PHYLOGENETIC COMPARATIVE METHODS: REGRESSION MODELS

Cara Evans/Catherine Sheard

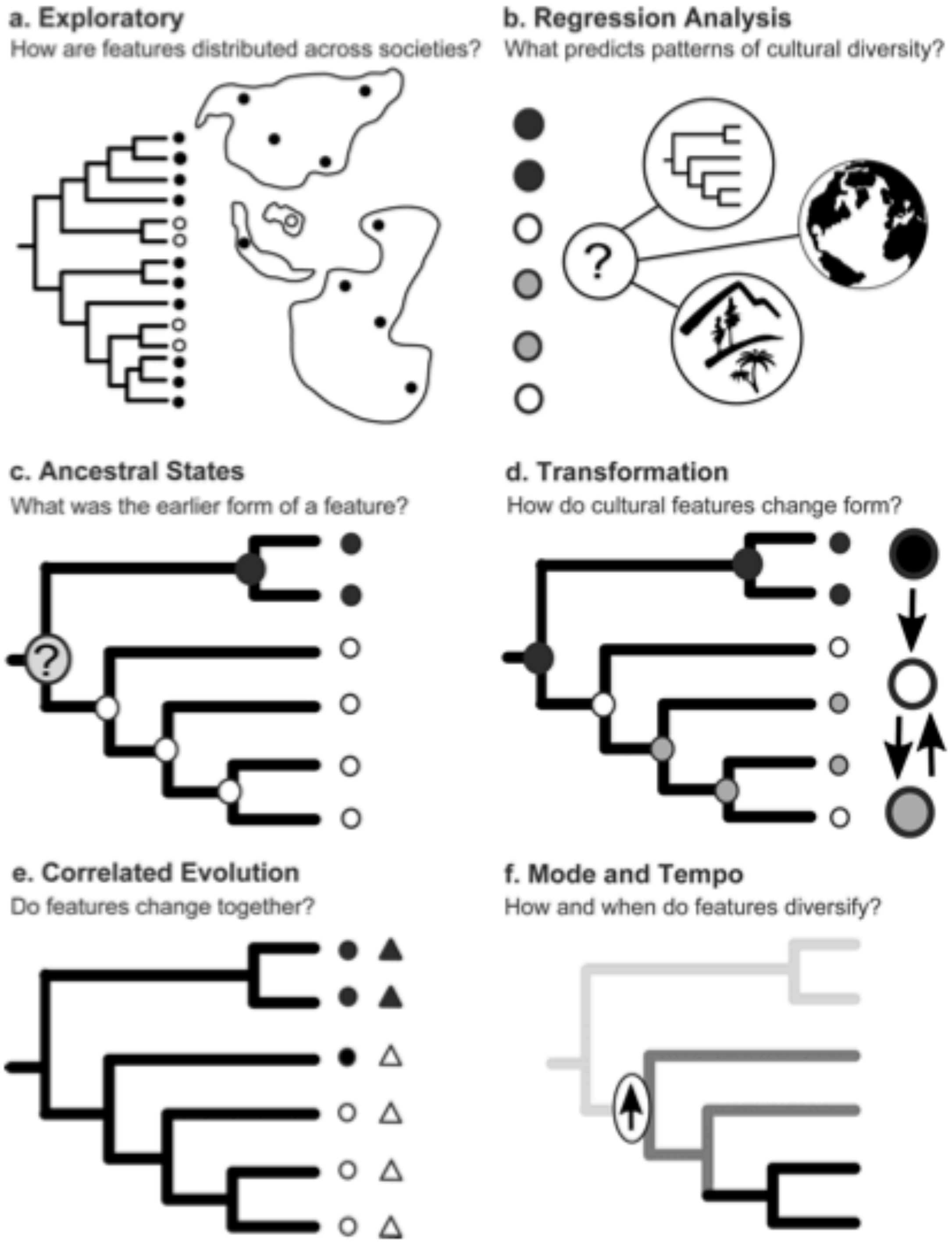
Quantitative Methods Spring School
Max Planck Institute for the Science of Human History
May 2017

OVERVIEW

- **Discussion of the types of applications for phylogenetic regression and phylogenetic multilevel models**

- **Comparison of some state of art methods for conducting phylogenetic regression**

- **A brief further introduction to a specific application used for conducting phylogenetic multilevel models (MCMCglmm), followed by a basic model run through of the application in R**


a. Exploratory
How are features distributed across societies?
b. Regression Analysis
What predicts patterns of cultural diversity?
?
c. Ancestral States
What was the earlier form of a feature?
?
d. Transformation
How do cultural features change form?
e. Correlated Evolution
Do features change together?
f. Mode and Tempo
How and when do features diversify?

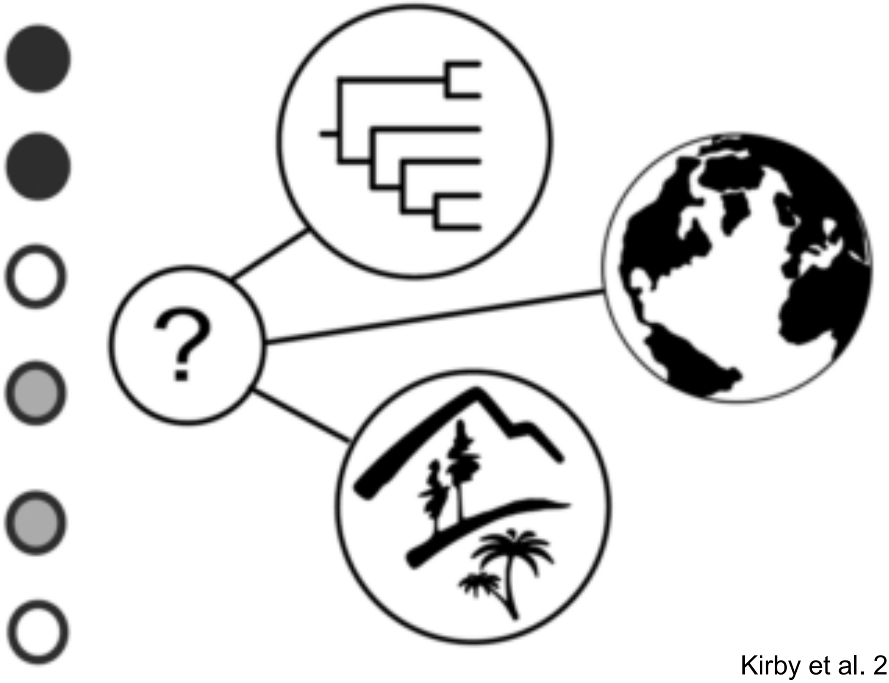
Kirby et al. (2016). *PLoS One*, *11*(7)

What question am I trying to ask with my data?

How do I ask that question?

REGRESSION ANALYSIS

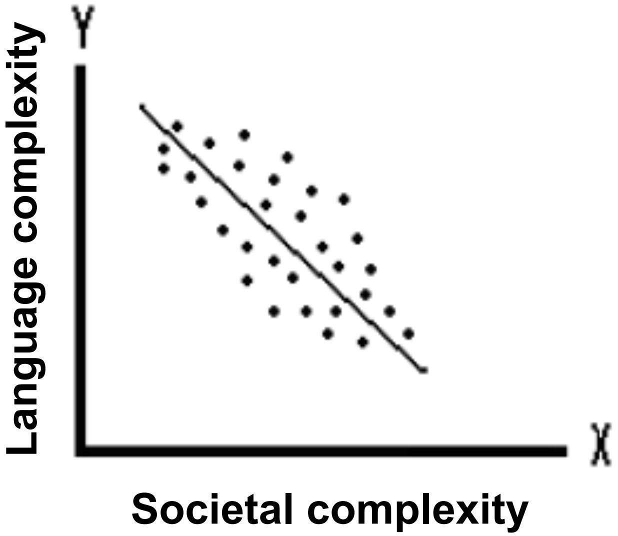
What predicts patterns of cultural diversity?

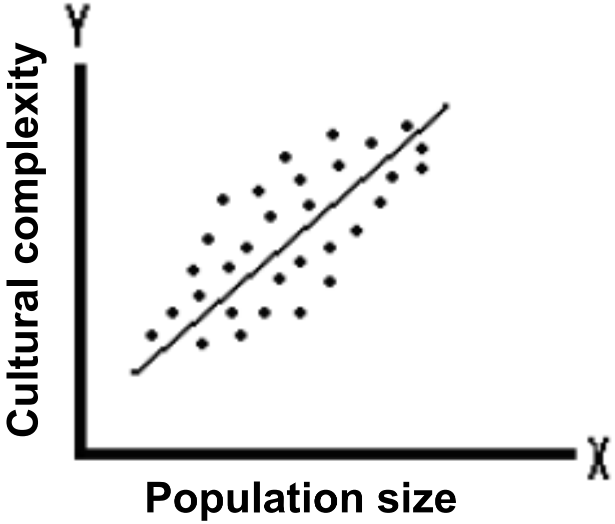


Kirby et al. 2016

REGRESSION ANALYSIS

What predicts patterns of cultural and linguistic diversity?



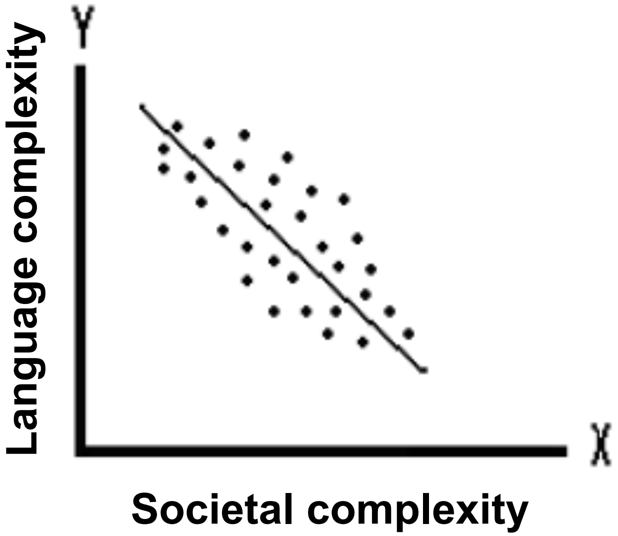


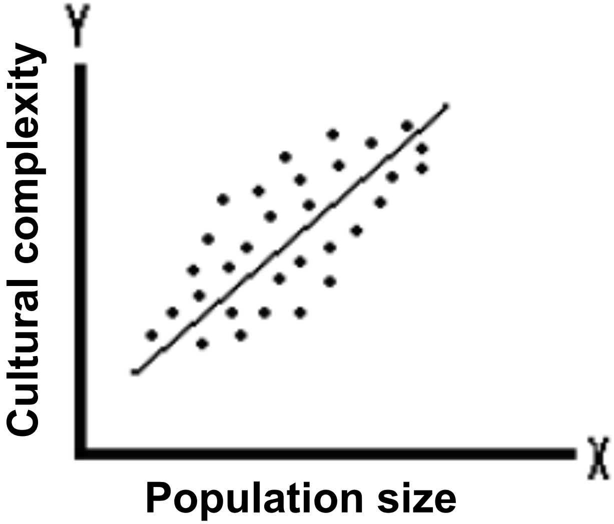
✓ **Do more complex societies have less complex languages?**

✓ **Do larger populations have higher levels of cultural complexity?**

REGRESSION ANALYSIS

What predicts patterns of cultural and linguistic diversity?





