Independent Project Model Specification

Wynne Moss

October 31, 2018

The data

SUMMARY OF DATASET: I quantified parasite communities across 10 different pond sites in the East Bay of California; each site was visited 4-6 times within the 2017 summer. At each visit I collected 10-12 individuals from 2 amphibian species. Individuals were measured and parasite infection was quantified.

GROUPING VARIABLES:

- Site (10 ponds) - this is a random variable.
- Visit (6 visits) - I want to treat this as a fixed effect, even though visits are nested (?) within sites (so should it go in as a predictor?)
- Species (2 species) - Again, I want to treat this as a fixed effect, although species are nested(?) within site visit

PREDICTOR VARIABLES:

- Species (2 species) - I think this is a better designation for species
- Body size (snout-vent-length) - A continuous variable (fixed effect) at the individual level
- Developmental stage - A continuous variable at the individual level…problematic since development is measured differently in newts vs. frogs
- Sex - A factor variable at the individual level

RESPONSE VARIABLE:

- The number of parasites found within an individual. For this analysis, just count number of Echinostoma parasites. Can be modeled as Poisson distributed with a negative binomial to account for aggregation. Infection status (1 or 0) could also be modeled as a binary response variable using logistic regression.

Questions

1) Does the impact of species and body size change over the course of the summer (interact with visit)?
2) How much of the variation in overall parasite load is explained by visit-level, species-level, site-level, or individual-level variation?

Display the structure of the data

```
dis <- read.csv("diss.data.2017.csv")
# colnames(dis)
str(dis)
```

```
## 'data.frame':    1049 obs. of  35 variables:
##  $ X                          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ HostCode                   : Factor w/ 1049 levels "PRNTH1_20170328_PSRE_001",..: 113 872 340 755
##  $ Date                       : int  20170513 20170328 20170607 20170327 20170328 20170513 20170513 
##  $ SiteCode                   : Factor w/ 10 levels "PRNTH1","PRNTH4",..: 2 9 4 8 1 3 3 9 9 4 ...
##  $ SpeciesCode                : Factor w/ 2 levels "PSRE","TATO": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CollectionCode             : Factor w/ 53 levels "PRNTH1_20170328",..: 7 45 18 39 1 12 12 45 45 18
##  $ Lifestage                  : Factor w/ 2 levels "Larva","Metamorph": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Dissector                  : Factor w/ 5 levels "AO","CM","DC",..: 2 3 2 3 3 2 2 3 3 2 ...
##  $ DissectionCondition        : Factor w/ 3 levels "Dead on Arrival",..: 3 3 3 1 3 3 3 3 3 2 ...
##  $ GosnerStage                : int  40 26 26 26 26 29 26 26 26 26 ...
##  $ TarichaLarvaeStage         : Factor w/ 6 levels "","2T","3T","4T",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ SVL                        : num  4.58 5.13 5.29 5.41 5.42 5.45 5.6 5.74 5.94 5.95 ...
##  $ TailLength                 : num  27.44 6.44 5.76 6.78 7.15 ...
##  $ TotalLength                : num  32 11.6 11.1 12.2 12.6 ...
##  $ Malformed                  : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex                        : Factor w/ 3 levels "Female","Male",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ collDate                   : Factor w/ 13 levels "2017-03-27","2017-03-28",..: 3 2 7 1 2 3 3 2 2 7
##  $ visit                      : int  2 1 3 1 1 2 2 1 1 3 ...
##  $ SecYr                      : logi  NA NA NA NA NA NA ...
##  $ tot.para                   : int  2 1 2 0 0 4 0 1 1 2 ...
##  $ BDinf                      : int  1 0 0 0 0 0 0 1 1 0 ...
##  $ aveZE                      : num  2.75 0 0 0 0 ...
##  $ Alaria                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Cephalogonimus             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Echinostoma                : int  0 0 0 0 0 2 0 0 0 1 ...
##  $ Gorgoderid_Metacercaria    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gyrinicola_batrachiensis   : int  0 0 0 0 0 3 0 0 0 0 ...
##  $ Manodistomum_syntomentera  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Megalobatrachonema_moraveci: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Nematode                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Oxyurid                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Ribeiroia_ondatrae         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Nyctotherus                : int  1 1 0 0 0 1 0 0 1 0 ...
##  $ Opalina                    : int  1 0 1 0 0 1 0 1 0 0 ...
##  $ Tritrichomonas             : int  0 0 1 0 0 0 0 0 0 1 ...
```

Model formulation

For now, I'll focus on question 1 with the response variable being `Echinostoma` (number of Echinostoma parasites found within that individual)

random intercept

$y_i \sim \text{Poisson}(\mu_{j[i]})$

$\alpha_{i[j]} \sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha)$

$\log(\mu_{j[i]}) = \alpha_{i[j]} + \beta_1\text{SpeciesCode} + \beta_2\text{visit} + \beta_{1,2}(\text{SpeciesCode x Visit}) + \beta_3\text{SVL} + \beta_{2,3}(\text{SVL x Visit}) + (1|\text{SiteCode})$

I am a stumped on the mathematical formulation now that I have random slope and intercept

random slope

$y_i \sim \text{Poisson}(\mu_{i[j]})$

$\log(\mu_{i[j]}) = \alpha_{i[j]} + \beta_1\text{SpeciesCode} + \beta_{i[j]}\text{visit} + \beta_{1,2}(\text{SpeciesCode x Visit}) + \beta_3\text{SVL} + \beta_{2,3}(\text{SVL x Visit})$

$\alpha_{i[j]} \sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha)$

$\beta_{i[j]} \sim \text{Normal}(\mu_\beta, ???)$

Priors for betas:

$\beta_0 \sim$ dnorm(0,10) # according to the model output?

$\beta_1 \sim$ dnorm(0,5) # I think that stan somehow automatically scales

$\beta_2 \sim$ dnorm(0,5)

$\beta_{1,2} \sim$ dnorm(0,5)

$\beta_{2,3} \sim$ dnorm(0,5)

For negative binomial:

$y_i \sim \text{NB}(r, p)$? I think sub r in for μ_i but I'm not sure how p is modeled…

$\text{Var}(y_i) \sim r_i + r_i^2/p$ (Allows variance to increase with mean)

$\log(r) = \beta_0 + \beta_1\text{SpeciesCode} + \beta_2\text{visit} + \beta_{1,2}(\text{SpeciesCode x Visit}) + \beta_3\text{SVL} + \beta_{2,3}(\text{SVL x Visit}) + (\text{visit}|\text{SiteCode})$

Stan formulation

In stan_glm, the model could be written as:

```
stan.fit <- stan_glmer(Echinostoma ~ visit*SpeciesCode + visit*SVL + (visit|SiteCode), data = dis,  fami
summary(stan.fit)
# launch_shinystan(stan.fit)
stan.samp <- sample(stan.fit)
samples <- extract(ppfit_bayes$stanfit)
```

Next steps: incorporate Brett's suggestions. Treat visit as a random effect (there are two levels of grouping), but perhaps have some other metric of visit (Julian date) as a fixed effect allowing it to interact with other variables to affect parasite count. Developmental stages could perhaps be included with an interaction with species, which would allow TATO to have some levels and PSRE to have other levels. e.g. I need to put them in the same column. Instead of using stan's negative binomial (which seems a bit problematic according to googling) I can include individual as a random effect.

- Will autocorrelation of fixed and random effects become an issue?