Conditional and marginal models for analysing light interception data

Rafael Moral[1], Wagner Bonat[2], John Hinde[3], Clarice Demétrio[1] and Marina Duarte[1]

[1]University of São Paulo, Brazil; [2]Federal University of Paraná, Brazil; [3]NUI Galway, Ireland

Abstract

Here, we analyse data from an experiment in Biodiversity and Ecosystem Functioning theory. The response variable is the percentage of intercepted light by the canopy in forest patches that were previously restored with three different numbers of species (treatments). Because the experimental design includes multilevel and longitudinal sampling, we propose two different approaches: (i) a conditionally specified beta mixed model; and (ii) a marginally specified model where we include the covariance structure directly. While the interpretation of the models resulting from these approaches differ, both strategies offer advantages for modelling these data, as well as presenting some potential drawbacks.

Keywords: BEF theory; beta mixed models; multivariate covariance generalized linear models

Introduction

According to the Biodiversity and Ecosystem Functioning (BEF) theory, with a higher species diversity there is an increase in the number and intensity of ecosystem functions, which include processes of energy flow, nutrient cycling, decomposition, and organic matter production (Cardinale et al., 2012). This is due to a more efficient use of resources, which allow for different pathways for ecological processes across time and space. For forest restoration practitioners, the application of this theory could be very useful. When developing a plan for forest restoration it is desirable to re-establish ecological processes that can maintain a forest over time, without the need for any kind of management, such as manuring, irrigation, or pest control. An important variable in this context is light interception, which is an indicator of self-sustainability of a forest. Higher light interception boosts photosynthesis and, consequently, biomass and carbon storage by trees, which are important processes of a self-sustainable system, and a target of forest restoration practices.

To assess whether higher tree diversity promotes (i) higher light interception and (ii) a more even distribution of light, both horizontally and vertically in a forest, an experiment was conducted in patches of restored Atlantic Forest in Brazil. There were three different levels of species richness, 20, 60, and 117 species, with four replicates in a completely randomized design. In each plot, twelve subplots were sampled at 0, 1, 2, 3, and 4 metres high, hence giving a form of longitudinal (height) study. The observed variable was the percentage of light interception by the canopy. The statistical analysis of this experiment is challenging, since we have a continuous bounded response variable along with multilevel and longitudinal structures.

The main goal of this work is to propose and compare two approaches to analyse continuous bounded data. First, we propose conditionally specified beta mixed models, where we include random effects to incorporate the correlation between observations made within the same plot, as well as among the longitudinal observations made within each subplot. Second, we propose marginally specified models, where we model the marginal covariance structure directly using a linear combination of known matrices. Finally, we discuss different computational strategies for fitting these models and pinpoint advantages and potential drawbacks of both approaches.

Methodology

Conditional approach

Let y_{ijkl} be the percentage of light intercepted by the canopy at the j-th replicate, $j = 1, \ldots, 4$, of the i-th treatment, $i = 1, 2, 3$, measured at the k-th subplot, $k = 1, \ldots, 12$, at the l-th height level, $l = 1, \ldots, 5$. The Beta distribution is a reasonable assumption for modelling the response variable, since it is bounded in the $(0, 1)$ interval, with density function

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\{(1-\mu)\phi\}} y^{\mu\phi - 1}(1-y)^{(1-\mu)\phi - 1}, \quad 0 < y < 1$$

with $\mathrm{E}(Y) = \mu \in (0, 1)$ and $\mathrm{Var}(Y) = \frac{\mu(1-\mu)}{\phi+1}$, where $\phi > 0$ (Cribari-Neto and Zeileis, 2010)