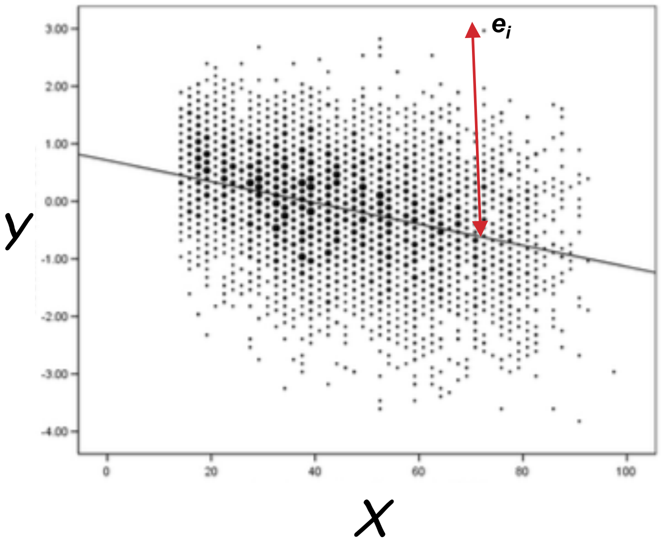
✓ **Do more complex societies have less complex languages?**

✓ **Do larger populations have higher levels of cultural complexity?**

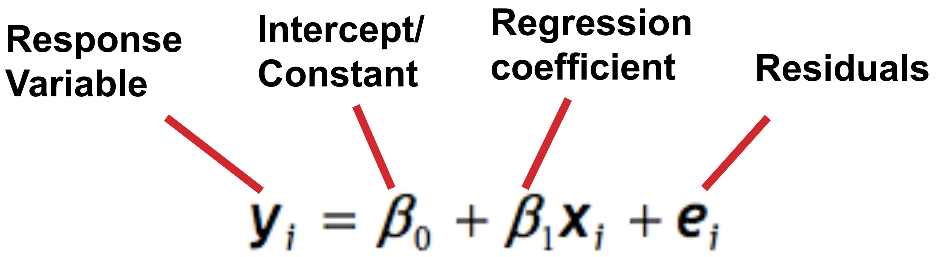
✗ **Ancestral states, causation, mode and tempo, transformation**

REGRESSION ANALYSIS

What predicts patterns of cultural and linguistic diversity?



Linear regression equation:



Response Variable

Intercept/ Constant

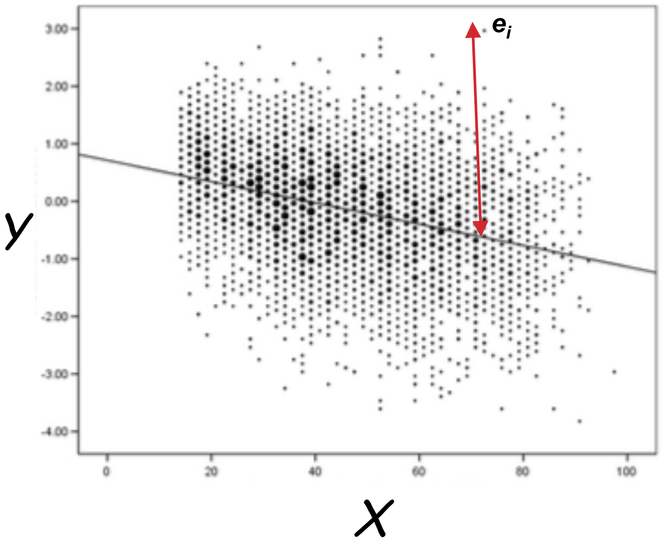
Regression coefficient

Residuals

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

REGRESSION ANALYSIS

What predicts patterns of cultural and linguistic diversity?



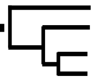
Linear regression equation:

Response Variable **Intercept/ Constant** **Regression coefficient** **Residuals**

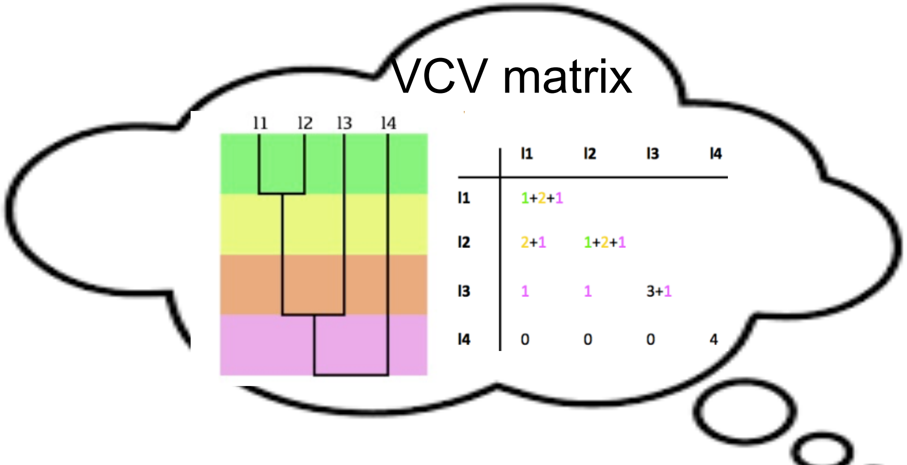
$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Phylogenetic regression equation:

$$y_i = \beta_0 + \beta_1 x_i + e_i \leftarrow$$



VCV matrix

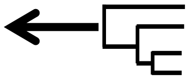


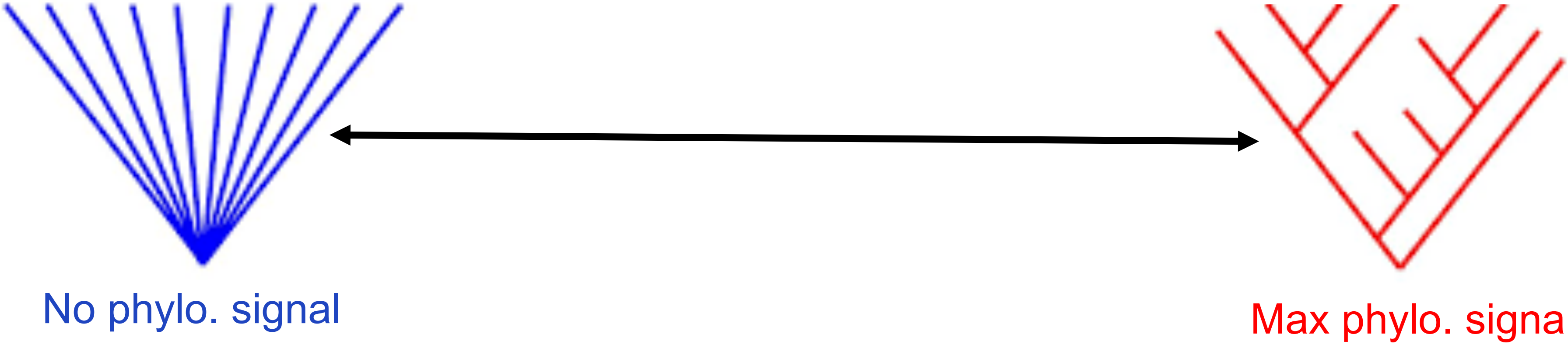
	l1	l2	l3	l4
l1	1+2+1			
l2	2+1	1+2+1		
l3	1	1	3+1	
l4	0	0	0	4

PHYLOGENETIC REGRESSION

Estimating phylogenetic signal:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$



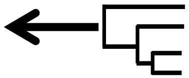


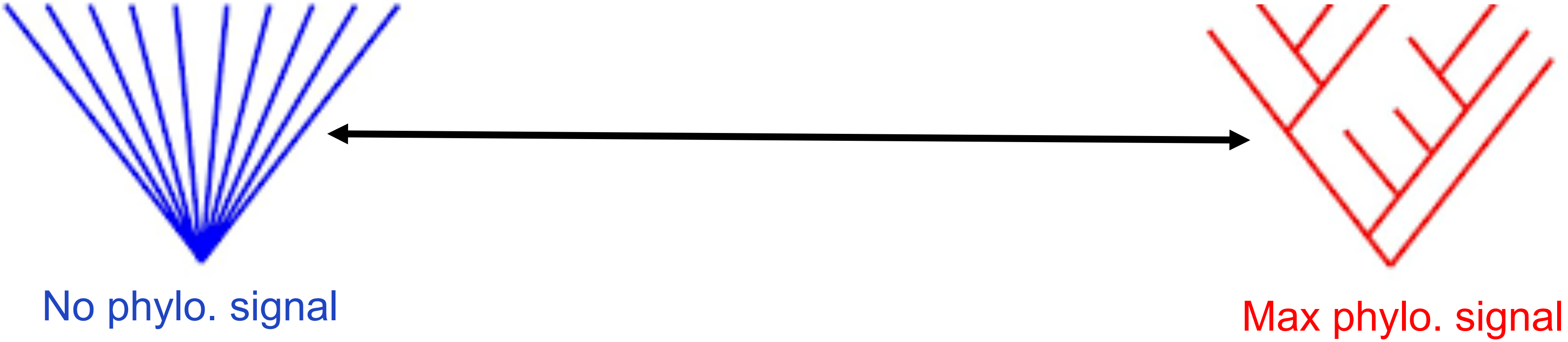
Revell, L. J. (2010). Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, *1*(4), 319-329.

