HW#2 Generalized Linear Mixed Models with TMB

Neal Marquez

In order to test the effects of mis-specifying a model on a parameters standard error estimate we will simulate data that resembles collecting data on ten counts of a species at 10 different sites where sites have some biological difference in habitat suitability. The data follows the following distribution.

$$log(\gamma_s) \sim \mathcal{N}(\mu, 1)$$
$$log(\lambda_{s,c}) \sim \mathcal{N}(log(\gamma_s), .5)$$
$$y_{s,c} \sim Poisson(\lambda_{s,c})$$

μ is the log mean of the expected counts and has a set value of 2.
$log(\gamma_s)$ is the log mean for site s.
$\lambda_{s,c}$ is the expected count observation for site s count c.

We will attempt to estimate μ using four different models.

1. a generalized linear model, with only an intercept term **glm**

$$\hat{\lambda_{s,c}} = exp(\hat{\mu})$$

2. a GLMM with only among-site variability **glmm_site**

$$\hat{\lambda_{s,c}} = exp(\hat{\mu} + \zeta_s)$$
$$\zeta \sim \mathcal{N}(0, \sigma^s)$$

3. a GLMM with only overdispersion **glmm_ind**

$$\hat{\lambda_{s,c}} = exp(\hat{\mu} + \epsilon_{s,c})$$
$$\epsilon \sim \mathcal{N}(0, \sigma^c)$$

4. a GLMM with both among-site variability and overdispersion **glmm_both**

$$\hat{\lambda_{s,c}} = exp(\hat{\mu} + \zeta_s + \epsilon_{s,c})$$
$$\zeta \sim \mathcal{N}(0, \sigma^s); \epsilon \sim \mathcal{N}(0, \sigma^c)$$

In order to get an idea of how often our confidence intervals for μ cover the true value we will simulate 1000 data sets using the specifications above and observe (1) the range of the confidence intervals of our parameter of concern and (2) how often the confidence interval covers the true value.

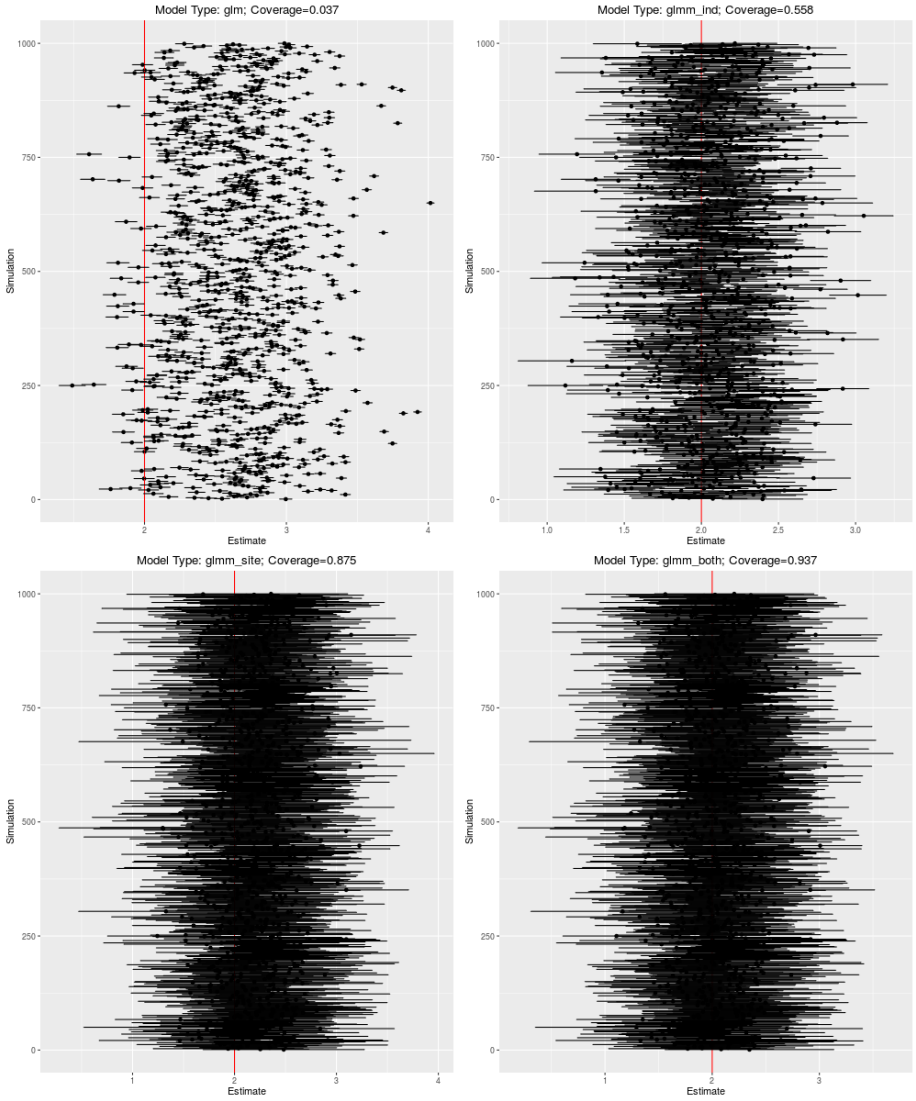

Figure 1: Uncertainty of μ estimate for each model/simulation

Results

Estimates for μ and uncertainty for each model were estimated using REML via INLA. ϵ and ζ were treated as random effects, where applicable.

Using the **glm** model yielded biased and erroneous results that seldom led to confidence intervals that covered the true value of μ. When we start adding in random effects accounting for the site level variance contributed to greater coverage of confidence intervals than including individual level variance. Ultimately using the correct model specifications leads to the uncertainty of parameter estimate $\hat{\mu}$ covering the true value of μ around 95% of the time.

This pattern likely arises because the assumption of the **glm** model is that your errors are independent and normally distributed. The **glmm_site** accounts for the site associativity that is present in the data and why we se ethe large increase in the confidence intervals, even more so than if we added uncertainty in the individual counts as in **glmm_ind**.