this article is to show how Nakagawa and Schielzeth's R^2_{GLMM} can be further extended to encompass random slopes models.

Nakagawa and Schielzeth's R^2_{GLMM}

Nakagawa & Schielzeth (2013) defined two R^2 statistics for GLMMs, marginal and conditional R^2_{GLMM}, that allow separation of the contributions of fixed and random effects to explaining variation in the responses. Marginal R^2_{GLMM} gauges the variance explained by the fixed effects as a proportion of the sum of all the variance components:

$$R^2_{GLMM(m)} = \frac{\sigma^2_f}{\sigma^2_f + \sum_{l=1}^{u} \sigma^2_l + \sigma^2_e + \sigma^2_d}, \qquad \text{eqn 1}$$

where σ^2_f is the variance attributable to the fixed effects, σ^2_l is the variance of the lth of u random effects, σ^2_e is the variance due to additive dispersion and σ^2_d is the distribution-specific variance. The residual variance, σ^2_ε, is defined

*Correspondence author. E-mail: paul.johnson@glasgow.ac.uk

as $\sigma_e^2 + \sigma_d^2$ for the purposes of this manuscript but see Nakagawa & Schielzeth (2013) for an alternative definition of dispersion. Conditional R^2 additionally includes in the numerator the variance explained by the random effects:

$$R^2_{\mathrm{GLMM}(c)} = \frac{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2}. \qquad \text{eqn 2}$$

It is the definition of the random effect variances, the σ_l^2, that requires generalization to allow $R^2_{\mathrm{GLMM\ (m)}}$ and $R^2_{\mathrm{GLMM\ (c)}}$ to be extended beyond random intercepts models. In Nakagawa and Schielzeth's formula, σ_l^2 is simply the variance of the l th random intercept. This formula is correct for random intercept models because each observation has the same random effect variance. However, in other random effects specifications, the random effect variance can differ between observations, and, as pointed out by Nakagawa and Schielzeth, this causes difficulties in computing a single random effect variance component.

Extension of R^2_{GLMM} to random slopes models

Consider the simplest and most familiar random slopes GLMM, a LMM with a single random intercept and a single random slope:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_{0j} + \alpha_{1j} x_{ij} + \varepsilon_{ij}, \qquad \text{eqn 3}$$

$$\begin{bmatrix} \alpha_{0j} \\ \alpha_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{\Sigma} \right), \qquad \text{eqn 4}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2_{\alpha 0} & \sigma_{\alpha 0 \alpha 1} \\ \sigma_{\alpha 0 \alpha 1} & \sigma^2_{\alpha 1} \end{bmatrix}, \qquad \text{eqn 5}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2_{\varepsilon}), \qquad \text{eqn 6}$$

where Y_{ij} and x_{ij} are, respectively, the response and predictor values (covariates) for the ith observation on the jth individual. Random deviation of the jth individual from the fixed global intercept, β_0, is represented by α_{0j}, while random deviation from the fixed global slope, β_1, is represented by α_{1j}. Because intercepts and slopes are typically correlated, three parameters are required to model the random effect, which are represented by the covariance matrix $\mathbf{\Sigma}$**.** The leading diagonal of $\mathbf{\Sigma}$ consists of the random intercept variance, $\sigma^2_{\alpha 0}$, and the random slope variance, $\sigma^2_{\alpha 1}$, while the off-diagonal element is the covariance, $\sigma_{\alpha 0 \alpha 1}$, between the random intercept and random slope. Finally, ε_{ij} is the residual of the ith observation on the jth individual and σ^2_{ε} is the residual variance. For LMMs, $\sigma^2_d = 0$, so that $\sigma^2_{\varepsilon} = \sigma^2_e$.