PHYLOGENETIC REGRESSION

Estimating phylogenetic signal:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$





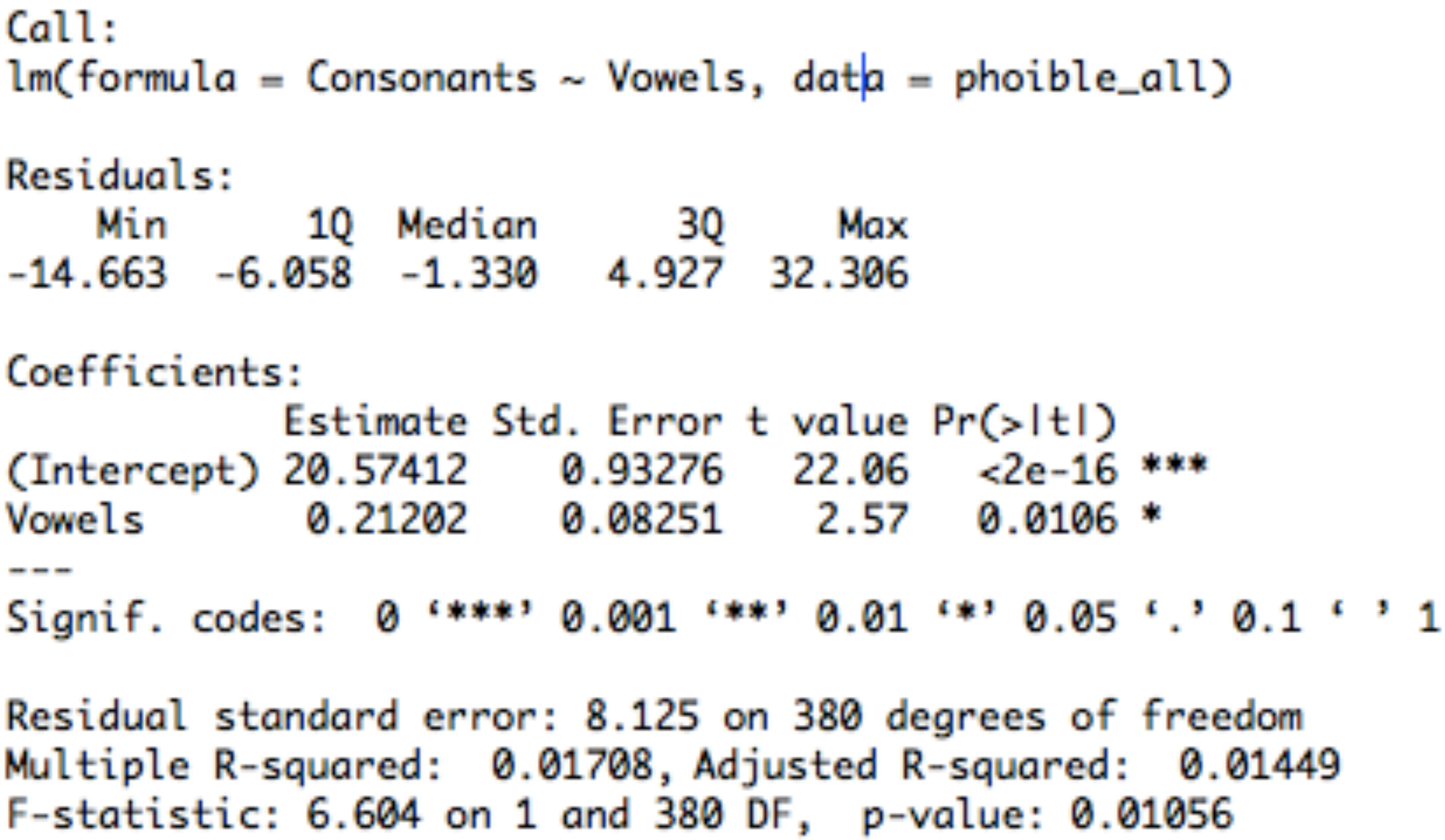
Assuming maximum phylogenetic signal when there is none is as bad as assuming no phylogenetic signal when there is some!

Methods that simultaneously estimate phylogenetic signal in the residual error with the regression parameters are preferred!

Revell, L. J. (2010). Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, *1*(4), 319-329.

ANNEMARIE'S EXAMPLE USING LM VS. PGLS

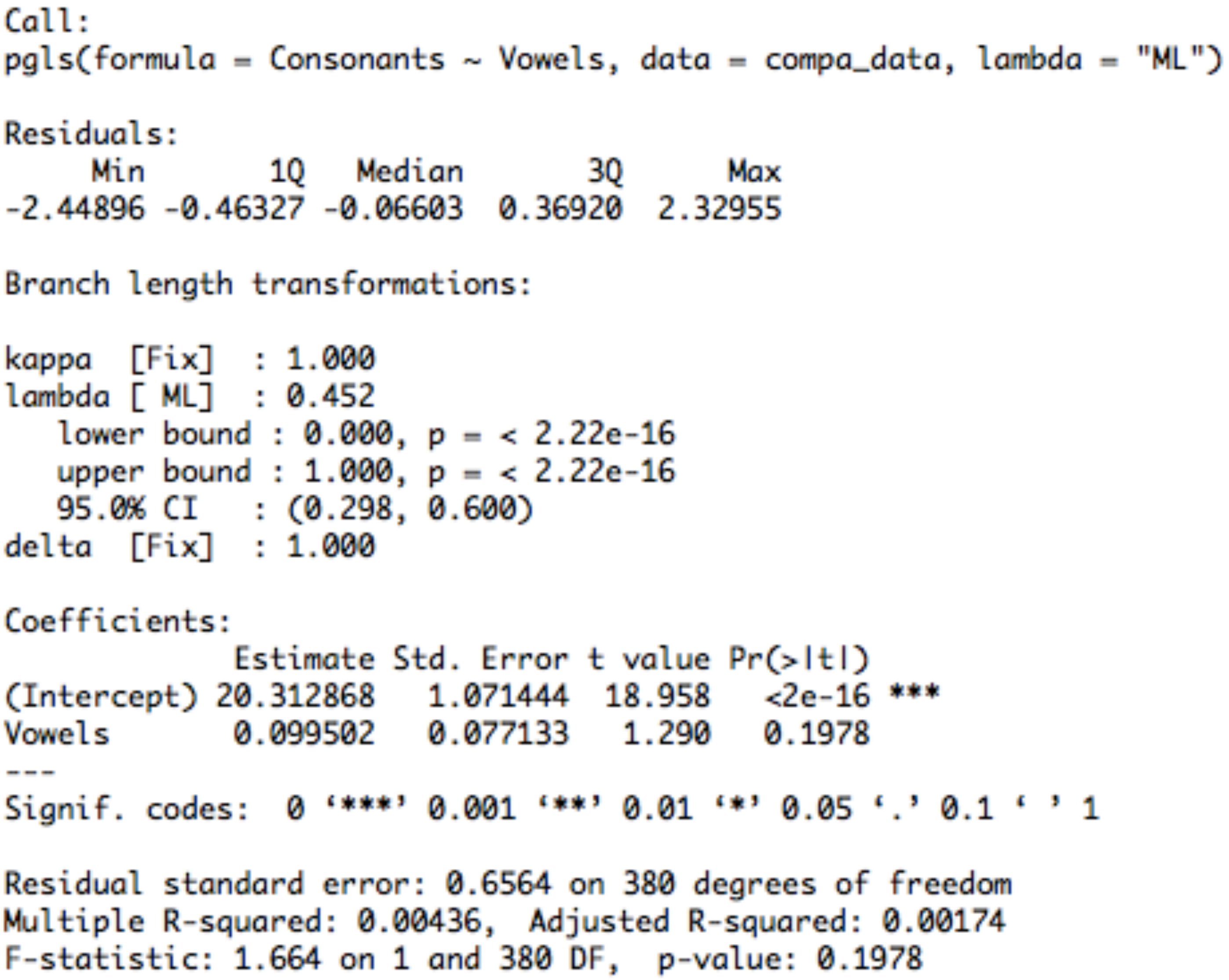
Do societies with more vowels also have more consonants?



```
Call:
lm(formula = Consonants ~ Vowels, data = phoible_all)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.663  -6.058  -1.330   4.927  32.306 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20.57412    0.93276   22.06   <2e-16 ***
Vowels       0.21202    0.08251    2.57   0.0106 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.125 on 380 degrees of freedom
Multiple R-squared:  0.01708,	Adjusted R-squared:  0.01449 
F-statistic: 6.604 on 1 and 380 DF,  p-value: 0.01056
```

ANNEMARIE’S EXAMPLE USING LM VS. PGLS

Do societies with more vowels also have more consonants?

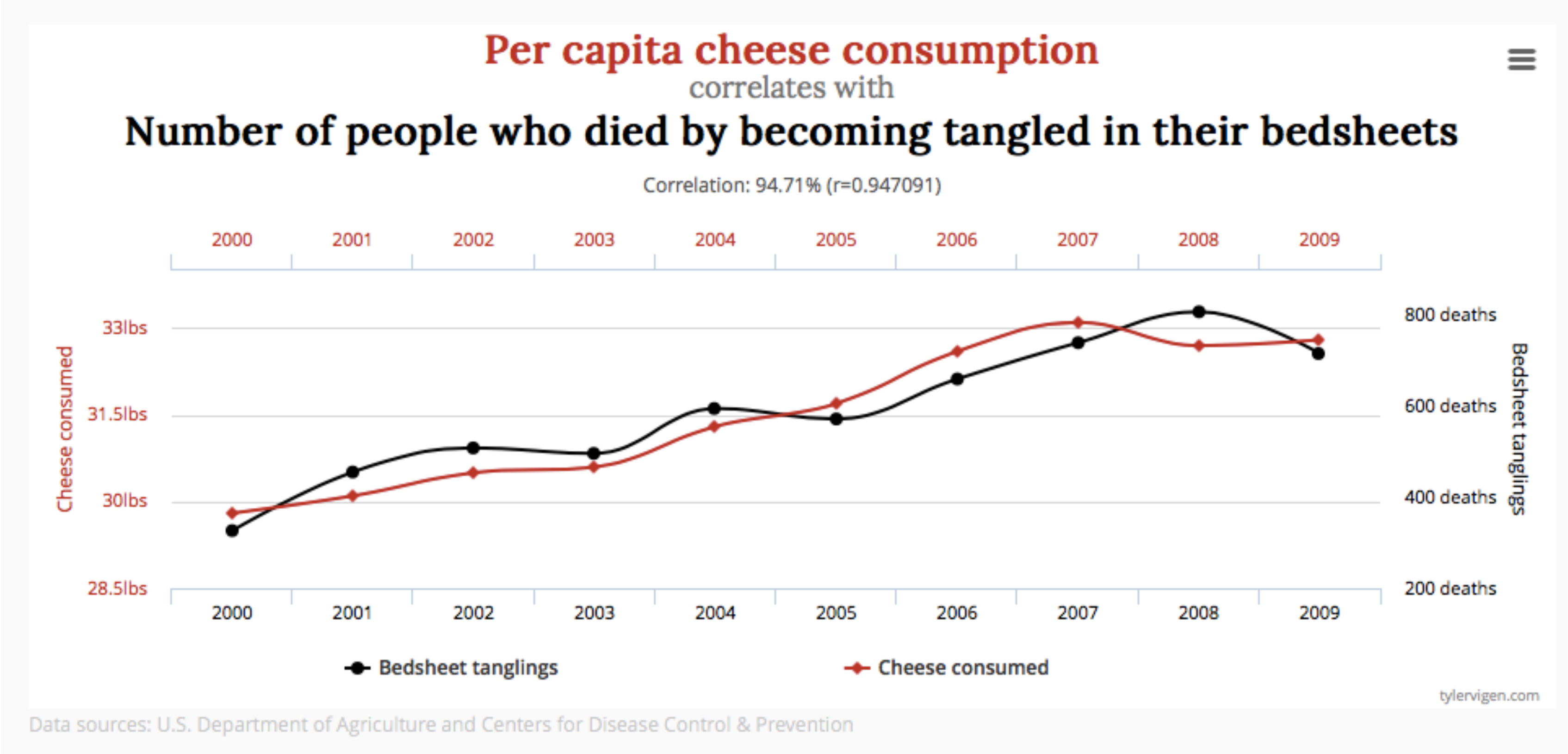


```
Call:
pgls(formula = Consonants ~ Vowels, data = compa_data, lambda = "ML")

Residuals:
     Min       1Q   Median       3Q      Max 
-2.44896 -0.46327 -0.06603  0.36920  2.32955 

Branch length transformations:

kappa  [Fix]  : 1.000
lambda [ ML]  : 0.452
   lower bound : 0.000, p = < 2.22e-16
   upper bound : 1.000, p = < 2.22e-16
   95.0% CI   : (0.298, 0.600)
delta  [Fix]  : 1.000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20.312868   1.071444  18.958   <2e-16 ***
Vowels       0.099502   0.077133   1.290   0.1978    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6564 on 380 degrees of freedom
Multiple R-squared: 0.00436,	Adjusted R-squared: 0.00174 
F-statistic: 1.664 on 1 and 380 DF,  p-value: 0.1978
```

REGRESSION CAVEATS

All of the usual regression caveats regarding model fit, checks and assumptions also apply in phylogenetic regression, including:

Correlation vs. causation



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

http://www.tylervigen.com/spurious-correlations

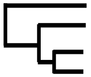
PHYLOGENETIC MULTIPLE REGRESSION

Models that include more than one predictor variable

One way to try to avoid finding spurious relationships is to include additional variables that might act as confounding factors or also predict the variable of interest….