Clearly longitudinal observations taken on the same subplot may be correlated, as may all observations made on the same plot. To accommodate these correlations we include random intercepts, b_{1ij}, for observations in the same plot, and random intercepts, b_{2ijk}, and slopes, b_{3ijk}, for observations on the same subplot. Then, we take the conditional distribution of $Y_{ijkl} | b_{1ij}, b_{2ijk}, b_{3ijk}$ as $\mathrm{Beta}(\mu_{ijkl}, \phi_{ijkl})$, with $b_{1ij} \sim \mathrm{N}(0, \sigma_1^2)$,

$$\left[\begin{array}{c} b_{2ijk} \\ b_{3ijk} \end{array}\right] \sim \mathrm{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma_2^2 & \sigma_{23} \\ \sigma_{23} & \sigma_3^2 \end{array}\right]\right),$$

and linear predictors

$$\begin{aligned} \log\left(\frac{\mu_{ijkl}}{1-\mu_{ijkl}}\right) &= \beta_{0i} + b_{1ij} + b_{2ijk} + (\beta_{1i} + b_{3ijk})h_l \\ \log(\phi_{ijkl}) &= \gamma_{0i} + \gamma_{1i}h_l \end{aligned}$$

for the mean and dispersion parameters. Here h_l are the height levels and β_{0i} and β_{1i} are different intercepts and slopes over height per treatment for the mean parameter, while γ_{0i} and γ_{1i} are different intercepts and slopes over height per treatment for the dispersion parameter.

We may write the log-likelihood as

$$l(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{3}\sum_{j=1}^{4} \log\left\{\int_{-\infty}^{\infty}\prod_{k=1}^{12}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{l=1}^{5} f(y_{ijkl}; \mu_{ijkl}, \phi_{jikl}) f(b_{2ijk}, b_{3ijk}) db_{2ijk} db_{3ijk}\right) db_{1ij}\right\},$$

with parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \boldsymbol{\gamma_0}, \boldsymbol{\gamma_1}, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_{23})^\top$. The implementation of this model and likelihood was made in R (R Core Team, 2017), and we used the Laplace approximation to compute the integrals in the log-likelihood.

Marginal approach

Let $\mathbf{y}$ be the $n \times 1$ vector of percentages of light interception and $\mathbf{X}$ the $n \times p$ design matrix with $\boldsymbol{\beta}$ the $p \times 1$ parameter vector, including different intercepts and slopes over height per treatment. Take $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = g_1^{-1}(\mathbf{X}\boldsymbol{\beta})$, with $g_1^{-1}(\cdot)$ the inverse logit function, and $\mathrm{Var}(\mathbf{Y}) = \boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{\mu})^{\frac{1}{2}}(\tau_0\mathbf{I})\mathrm{V}(\boldsymbol{\mu})^{\frac{1}{2}}$, with $\mathrm{V}(\boldsymbol{\mu}) = \mathrm{diag}\{\mu_i(1-\mu_i)\}$, $i = 1, \ldots, n$, and $\tau_0 = \frac{1}{1+\phi}$, with ϕ a dispersion parameter. This corresponds to the variance-covariance matrix of a beta regression model for independent observations. We now change the identity matrix $\mathbf{I}$ to a non-diagonal matrix $\boldsymbol{\Omega}(\boldsymbol{\tau})$, with $\boldsymbol{\tau}$ a $D \times 1$ dispersion parameter vector, giving $\mathrm{Var}(\mathbf{Y}) = \boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{\mu})^{\frac{1}{2}}(\boldsymbol{\Omega}(\boldsymbol{\tau}))\mathrm{V}(\boldsymbol{\mu})^{\frac{1}{2}}$. We can model $\boldsymbol{\Omega}(\boldsymbol{\tau})$ as a linear combination of known matrices $Z_1, \ldots, Z_D$, $g_2(\boldsymbol{\Omega}(\boldsymbol{\tau})) = \sum_{d=0}^{D} \tau_d Z_d$, where $g_2(\cdot)$ is a covariance link function. Here, we build the matrices such that observations within the same subsample are correlated, and observations taken on the same plot are also correlated.