The difficulty of defining σ_l^2 for this model arises from the dependence of the random effect variance component on x_{ij}, which implies that σ_l^2 cannot be defined from $\mathbf{\Sigma}$ alone, but requires input from the x_{ij}. An observation-specific random effect variance, σ^2_{lij}, can be defined, given x_{ij}, as

$$\sigma^2_{lij} = \mathrm{var}(\alpha_{0j} + \alpha_{1j} x_{ij}), \qquad \text{eqn 7}$$

showing the dependence of σ^2_{lij} on x_{ij}. For example, when $x_{ij} = 0$ (i.e. at the intercept),

$$\sigma^2_{lij} = \mathrm{var}(\alpha_{0j}) = \sigma^2_{\alpha 0}, \qquad \text{eqn 8}$$

while when $x_{ij} = 1$,

$$\begin{aligned} \sigma^2_{lij} &= \mathrm{var}(\alpha_{0j} + \alpha_{1j}) \\ &= \mathrm{var}(\alpha_{0j}) + \mathrm{var}(\alpha_{1j}) + 2\mathrm{cov}(\alpha_{0j}, \alpha_{1j}) \\ &= \sigma^2_{\alpha 0} + \sigma^2_{\alpha 0} + 2\sigma_{\alpha 0 \alpha 1} \end{aligned} \qquad \text{eqn 9}$$

(Snijders & Bosker 2012). In the most extreme case where the x_{ij} values are unique, there will be as many random effect variances as observations. The first step to estimating the random effect variance component is to estimate each σ^2_{lij}. The random effect portion of the model, $\alpha_{0j} + \alpha_{1j} x_{ij}$, can then be viewed as a mixture of n normal distributions with a common mean of zero but up to n different variances, where n is the number of observations. When the mean is constant, the variance of a mixture is simply the mean of the individual variances (Behboodian 1970). The mean random effect variance is therefore

$$\overline{\sigma_l^2} = \left(\sum_j \sum_i \sigma^2_{lij}\right)/n. \qquad \text{eqn 10}$$

A simple and general formula for $\overline{\sigma_l^2}$ given any value of x_{ij} can be derived as follows. For any random effects specification, let $\mathbf{Z}$ be the design matrix of the random effects of a GLMM with n rows and k columns corresponding to the k random effects, and $\mathbf{\Sigma}$ the covariance matrix of the random effects of dimension k. For example, in the simple random slopes model in equations 3-6, the first column of $\mathbf{Z}$ is a vector of ones corresponding to the random intercept, while the second is the predictor variable, the x_{ij}. The vector of observation-level random effect variances is the leading diagonal of the $n \times n$ matrix $\mathbf{Z\Sigma Z}'$, where $\mathbf{Z}'$ is the transpose of $\mathbf{Z}$ (Laird & Ware 1982). The mean random effect variance, $\overline{\sigma_l^2}$, is the mean of this vector, that is,

$$\overline{\sigma_l^2} = \mathrm{Tr}(\mathbf{Z\Sigma Z}')/n, \qquad \text{eqn 11}$$

where the Tr denotes the trace operation, which sums the leading diagonal. An index notation version of the matrix notation equation 11 is contained within equation 20 of Snijders & Bosker (1994). The advantage of the matrix version is computational simplicity. Equation 11 gives the same results as Nakagawa & Schielzeth's method for random intercepts models but can also be used for random slopes models as well as models with no intercept. An estimate of $\overline{\sigma_l^2}$ for use in Equations 1 and 2 can be easily computed from the estimated covariance matrix of the lth random effect. Examples of the application of this procedure to estimating R^2_{GLMM} from random slopes GLMMs using R are provided as Data S1.

The Supplementary R code also illustrates a simplified method of estimating the term β_0 in equation A6 of Nakagawa & Schielzeth (2013), which approximates σ^2_d for a Poisson GLMM. Rather than refit the model after centring or dropping the covariates as recommended, β_0 can be more easily estimated by taking the mean of $\mathbf{X}\hat{\boldsymbol{\beta}}$, the linear predictor, where $\mathbf{X}$ is the design matrix for the fixed effects and $\hat{\boldsymbol{\beta}}$ is the vector of fixed effect estimates.

These extensions to R^2_{GLMM} have been incorporated into the *r.squaredGLMM* function in version 1.10.0 of the *MuMIn* package (Bartoń 2014).