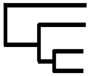
Simple regression model (1 predictor variable)

$$y_i = \beta_0 + \beta_1 x_i + e_i \longleftarrow$$



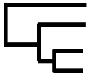
Multiple regression model (2 predictor variables)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \longleftarrow$$



Multiple regression model (N predictor variables + interaction terms)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_p x_{pi} + e_i \longleftarrow$$



(But be careful not to over-parameterize the model!! See: Anderson, D. R., & Burnham, K. P. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*.)

The ecology of religious beliefs

Carlos A. Botero[a,b,1], Beth Gardner[c], Kathryn R. Kirby[d], Joseph Bulbulia[e], Michael C. Gavin[f], and Russell D. Gray[g,h,i]

[a]Initiative for Biological Complexity, Department of the Interior Southeast Climate Science Center, and [c]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695; [b]Department of Biology, Washington University in St. Louis, St. Louis, MO 63130; [d]Department of Ecology & Evolutionary Biology and Department of Geography and Program in Planning, University of Toronto, Toronto, ON, Canada M5S 3E8; [e]School of Art History, Classics and Religious Studies, Victoria University of Wellington, Wellington 6140, New Zealand; [f]Department of Human Dimensions of Natural Resources, Colorado State University, Fort Collins, CO 80523; [g]School of Psychology, University of Auckland, Auckland 1142, New Zealand; [h]School of Philosophy, Research School of the Social Sciences, Australian National University, 0200 Canberra, Australia; and [i]Department of Linguistic and Cultural Evolution, Max Planck Institute for History and the Sciences, 07745 Jena, Germany

Edited by Ara Norenzayan, University of British Columbia, Vancouver, Canada, and accepted by the Editorial Board October 10, 2014 (received for review May 11, 2014)

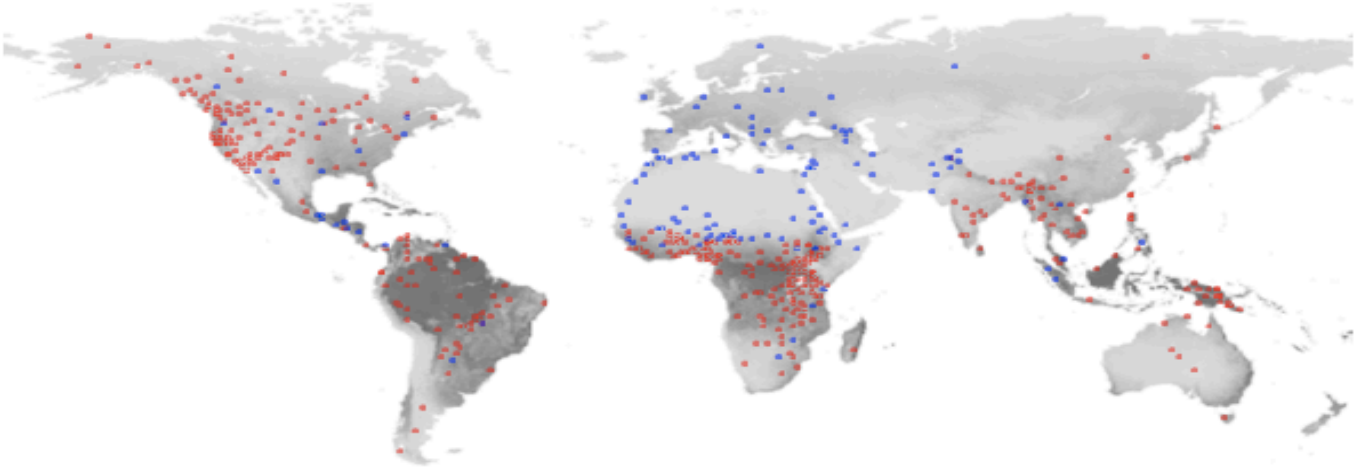
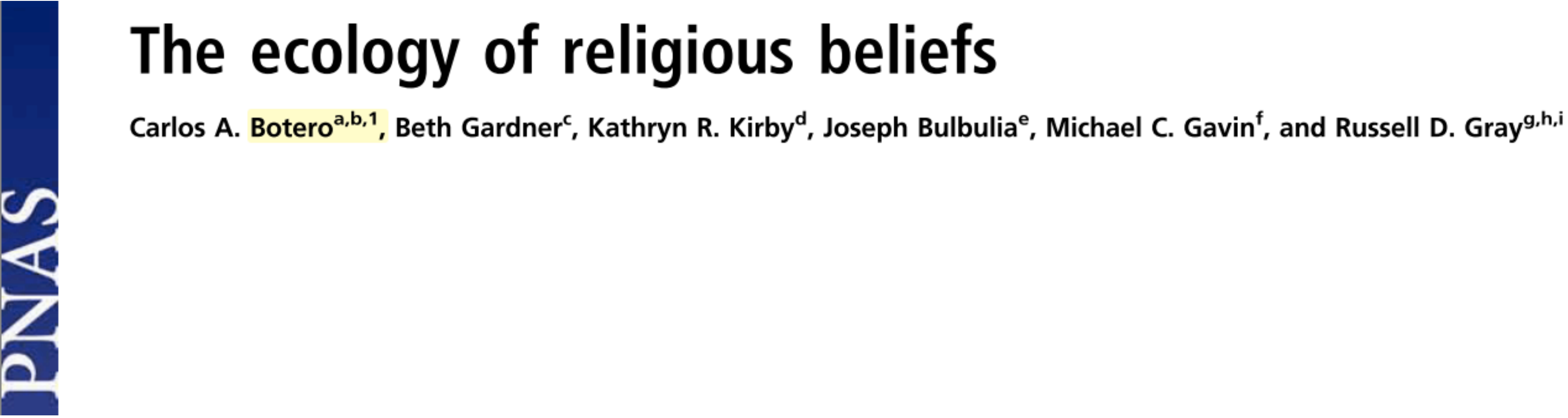


Fig. 1. Global distribution of societies that exhibit beliefs in moralizing high gods (blue) or not (i.e., atheism or beliefs in nonmoralizing deities or spirits in red). The underlying map depicts the mean values of net primary productivity (i.e., the net balance of monthly consumption relative to production of carbon dioxide by living plants) in gray scale. Darker localities reflect places with greater potential for overall plant growth.

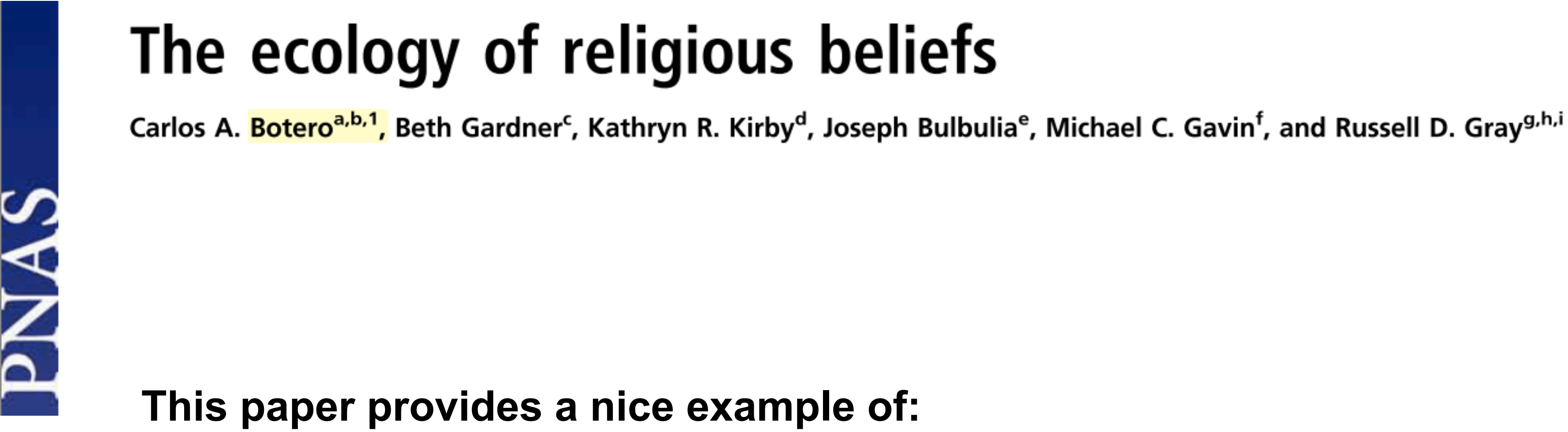


The ecology of religious beliefs

Carlos A. Botero[a,b,1], Beth Gardner[c], Kathryn R. Kirby[d], Joseph Bulbulia[e], Michael C. Gavin[f], and Russell D. Gray[g,h,i]

This paper provides a nice example of:

-Data checks and principle components analysis (PCA) undertaken to account for multicolinearity (correlations) between the many included predictor variables



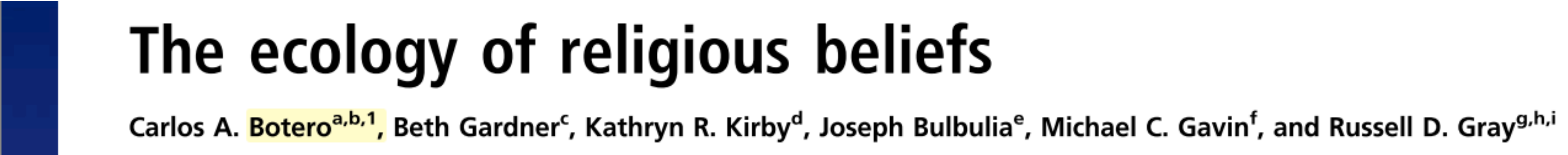
The ecology of religious beliefs

Carlos A. Botero[a,b,1], Beth Gardner[c], Kathryn R. Kirby[d], Joseph Bulbulia[e], Michael C. Gavin[f], and Russell D. Gray[g,h,i]

This paper provides a nice example of:

-Data checks and principle components analysis (PCA) undertaken to account for multicolinearity (correlations) between the many included predictor variables (also Capillini et al (2015) –see refs – for example using variance inflation factors to check for multicolinearity..).