We now form two estimating equations, namely a quasi-score function for the regression parameters, $\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \frac{d}{d\boldsymbol{\beta}}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, and a Pearson estimating function for the dispersion parameters, $\psi_{\boldsymbol{\tau}}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \text{tr}\left\{-\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \boldsymbol{\tau}}\left((\mathbf{y} - \boldsymbol{\mu})^\top(\mathbf{y} - \boldsymbol{\mu}) - \boldsymbol{\Sigma}\right)\right\}$. The asymptotic distribution of the joint solution $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}})^\top$ is $\text{N}(\boldsymbol{\theta}, \text{J}_{\boldsymbol{\theta}}^{-1})$, where $\text{J}_{\boldsymbol{\theta}}^{-1}$ is the inverse of the Godambe information matrix $\text{J}_{\boldsymbol{\theta}}^{-1} = \text{S}_{\boldsymbol{\theta}}^{-1}\text{V}_{\boldsymbol{\theta}}(\text{S}_{\boldsymbol{\theta}}^{-1})^\top$, with $\text{S}_{\boldsymbol{\theta}}$ and $\text{V}_{\boldsymbol{\theta}}$ the sensitivity and variability matrices, respectively.

Bonat and Jørgensen (2016) devise a modified Chaser algorithm to estimate the parameters, giving iterative updates as

$$\begin{aligned}\boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - \text{S}_{\boldsymbol{\beta}}^{-1}\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\tau}^{(i)}) \\ \boldsymbol{\tau}^{(i+1)} &= \boldsymbol{\tau}^{(i)} - \alpha \text{S}_{\boldsymbol{\tau}}^{-1}\psi_{\boldsymbol{\tau}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\tau}^{(i)}),\end{aligned}$$

where α a tuning parameter. This is implemented in the R package `mcglm` (Bonat, 2017).

Results and Discussion

Using both approaches, we fitted the model considering all effects described above and nested sub-models, removing covariates from the mean and/or dispersion parameters, as well as random effects/covariance parameters. We compared the fitted models using the AIC (for the conditional approach) and the pseudo AIC, based on the Gaussian pseudo-log-likelihood (for the marginal approach).

Looking at the AIC and pseudo AIC values (see Table 1), for both modelling approaches model $M6$ is selected, i.e., different intercepts and slopes over height per treatment are included in both the linear predictors for the mean and dispersion. Moreover, all covariance parameters are included except the covariance between all observations taken on the same plot, which is the plot random effect for the conditional modelling approach.

Table 1: Specification of seven different models, fitted using both the conditional and marginal approaches, relating to the inclusion of different intercepts and slopes over height per treatment in the mean (μ) and dispersion (ϕ) parameters, with corresponding AIC and pseudo AIC values

Model	Covariates		Variance components				Conditional	Marginal
	μ	ϕ	σ_1^2/τ_6	σ_2^2/τ_7	σ_3^2/τ_8	σ_{23}/τ_9	AIC	pAIC[1]
$M1$	×	×	×	×	×	×	−701.27	−159.36
$M2$	yes	×	×	×	×	×	−875.78	−398.66
$M3$	yes	×	×	yes	yes	yes	−1188.09	−751.62
$M4$	yes	×	yes	yes	yes	yes	−1187.53	−751.20
$M5$	yes	yes	×	×	×	×	−918.38	−451.20
$M6$	yes	yes	×	yes	yes	yes	−1231.05	−768.94
$M7$	yes	yes	yes	yes	yes	yes	−1230.48	−768.38

[1] pseudo AIC

We note that the proportion of light intercepted by the canopy gets smaller as the height increases (see Figure 1). Patches with a higher number of species intercept more light, giving evidence that higher tree diversity promotes more light interception and therefore makes it more likely for systems to be self-sustainable. Additionally, the variability increases over height and is smaller for patches with higher diversity (see the estimates for the dispersion parameter covariates, Table 2), giving evidence of a niche complementarity effect.

The two modelling strategies considered here provided the same qualitative answers to the research questions. However, the interpretation of the covariance parameters is different, and this is due to model formulation. It is computationally faster to fit the marginal models,

