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

Names: Danny Szydlowski, Tyler Butts, Sean Bertalot

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?
The variance here is statistically homogenous but still looks pretty heterogeneous. When visualizing the variances,there is quite a bit of variation between the hives. This is likely due to differences in which hives are infected and the degree of infection between the hives causing differences in variance of spore density.

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?
The log transformation is typically used to make variances among groups more homogenous and it did a far better job then the other transformations here. Looking at the Levene test results the log transformation increased the p-value so the variances were even more statistically homogenous

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?
The residuals do not seem homogenous among hives; they actually seem to vary a good amount.

Q4. What are the advantages of including hive as a random effect, rather than as a fixed effect?
If we apply hive as a random effect we are able to better control for

differences in model fit among hives. This is also useful to control for non-independence among hives

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 want to try hive as a random effect because each individual hive is only one in a population of hives, and that way we can better control for differences in model residuals across hive.

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.

```
beeLME <- lmer(sporeTrans~Infection* log(BeesN) +(1|Hive), data = dat, REML = TRUE)
```

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?

```
# The full model with the random effect is better than the simple linear model because p << 0.0001
```

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?

```
fitted(beeLM)
fitted(beeLME)


FLME <- fitted(beeLME) # get the residuals for the linear model
RLME <- residuals(beeLME)

plot(RLME~FLME, ylab = "residuals", xlab = "fitted")

plot(RLME~dat$sporeTrans)
plot(RLME~dat$BeesN)
plot(RLME~dat$Hive)
```

The residuals vs. fitted values look pretty good, but the residuals for the spore density seem a bit more variable on the lefthand side of the graph. The residuals vs. the number of bees seem fairly evenly spread, as do those for the hive.

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 model is not improved by dropping an interaction term because the p-value = 0.78

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

```
beeLME <- lmer(sporeTrans~Infection* log(BeesN) +(1|Hive), data = dat, REML = FALSE)
beeLME_REMLFalse_noInter <- lmer(sporeTrans~Infection+ log(BeesN) +(1|Hive), data = dat, REML = FALSE)
beeLME_REMLFalse <- lmer(sporeTrans~Infection* log(BeesN) +(1|Hive), data = dat, REML = FALSE)
beeLME_REMLFalse_onlyInter <- lmer(sporeTrans~Infection: log(BeesN) +(1|Hive), data = dat, REML = FALSE)

# tried one additional model with only the interaction term
AIC(beeLME, beeLME_REMLFalse_noInter, beeLME_REMLFalse, beeLME_REMLFalse_onlyInter)
```

	df	AIC
#beeLME	6	294.0506
#beeLME_REMLFalse_noInter	5	292.1286
#beeLME_REMLFalse	6	294.0506
#beeLME_REMLFalse_onlyInter	5	292.2628

The model without an interaction term between Infection and the number of bees (transformed) is most supported by AIC. Because delta AIC between the top model and the other models is <2, they are not statistically different models and we would need to consider all in analysis. However, if we were to fit these models with REML the results would likely be different. Our final set of fixed effects was infection and the log-transformed number of bees without an interaction term

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 model seems to work pretty well across hives (which is likely because of the random effect) but does not seem to work well across the number of bees. A random effect could be created for categories of bee density, or there could just be a lot of variation in the number of bees per hive

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

The density of spores per bee increases with the number of infected bees and with the number of bees in the hive. This is likely because a greater number of bees the disease can spread more easily. The model also explains approximately 82% of the variation in spore density from the random effect of hives alone, which according to the model is related to bee density.

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?

Correlation = 0.82

Because the correlation is so high (0.82) we don't gain
much new information from sampling the same hive multiple times