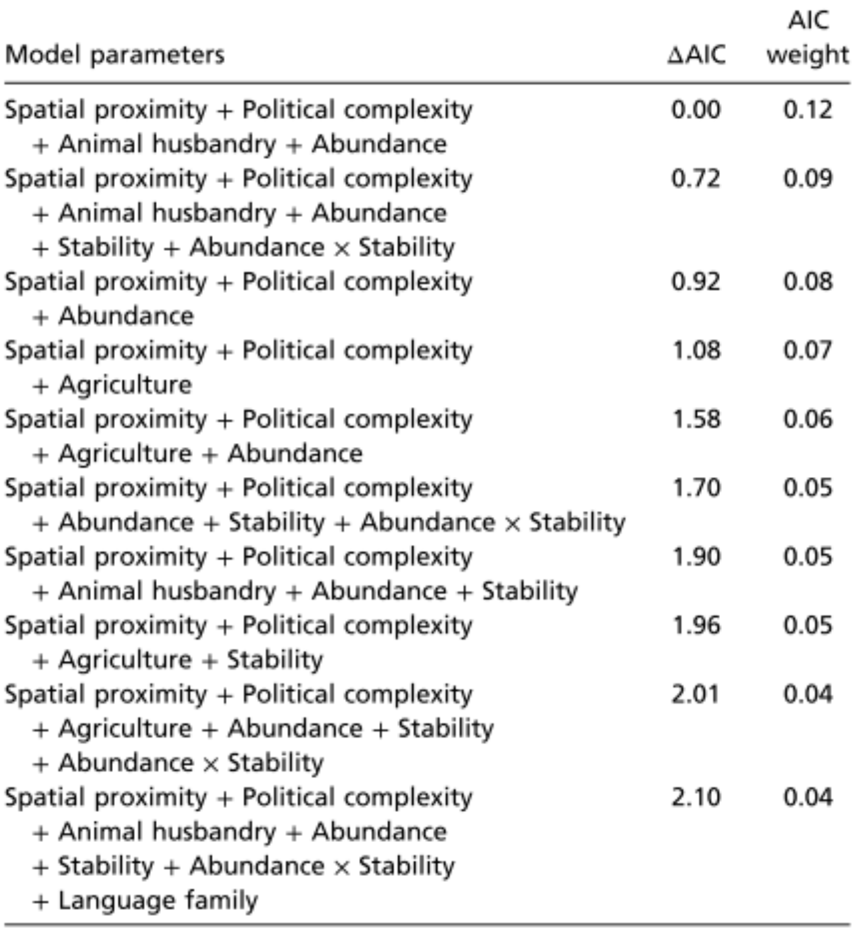
-Multi-model inference and model averaging in situations where there is no clear best model



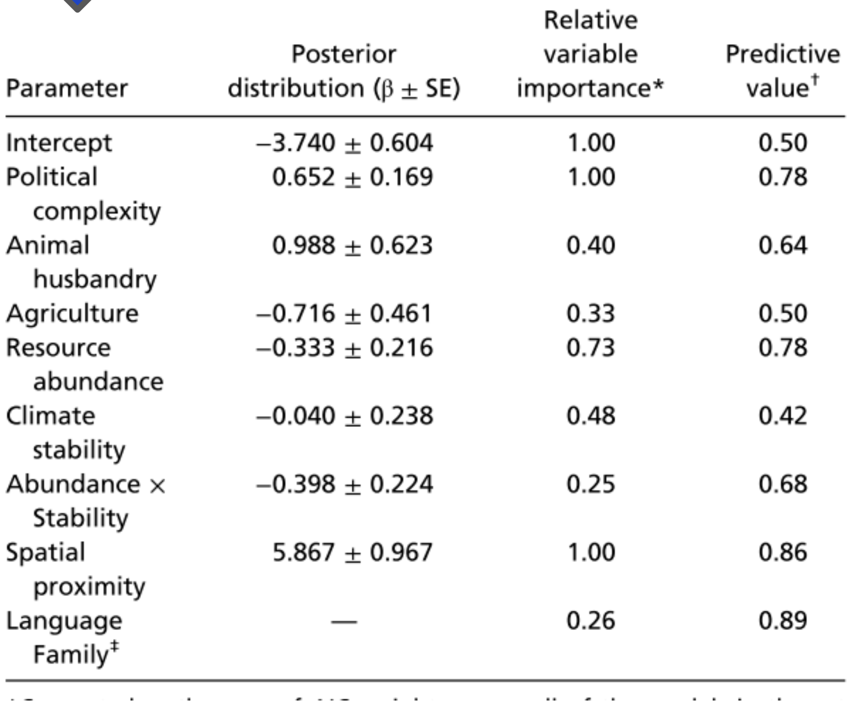
The ecology of religious beliefs

Carlos A. Botero[a,b,1], Beth Gardner[c], Kathryn R. Kirby[d], Joseph Bulbulia[e], Michael C. Gavin[f], and Russell D. Gray[g,h,i]

Model selection: Multi-model inference



Model parameters	ΔAIC	AIC weight
Spatial proximity + Political complexity + Animal husbandry + Abundance	0.00	0.12
Spatial proximity + Political complexity + Animal husbandry + Abundance + Stability + Abundance × Stability	0.72	0.09
Spatial proximity + Political complexity + Abundance	0.92	0.08
Spatial proximity + Political complexity + Agriculture	1.08	0.07
Spatial proximity + Political complexity + Agriculture + Abundance	1.58	0.06
Spatial proximity + Political complexity + Abundance + Stability + Abundance × Stability	1.70	0.05
Spatial proximity + Political complexity + Animal husbandry + Abundance + Stability	1.90	0.05
Spatial proximity + Political complexity + Agriculture + Stability	1.96	0.05
Spatial proximity + Political complexity + Agriculture + Abundance + Stability + Abundance × Stability	2.01	0.04
Spatial proximity + Political complexity + Animal husbandry + Abundance + Stability + Abundance × Stability + Language family	2.10	0.04



Parameter	Posterior distribution (β ± SE)	Relative variable importance*	Predictive value†
Intercept	−3.740 ± 0.604	1.00	0.50
Political complexity	0.652 ± 0.169	1.00	0.78
Animal husbandry	0.988 ± 0.623	0.40	0.64
Agriculture	−0.716 ± 0.461	0.33	0.50
Resource abundance	−0.333 ± 0.216	0.73	0.78
Climate stability	−0.040 ± 0.238	0.48	0.42
Abundance × Stability	−0.398 ± 0.224	0.25	0.68
Spatial proximity	5.867 ± 0.967	1.00	0.86
Language Family‡	—	0.26	0.89

Anderson, D. R., & Burnham, K. P. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*

The ecology of religious beliefs

Carlos A. Botero[a,b,1], Beth Gardner[c], Kathryn R. Kirby[d], Joseph Bulbulia[e], Michael C. Gavin[f], and Russell D. Gray[g,h,i]

This paper also highlights:

-That oftentimes when analysing cross cultural and linguistic data we need models with non-Gaussian (i.e., non-normal) error distributions (e.g., we might need to model presence/absence of a binary trait, or a Poisson distribution of counts)

The ecology of religious beliefs

Carlos A. Botero[a,b,1], Beth Gardner[c], Kathryn R. Kirby[d], Joseph Bulbulia[e], Michael C. Gavin[f], and Russell D. Gray[g,h,i]

This paper also highlights:

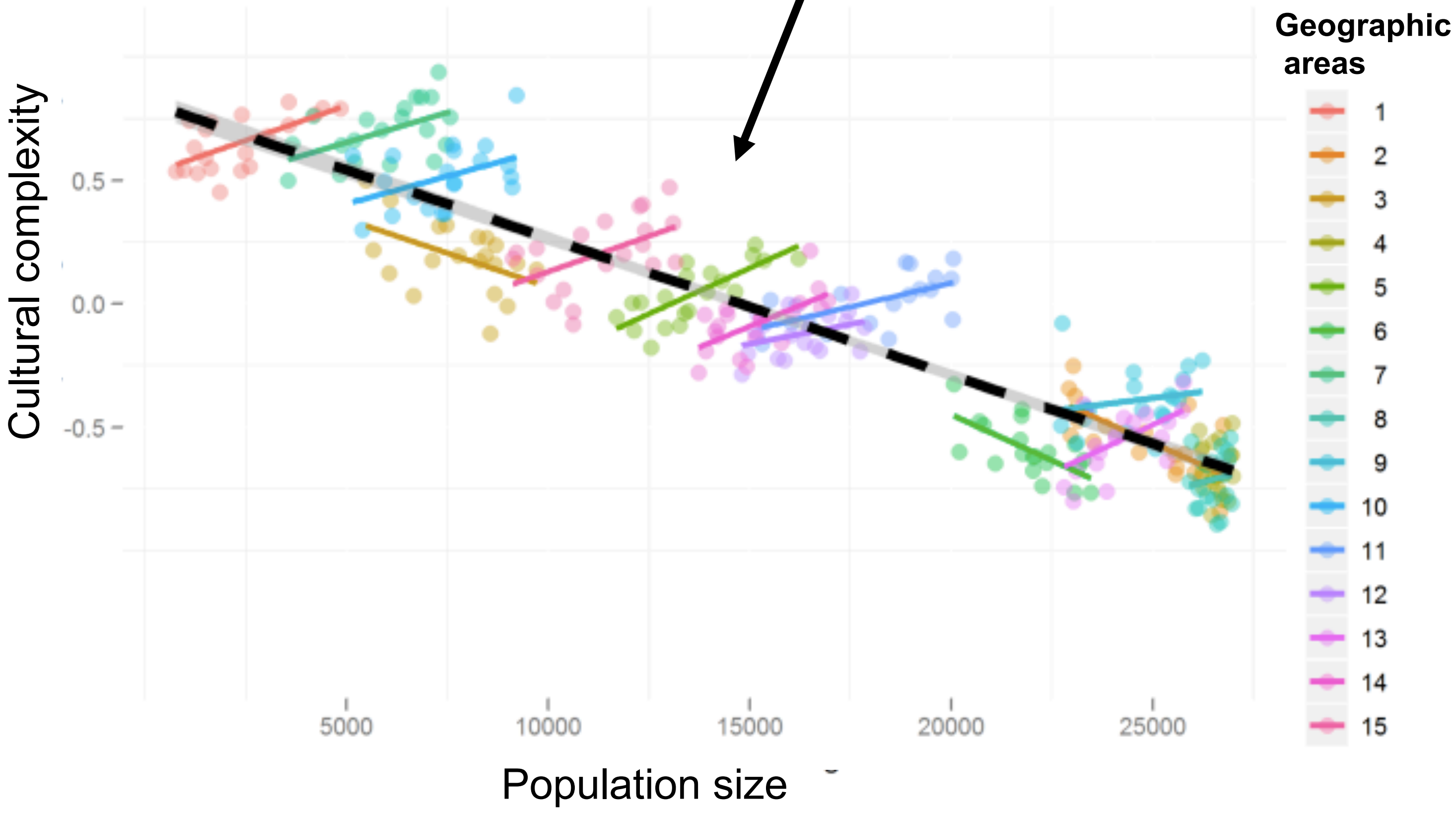
-That oftentimes when analysing cross cultural and linguistic data we need models with non-Gaussian (i.e., non-normal) error distributions (e.g., we might need to model presence/absence of a binary trait, or a Poisson distribution of counts)

-Although not a phylogenetic model, the inclusion of language family as a random effect, and the issue of modelling spatial clustering among societies, also highlights the requirement for multilevel models in cross-cultural phylogenetic analyses.