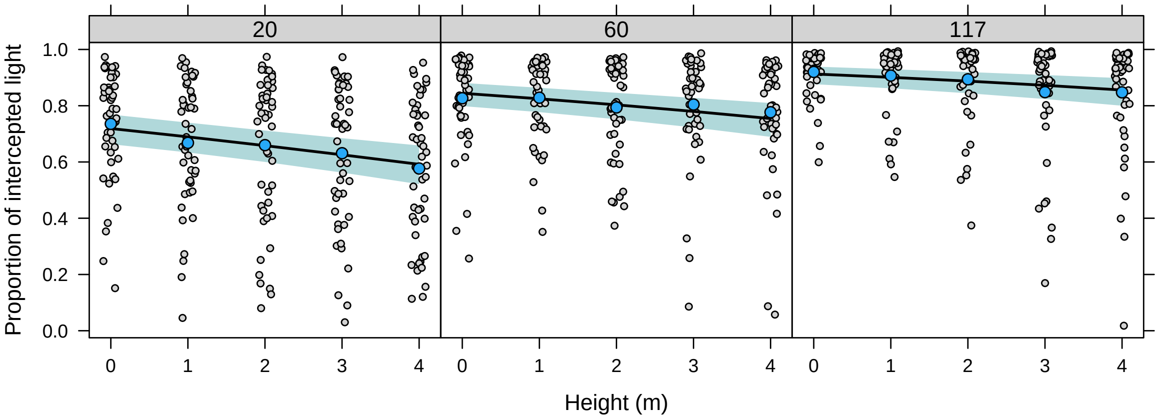

Figure 1: Light interception data: observed responses (grey points), means (blue points), fitted curves (black lines) and confidence intervals (blue shaded) estimated by the conditional model, for each treatment (number of species).

Table 2: Parameter estimates for intercepts and slopes for the mean parameter ($\boldsymbol{\beta}$) for the conditional and marginal models, and dispersion parameter ($\boldsymbol{\gamma}$) and variance components for the conditional model and covariance parameters ($\boldsymbol{\tau}$) for the marginal model.

Parameter	Estimates		Parameter	Estimates	
	Conditional	Marginal		Conditional	Marginal
β_{01}	1.07 (0.15)	0.96 (0.13)	γ_{01}/τ_0	2.55 (0.22)	0.07 (0.03)
β_{02}	1.78 (0.14)	1.59 (0.15)	γ_{02}/τ_1	3.12 (0.17)	−0.01 (0.04)
β_{03}	2.77 (0.14)	2.46 (0.19)	γ_{03}/τ_2	3.93 (0.23)	−0.04 (0.04)
β_{11}	−0.18 (0.05)	−0.16 (0.04)	γ_{11}/τ_3	−0.14 (0.09)	0.01 (0.00)
β_{12}	−0.08 (0.04)	−0.08 (0.05)	γ_{12}/τ_4	−0.06 (0.06)	−0.01 (0.02)
β_{13}	−0.21 (0.05)	−0.20 (0.06)	γ_{13}/τ_5	−0.40 (0.10)	0.00 (0.02)
			σ^2_2/τ_7	0.76 (0.12)	0.12 (0.02)
			σ^2_3/τ_8	0.07 (0.01)	0.01 (0.00)
			ρ_{23}/τ_9	−0.39 (0.10)	−0.01 (0.01)

because there is no need to approximate the high-dimensional integrals in the likelihood function. However, since the model is defined by using first and second moment assumptions, there is no corresponding probability distribution and how to carry out simulation studies here is subject of ongoing research. Further work also includes optimization of the code written for fitting the conditional models, and further exploration of the behaviour of both modelling approaches.

References

Bonat, W.H. (2017) mcglm: Multivariate covariance generalized linear models, R package version 0.3.0. URL `https://cran.r-project.org/package=mcglm`

Bonat, W.H.; Jørgensen, B. (2016) Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society C*, 65:649–675.

Cardinale, B.J.; Duffy, J.E.; Gonzalez, A.; Hooper, D.U.; Perrings, C.; Venail, P.; Narwani, A.; Mace, G.M.; Tilman, D.; Wardle, D.; Kinzig, A.P.; Daily, G.C.; Loreau, M.; Grace, J.B.; Larigauderie, A.; Srivastava, D.S.; Naeem, S. (2012) Biodiversity loss and its impact on humanity. *Nature*, 486:59–67.

Cribari-Neto, F.; Zeileis, A. (2010) Beta regression in R. *Journal of Statistical Software*, 34:1–24.

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`