Discussion

The extension described above allows both marginal and conditional R^2_{GLMM} to be estimated from a random slopes model, obviating the need to approximate R^2_{GLMM} from the corresponding random intercepts model as recommended by Nakagawa & Schielzeth (2013). It is clearly preferable to estimate R^2_{GLMM} from the correct model given that there is no computational cost but is the improvement in either marginal or conditional R^2_{GLMM} likely to be substantial? Nakagawa & Schielzeth (2013) suggest that marginal and condition R^2_{GLMM} will usually be very similar when approximated from a random intercepts fit, and Snijders & Bosker (2012) make a similar claim for their related R^2_1 and R^2_2 statistics. Not surprisingly, the gain in accuracy in both R^2_{GLMM} statistics will depend on how well the random intercepts model approximates the random slopes model. The accuracy of the marginal R^2_{GLMM} approximation will depend on the accuracy of the global slope (or slopes) estimate from the random intercepts model, because the scale of the global slope (or slopes) estimate determines σ^2_f (Nakagawa & Schielzeth 2013), which in turn determines marginal R^2_{GLMM}. For balanced data, where the numbers of observations and the covariate distributions are balanced between groups, this approximation should be good, so the estimates of the global slope and marginal R^2_{GLMM} are likely to be very similar under both models. However, unbalanced data are common in ecology, for example where sampling strategies are constrained in space by variable access to sampling sites or in time by fluctuating resources, and in such cases the improvement in marginal R^2_{GLMM} could be considerable. For example, if one individual (or site, etc.) yields an unusually large number of observations, the global slope estimate will be biased towards that individual in a random intercepts model but not in a random slopes model. Examples of both scenarios are given in the Supplementary R code (Data S1).

Improvement in conditional R^2_{GLMM} is easier to predict and explain. Regardless of the adequacy of the marginal R^2_{GLMM} approximation, if the random slopes model fits substantially better than the random intercepts model, it should have lower residual variance (or less overdispersion, in the context of overdispersed Poisson or binomial GLMMs) and therefore higher conditional R^2_{GLMM}.

This extension will apply to other statistics that incorporate a random effects variance component calculated from a random slopes model, including the intraclass correlation coefficient (ICC), which gauges variance between groups (e.g. individuals or sites) as a proportion of the total variance. ICC can be used to measure intraindividual repeatability, also known as consistency, and has been applied widely in ecology and evolutionary biology (Nakagawa & Schielzeth 2010). Like R^2, ICC has also been generalized to random intercepts GLMMs by Nakagawa & Schielzeth (2010), but not to random slopes GLMMs. Equation 11 could also be applied to calculating repeatability (Nakagawa & Schielzeth 2010) by fixing a column of $\mathbf{Z}$ to a single value. For example, age dependence in phenotypic consistency could be investigated by estimating ICC conditioned on a range of ages.

In conclusion, the extension of R^2_{GLMM} to random slopes GLMMs substantially widens the range of models to which this useful measure can be applied.

Acknowledgements

I am grateful to S. Nakagawa, K. Bartoń and J. Lindström for helpful discussions, and to H. Schielzeth and two anonymous reviewers, whose comments greatly improved this manuscript. This work was supported by a BBSRC project grant (BB/K004484/1).

Data accessibility

R scripts: uploaded as online supporting information.

References

Bartoń, K. (2014) MuMIn: Multi-model inference. R package version 1.10.0. Retrieved May 14, 2014, from http://cran.r-project.org/package = MuMIn

Behboodian, J. (1970) On a Mixture of Normal Distributions. *Biometrika*, **57**, 215–217.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, **24**, 127–135.

Laird, N.M. & Ware, J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*, 1st edn. Cambridge University Press, Cambridge, UK.

Nakagawa, S. & Schielzeth, H. (2010) Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews of the Cambridge Philosophical Society*, **85**, 935–956.

Nakagawa, S. & Schielzeth, H. (2013) A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, 133–142.

Nussey, D.H., Wilson, A.J. & Brommer, J.E. (2007) The evolutionary ecology of individual phenotypic plasticity in wild populations. *Journal of Evolutionary Biology*, **20**, 831–844.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved April 10, 2014, from http://www.r-project.org/

Snijders, T.A.B. & Bosker, R.J. (1994) Modeled variance in two-level models. *Sociological Methods & Research*, **22**, 342–363.

Snijders, T.A.B. & Bosker, R.J. (2012) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd edn. Sage, London.

Received 13 January 2014; accepted 25 June 2014
Handling Editor: Robert B. O'Hara

Supporting Information

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

Data S1. R code illustrating the calculation of R^2_{GLMM}.