PHYLOGENETIC MULTILEVEL REGRESSION

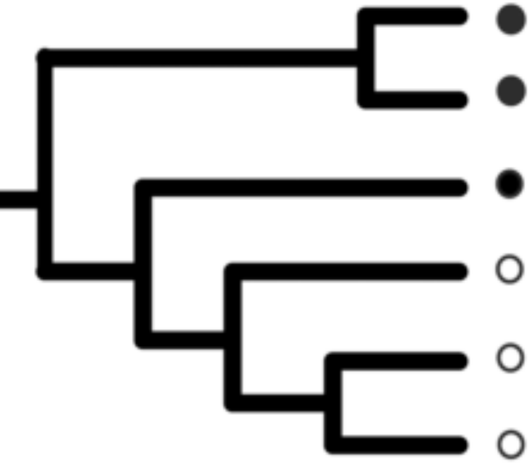
Clustering factors

We need a random effect to model the clustering by geographic areas



PHYLOGENETIC MULTILEVEL REGRESSION

Repeated measures

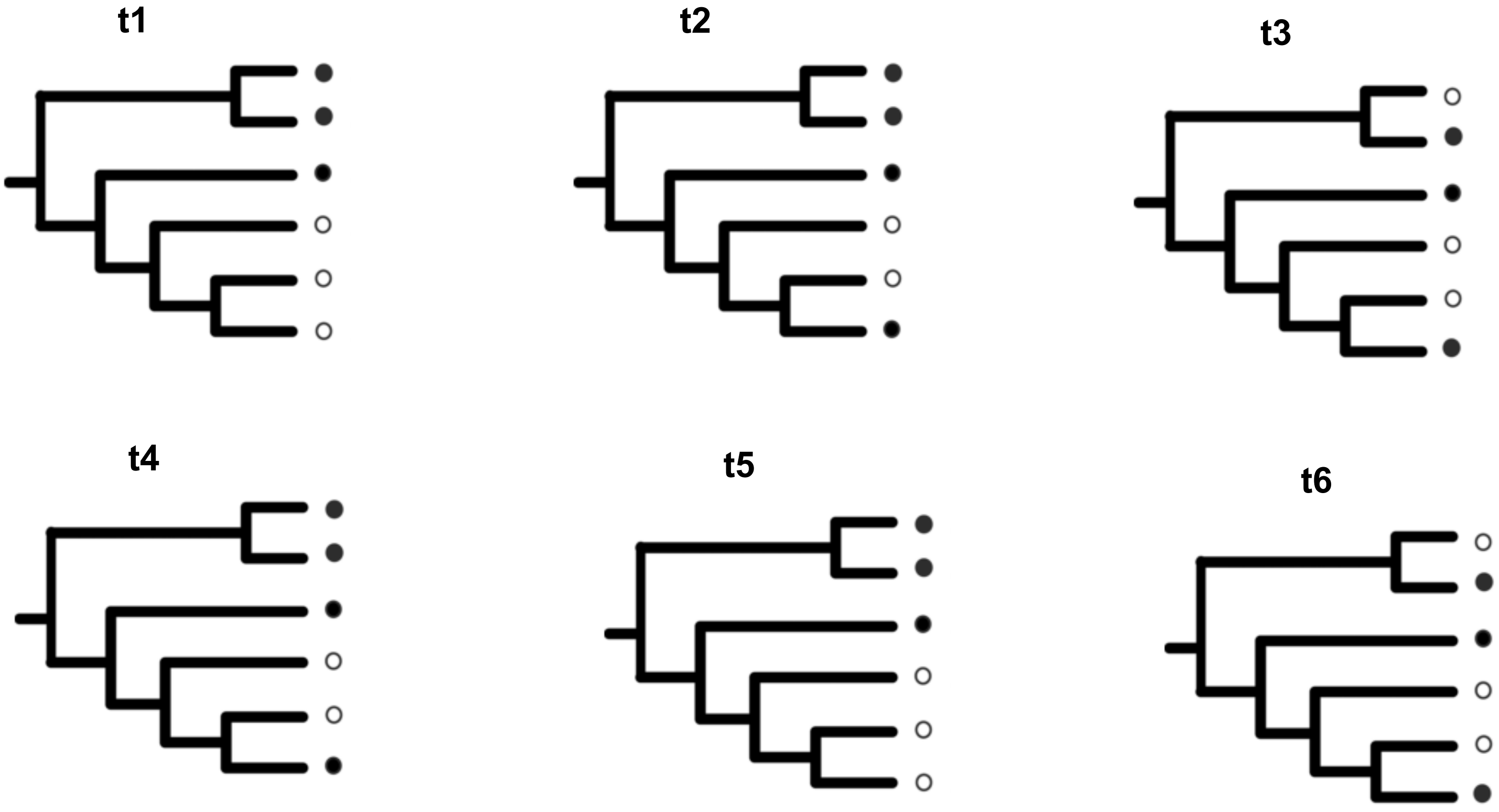


So far we’ve only considered datasets where each linguistic group/society is represented just once

PHYLOGENETIC MULTILEVEL REGRESSION

Repeated measures

But what if there are repeated measures taken across time (t_i)?

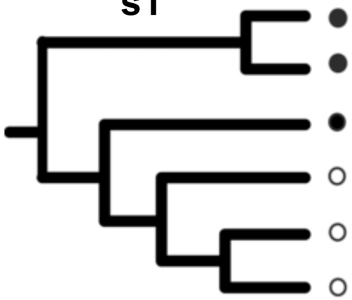


PHYLOGENETIC MULTILEVEL REGRESSION

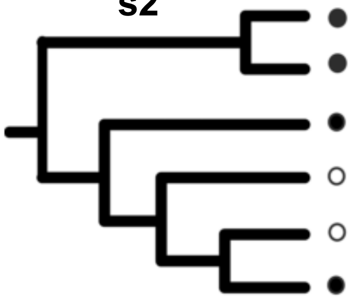
Repeated measures

But what if there are repeated measures taken across time (t_i)? Or space (s_i)?

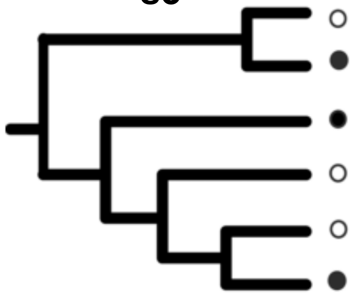
t1
s1



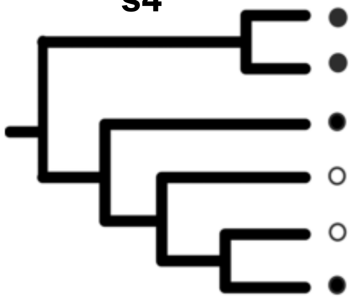
t2
s2



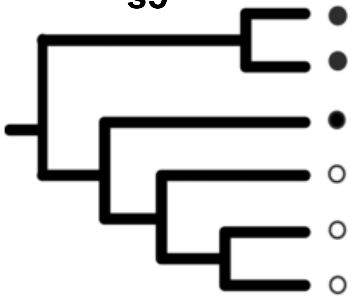
t3
s3



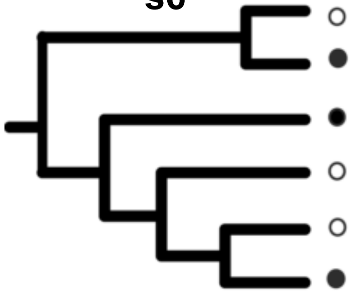
t4
s4



t5
s5



t6
s6

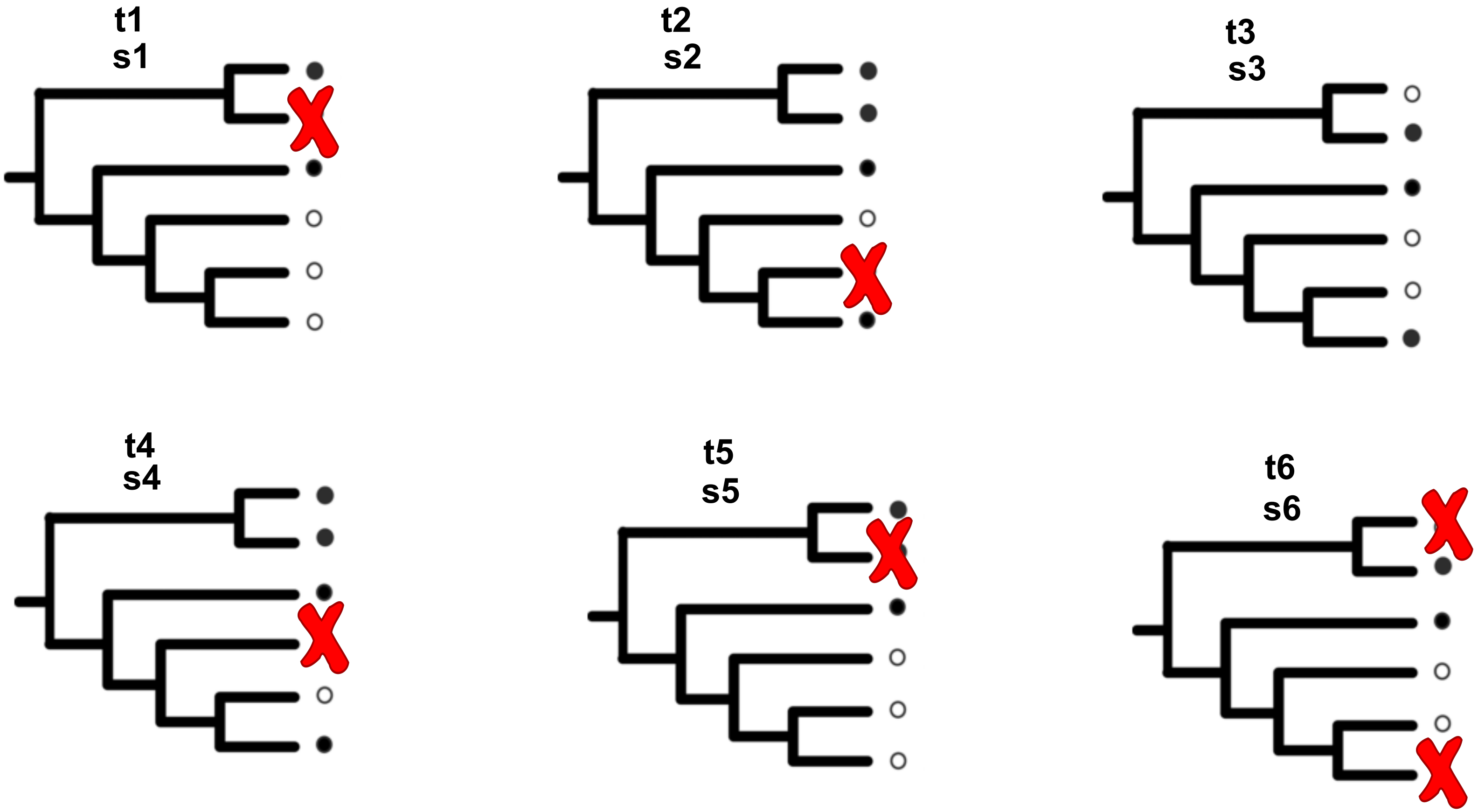


PHYLOGENETIC MULTILEVEL REGRESSION

Repeated measures

But what if there are repeated measures taken across time (t_i)? Or space (s_i)?

