Homework 4 - Mixed effects models
Due October 10 at 9:00am

Names:__Emily Adler & Riley Book____________________________

Background: American Foulbrood (AFB) is an infectious disease affecting the larval stage of honeybees (*Apis mellifera*) and is the most widespread and destructive of the brood diseases. The causative agent is *Paenibacillus larvae* and the spore forming bacterium infects queen, drone, and worker larvae. Only the spore stage of the bacterium is infectious to honey bee larvae. The spores germinate into the vegetative stage soon after they enter the larval gut and continue to multiply until larval death. The spores are extremely infective and resilient, and one dead larva may contain billions of spores.

Although adult bees are not directly affected by AFB, some of the tasks carried out by workers might have an impact on the transmission of AFB spores within the colony and on the transmission of spores between colonies. When a bee hatches from its cell, its first task is to clean the surrounding cells, and its next task is tending and feeding of larvae. Here, the risk of transmitting AFB spores is particularly great if larvae that succumbed to AFB are cleaned prior to feeding susceptible larvae.

Because AFB is extremely contagious, hard to cure, and lethal at the colony level, it is of importance to detect outbreaks, before they spread and become difficult to control. Reliable detection methods are also important for studies of pathogen transmission within and between colonies. Of the available methods, sampling adult bees has been shown the most effective. Hornitzky and Karlovskis (1989) introduced the method of culturing adult honey bees for AFB, and demonstrated that spores can be detected from colonies without clinical symptoms. Recently, culturing of *P. larvae* from adult honey bee samples has been shown to be a more sensitive tool for AFB screening compared to culturing of honey samples. When samples of adult bees are used, the detection level of *P. larvae* is closely linked to the distribution of spores among the bees.

For this reason, we will model the density of *P. larvae* with the potential explanatory variables as number of bees in the hive, presence or absence of AFB, and hive identity.
Instructions: Turn in the assignment via Canvas as a link to a GitHub repository containing a single PDF file (this worksheet with your answers) and a commented .R file(s) and your code. The repository should contain (at least) the following folders: code, data, and figures (or outputs) with the appropriate files in each folder.

Q1. Does variance of spore density appear homogenous among hives? Why or why not?

No, variance of spore density does not appear homogenous among hives. Measures of spore density within hives 12, 13, 14, and, to a lesser degree, 6 appear much more variable than measures of spore density within the other hives.

Q2. Try some transformations of the response variable to homogenize the variances (or at least improve it). Which transformation of spore density seems reasonable? Why?

Log10 of spore density +1 seems reasonable. It reduces the range of spore densities among hives and variance of spore densities within hives.

Q3. Develop a simple linear model for transformed spore density. Include infection (fInfection01), number of bees (sBeesN) and their interaction as explanatory variables. Check for a hive effect by plotting standardized residuals (see the residuals(yourmodel, type='pearson') function) against hive ID (fhive). Show your code and your plots. Do residuals look homogenous among hives?

For mod.1 (without Hive included as a fixed effect or random effect),
the residuals do not look homogenous among hives.

Q4. What are the advantages of including hive as a random effect, rather than as a fixed effect?

Using hive as a random effect is important since there are multiple observations from the same hive, so those observations are correlated. There are 24 hives so using hive as a fixed effect would be too many degrees of freedom. Using hive as a random effect allows for correlation between the hive observations so we only need to estimate one variance.

Apply the Zuur protocol (10-step version outlined here, as used with the barn owl nesting data in Zuur Ch. 5):

Step 1: Fit and check a "beyond optimal" linear regression (already done above)
Step 2: Fit a generalized least squares version of the "beyond optimal" model (no need: we will use the linear regression model).

Q5. Step 3. Choose a variance structure or structures (the random effects). What random effects do you want to try?

We will include hive as a random effect.

We will now fit a mixed effects (ME) model. Zuur et al. used the nlme package in R, but Douglas Bates now has a newer package that is widely used and that is called lme4. The benefits of lme4 include greater flexibility in the structure of the random effects, the option to use non-Gaussian error structures (for generalized linear mixed effects models, or GLMMs), and more efficient code to fit models. The main difference between nlme's lme() function and the lmer() function in lme4 is in how random effects are specified:

```
        model <- lmer(response ~ explanantoryvars + (1|random),
data=mydata) # a random intercept model
```

model <- lmer(response ~ explanantoryvars + (slope|random), data=mydata) # a random intercept and slope model

One of the frustrations some people run into is that the lme4 package doesn't provide p-values. This stems from disagreements and uncertainty about how best to calculate p-values. However, p-values can be dervied from the lmerTest package.

Q6. Step 4. Fit the "beyond optimal" ME model(s) with lmer() in the lme4 package (transformed spore density is response, fInfection01, sBeesN, and interaction are the explanatory variables). Show your code.

Q7. Step 5. Compare the linear regression and ME model(s) with a likelihood ratio test, including correction for testing on the boundary if needed. Use the anova() command. This will re-fit your lmer model with maximum likelihood, but this is OK (note there are some debates about exactly how to best compare an lm and lmer model). Show your work and the results. Which random effect structure do you choose based on the results?

We would choose the random intercept model since the AIC value is the lowest.

Q8. Step 6. Check the model: plot standardized residuals vs. fitted values and vs. each predictor. (You can get standardized residuals with residuals(yourmodel, type='pearson')). How do they look?

It looks ok, the spread of the residuals is about twice as large on the low end of the fitted values compared to the high end.

Vs the number of bees in the hive, the range of the residuals bounce around zero but the residuals at 4500 and 7000 are much more variable.

Vs Infection they are similar, there is a larger spread and an outlying residual when there is no infection.

Q9. Step 7. Re-fit the full model with ML (set REML=FALSE) and compare against a reduced model without the interaction term, also fit with ML. Use anova() to compare the models. Which model do you choose? Why?

The reduced model, without the interaction term BeesN * Infection. The interaction didn't prove to be significant when we ran the anova.

Q10. Step 8. Iterate #7 to arrive at the final model. Show your work. What is your final set of fixed effects?

We are left with only infection as a fixed effect.

Q11. Step 9. Fit the final model with REML. Check assumptions by plotting a histogram of residuals, plotting Pearson standardized residuals vs. fitted values, and plotting Pearson standardized residuals vs. explanatory variables. Are there issues with the model? If so, how might you address them?

The histogram looks normal.
There is still a wider range of residuals with the lower fitted vales than the higher values. Heteroskedastic.
There are outliers within the no infection hives. The observed values of spores per bee are lower than the model predicts for spores in categorically uninfected hives. To address these, we could try different transformation or consider other model types such as a GLMM that may be better suited to count data.

Q12. Step 10. Interpret the model. The summary() command is useful here. What have you learned about American Foulbrood?

The number of bees per hive does not affect the density of spores in the bees, but if a hive is infected, they will have a higher density of spores per bee.

Q13. Calculate the correlation between observations from the same hive as variance(fhive random effect)/(variance(fhive random effect) + variance(residual)). Given the correlation among observations from the same hive, do you think it's a good use of time to sample each hive multiple times? Why or why not?

The correlation is 1.

No, you know that the values from the same hive are highly correlated so it would be a better use of time to sample more hives.