And what if some measurements are missing leading to data imbalance?

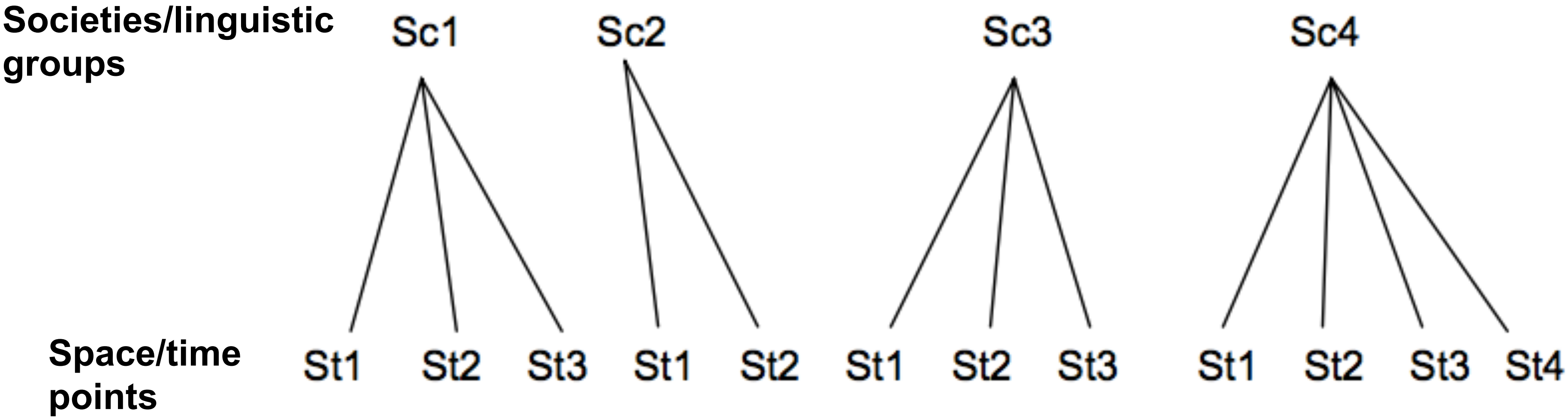


PHYLOGENETIC MULTILEVEL REGRESSION

Repeated measures

But what if there are repeated measures taken across time (t_i)? Or space (s_i)?

And what if some measurements are missing leading to data imbalance?



Societies/linguistic groups

Sc1 Sc2 Sc3 Sc4

Space/time points

St1 St2 St3 St1 St2 St1 St2 St3 St1 St2 St3 St4

<u>We need multilevel models to deal with these types of issues!</u>

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

Random effect/variance component

STATISTICAL PACKAGES AVAILABLE FOR PERFORMING PHYLOGENETIC REGRESSION

- Phylogenetic Generalized Least Squares (PGLS): R package caper (Orme D. 2013: https://cran.r-project.org/ web/packages/caper/)

- BayesTraits: Pagel, M., & Meade, A. 2017 BayesTraits (Version 3.0). Reading, UK. www.evolution.rdg.ac.uk

- Phylogenetic multilevel model (pGLMM): R package MCMCglmm (Hadfield 2010: https://cran.rproject.org/web/packages/MCMCglmm/MCMCglmm)

This is not an exhaustive list! But these are some of the current state-of-art packages you probably want to consider!

STATISTICAL PACKAGES AVAILABLE FOR PERFORMING PHYLOGENETIC REGRESSION

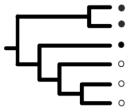
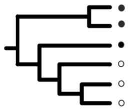
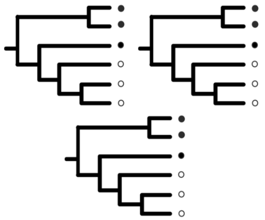
	Single-level **Gaussian (i.e., continuous dependent variable)**
PGLS	✓
BayesTraits	✓
MCMCglmm	✓

Nb. MCMCglmm incorportates phylo. signal as a random effect/variance component, PGLS/BT in the residual error

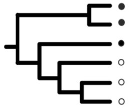
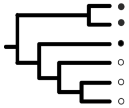
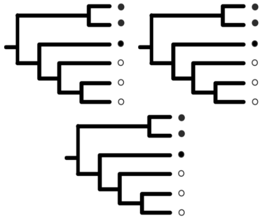
STATISTICAL PACKAGES AVAILABLE FOR PERFORMING PHYLOGENETIC REGRESSION

	Single-level Gaussian (i.e., continuous dependent variable)	Single-level Non-Gaussian (i.e., binary, binomial, Poisson)
PGLS	✓	✗
BayesTraits	✓	✗
MCMCglmm	✓	✓

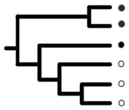
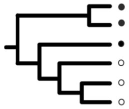
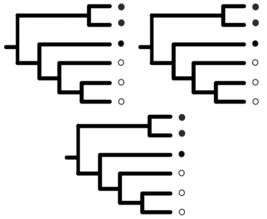
STATISTICAL PACKAGES AVAILABLE FOR PERFORMING PHYLOGENETIC REGRESSION

	Single-level  Gaussian (i.e., continuous dependent variable)	Single-level  Non-Gaussian (i.e., binary, binomial, Poisson)	Multilevel  Gaussian&non-Gaussian
PGLS	✓	✗	✗
BayesTraits	✓	✗	✗
MCMCglmm	✓	✓	✓

STATISTICAL PACKAGES AVAILABLE FOR PERFORMING PHYLOGENETIC REGRESSION

	Single-level  Gaussian (i.e., continuous dependent variable)	Single-level  Non-Gaussian (i.e., binary, binomial, Poisson)	Multilevel  Gaussian&non-Gaussian	Incorporates posterior tree sample??
PGLS	✓	✗	✗	✗
BayesTraits	✓	✗	✗	✓
MCMCglmm	✓	✓	✓	✓ ?

STATISTICAL PACKAGES AVAILABLE FOR PERFORMING PHYLOGENETIC REGRESSION

	Single-level  Gaussian (i.e., continuous dependent variable)	Single-level  Non-Gaussian (i.e., binary, binomial, Poisson)	Multilevel  Gaussian&non-Gaussian	Incorporates posterior tree sample??	Estimate Phylo. signal
PGLS	✓	✗	✗	✗	✓
BayesTraits	✓	✗	✗	✓	✓
MCMCglmm	✓	✓	✓	✓ ?	✓

….. AND THE WINNER IS…..

Cool kid ‘MCMCglmm’ (Hadfield 2010) ☺

MCMCglmm (Hadfield, 2010)

Why should you hang out with the cool kid?

It can incorporate a large number of different error structures

Including (but not limited to!):

Continuous (Gaussian)

Binary

Poisson

Binomial

Zero-inflated Binomial/Poisson

Multivariate response variables

(Could be useful if interested in modelling more than one response/dependent variable simultaneously: see “Houslay, T. M., & Wilson, A. (2017). Avoiding the misuse of BLUP in behavioral ecology. Published online” for discussion and some worked examples using MCMCglmm)

MCMCglmm (Hadfield, 2010)

Why should you hang out with the cool kid?

An estimate of phylogenetic signal – equivalent to Pagel's λ – is easily obtainable from the posterior distribution of the model variance (see the tutorial example!)

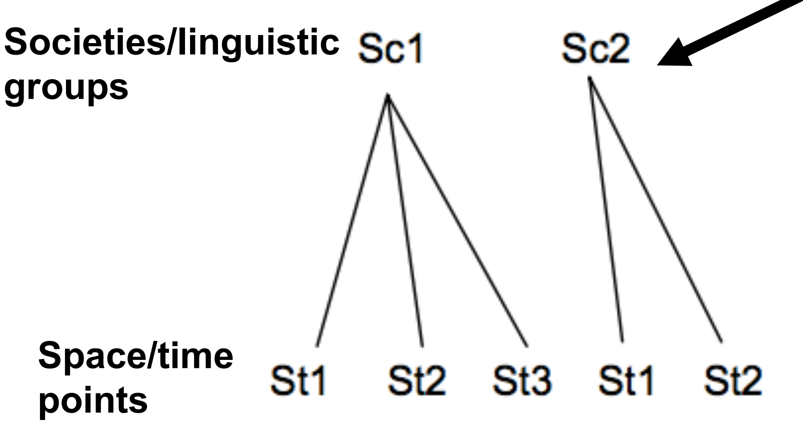
+ Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology*, *23*(3), 494-508.

MCMCglmm (Hadfield, 2010)

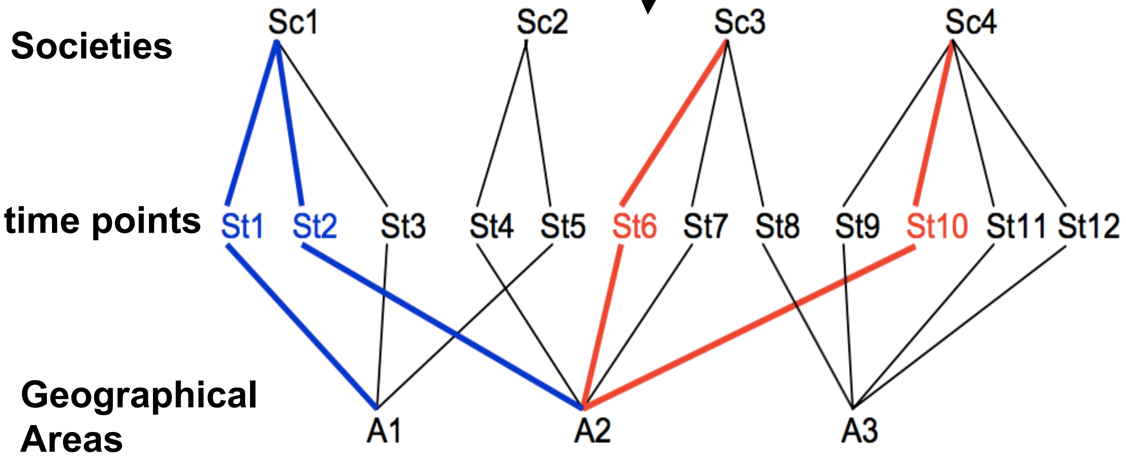
Why should you hang out with the cool kid?

It can easily accommodate a range of model random effects structures

Simple nested random effect structure



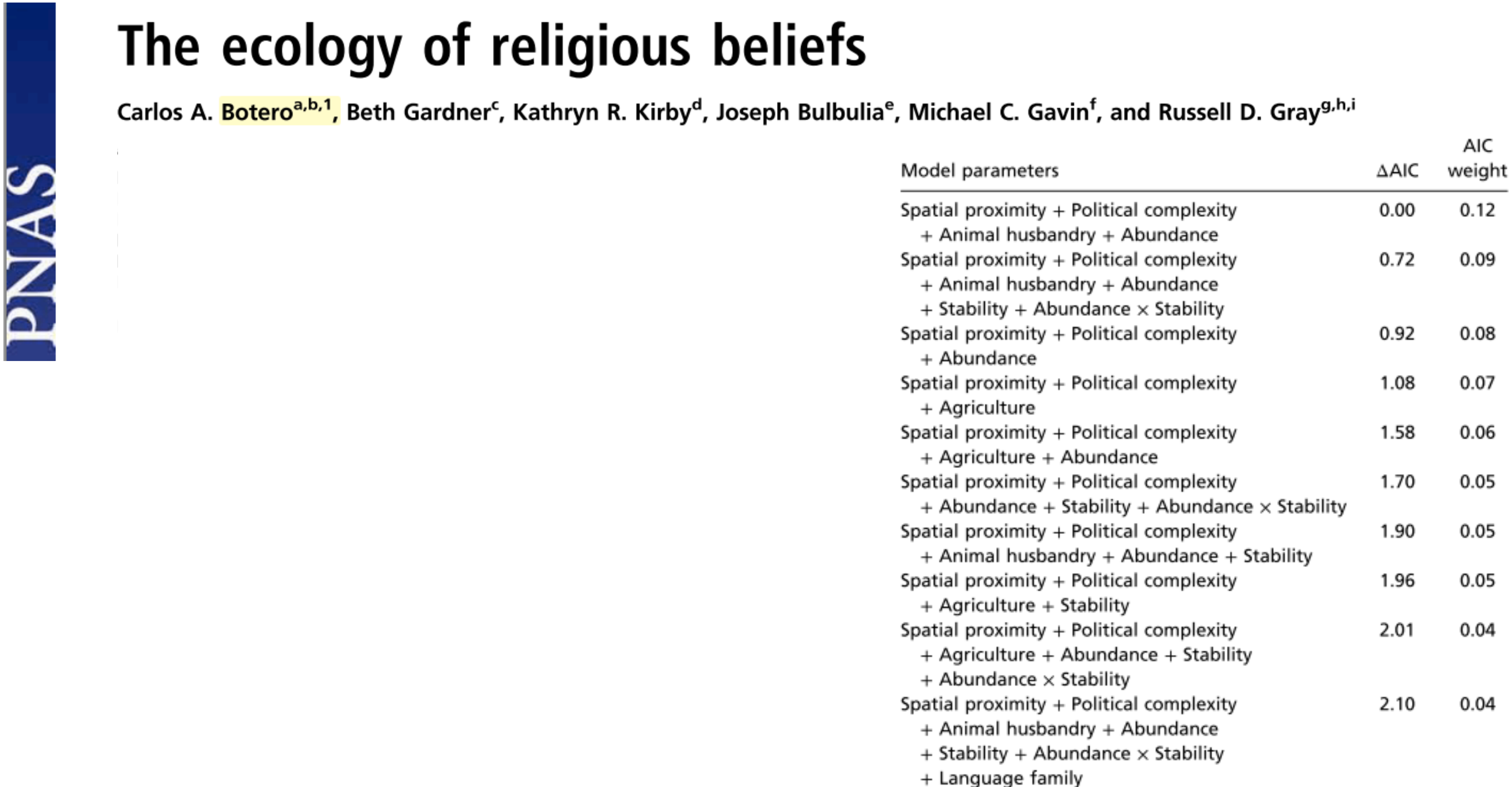
More complex 'crossed' random effect structures



MCMCglmm (Hadfield, 2010)

Why should you hang out with the cool kid?

It is compatible with the R multi-model inference package MuMIn



The ecology of religious beliefs

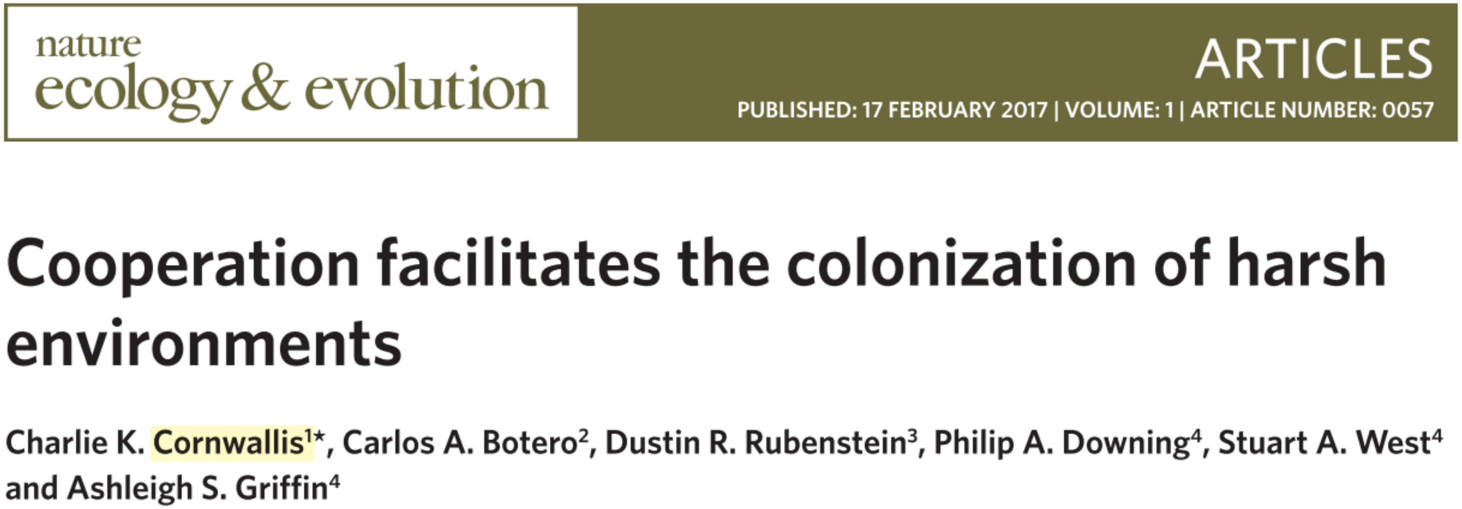
Carlos A. Botero[a,b,1], Beth Gardner[c], Kathryn R. Kirby[d], Joseph Bulbulia[e], Michael C. Gavin[f], and Russell D. Gray[g,h,i]

Model parameters	ΔAIC	AIC weight
Spatial proximity + Political complexity + Animal husbandry + Abundance	0.00	0.12
Spatial proximity + Political complexity + Animal husbandry + Abundance + Stability + Abundance × Stability	0.72	0.09
Spatial proximity + Political complexity + Abundance	0.92	0.08
Spatial proximity + Political complexity + Agriculture	1.08	0.07
Spatial proximity + Political complexity + Agriculture + Abundance	1.58	0.06
Spatial proximity + Political complexity + Abundance + Stability + Abundance × Stability	1.70	0.05
Spatial proximity + Political complexity + Animal husbandry + Abundance + Stability	1.90	0.05
Spatial proximity + Political complexity + Agriculture + Stability	1.96	0.05
Spatial proximity + Political complexity + Agriculture + Abundance + Stability + Abundance × Stability	2.01	0.04
Spatial proximity + Political complexity + Animal husbandry + Abundance + Stability + Abundance × Stability + Language family	2.10	0.04

MCMCglmm (Hadfield, 2010)

Why should you hang out with the cool kid?

It appears to offer some functionality for conducting multilevel ancestral state reconstruction!



nature ecology & evolution

ARTICLES

PUBLISHED: 17 FEBRUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0057

Cooperation facilitates the colonization of harsh environments

Charlie K. Cornwallis[1]*, Carlos A. Botero[2], Dustin R. Rubenstein[3], Philip A. Downing[4], Stuart A. West[4] and Ashleigh S. Griffin[4]

I haven’t yet looked at the details, but the package appears to have been used to examine the environmental conditions and mating system that preceded the evolution of cooperative breeding in birds…

MCMCglmm (Hadfield, 2010)

The downside….

Its user manual and course notes are a bit cryptic for somebody not overly familiar with math and Bayesian modelling! And model/prior specifications can quickly become tricky on complex data!

But, the user manual is v. detailed + the author (and others) are v. active in online forums + there is a growing literature on how to use the package and an ever increasing number of publications (in the biological sciences!) where it has been used! – I've provided a few citations but a quick online search will reveal many more!

MCMCglmm (Hadfield, 2010)

Useful resources for familiarization with the package:

Overviews/tutorials:

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1-22.

Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology*, *23*(3), 494-508.

MCMCglmm (course notes):https://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf

de Villemereuil (2010) Tutorial: Estimation of a biological trait heritability using the animal model: How to use the MCMCglmm R package: http://devillemereuil.legtux.org/wp-content/uploads/2012/12/tuto_en.pdf

Garamszegi, L. Z. (2014). Modern phylogenetic comparative methods and their application in evolutionary biology. *Concepts and Practice. London, UK: Springer*.

MCMCglmm (Hadfield, 2010)

Useful resources for familiarization with the package:

Using MCMCglmm with a posterior tree sample
https://github.com/TGuillerme/mulTree

Healy, K., Guillerme, T., Finlay, S., Kane, A., Kelly, S. B., McClean, D., ... & Cooper, N. (2014). Ecology and mode-of-life explain lifespan variation in birds and mammals. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1784), 20140298.

Using MCMCglmm with a binary response variable
Advice about priors:
de Villemereuil, P., Gimenez, O. & Doligez, B. (2012). Comparing parent-offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: a simulation study for Gaussian and binary traits. *Methods Ecol. Evol.*, **4**, 260–275

Hadfield, J.D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.*, **33**, 1–22.

A nice example:
Capellini, I., Baker, J., Allen, W. L., Street, S. E., & Venditti, C. (2015). The role of life history traits in mammalian invasion success. *Ecology letters*, *18*(10), 1099-1107.