Assignment 11 — due May 1, 2019

1. (This problem comes from a past qualifying exam.) Microbial technologies is a growing research area within agricultural biotechnology. There are growing efforts to reduce our reliance on synthetic fertilizers by the use of nitrogen use efficiency (NUE)-enhancing microbiomes. A UW-Madison lab studying *A. thaliana* carried out an experiment to study seed yield as a measure of NUE for two genotypes (varieties) of *A. thaliana* exposed to 3 different compositions of microorganisms grown within shallow pots. The experiment was carried out in 4 different greenhouses across the UW campus. Each greenhouse contained 3 benches. Each bench had 2 pots with a single randomly assigned variety of the *A. thaliana* plant in each pot. The experimenters made sure that both varieties were represented on each bench. The 3 microorganisms compositions were randomly assigned to the 3 benches within each greenhouse, with 1 bench per composition. The researchers computed and recorded a composite measure of seed yield of each plant in each pot at the conclusion of the experiment.

   Let y_{ijk} denote the seed yield measured for the plant with genotype i ($i = 1, 2$) grown with microorganism composition j ($j = 1, 2, 3$) in greenhouse k ($k = 1, 2, 3, 4$). Consider the following model (model I):

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + \gamma_k + \tau_{jk} + \varepsilon_{ijk}, \quad i = 1, 2, \quad j = 1, 2, 3, \quad k = 1, 2, 3, 4,$$

   where $\gamma_k \sim \mathcal{N}(0, \sigma^2_\gamma)$, $\tau_{jk} \sim \mathcal{N}(0, \sigma^2_\tau)$, $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_\varepsilon)$, and σ^2_γ, σ^2_τ, $\sigma^2_\varepsilon > 0$. All of these random terms are mutually independent, and the remaining terms in the model are unknown fixed parameters.

   (a) Under model I, what is the correlation between the seed yields of two plants growing together on the same bench?

   (b) Rewrite model I as: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ with

$$\begin{aligned}\mathbf{y} \;=\; & (y_{111}, y_{211}, y_{121}, y_{221}, y_{131}, y_{231}, y_{112}, y_{212}, y_{122}, y_{222}, y_{132}, y_{232}, \\ & y_{113}, y_{213}, y_{123}, y_{223}, y_{133}, y_{233}, y_{114}, y_{214}, y_{124}, y_{224}, y_{134}, y_{234})\,.\end{aligned}$$

   Explicitly specify $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{Z}$ and $\mathbf{u}$.

   For the remaining parts below, consider the following `R` commands and output, where `y` denotes the data vector $\mathbf{y}$ from part (b), `GH`, `MC` and `GENO` are factors in `R` corresponding to the experimental factors greenhouse, microorganism composition, and genotype, respectively.

```
m = lm(y ~ GH*MC*GENO)
anova(m)
Analysis of Variance Table
Response: y
          Df Sum Sq  Mean Sq
GH         3  113.3     37.8
MC         2  321.8    160.9
GENO       1    2.5      2.5
GH:MC      6  116.4     19.4
GH:GENO    3   11.7      3.9
MC:GENO    2   75.1     37.5
GH:MC:GENO 6   14.5      2.4
```

   Answer the following questions based on the output above, and model I.

   (c) Using parameters of model I, write down explicitly the null hypothesis of no microorganism composition effect.

   (d) Test for microorganism composition main effect.

   (e) Test for genotype main effect.

   (f) Test for microorganism composition by genotype interaction effect.

(g) Given the overall objective of "reducing our reliance on synthetic fertilizers by the use of nitrogen use efficiency (NUE)-enhancing microbiomes", critique this designed experiment.

(h) The researchers are allowed to replace the 3 small benches with one large bench that can hold 6 pots within each greenhouse. Then the 2 genotypes and 3 microorganism compositions are randomly assigned to the pots within each greenhouse by making sure each combination is represented once within each greenhouse. Comment on this design compared to the original design and specify how your tests for parts (d) (e) and (f) would change. Be explicit and carry out the tests whenever you can.

2. It is common in fitting a generalized linear model to observe overdispersion. For example, suppose that $Y_i \sim \mathcal{P}(\mu_i)$ where $\log(\mu_i) = \beta_0 + \beta_1 x_i$. For a Poisson distribution it should be the case that $\text{var}(Y_i) = \mathbb{E}(Y_i)$, but often we find that the variance is larger than the expected value. A GLMM can be used to model this. Specifically, suppose that we change the model such that $Y_i|u_i \sim \mathcal{P}(\mu_i)$ where $\log(\mu_i) = \beta_0 + \beta_1 x_i + u_i$ and where $u_i \sim \mathcal{N}(0, \sigma_U^2)$.

   (a) Show that $\log \mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + \sigma_U^2/2$.

   (b) Determine $\text{var}(Y_i)$ and show that it is strictly greater than $\mathbb{E}(Y_i)$.

3. Here is a (modified) code fragment from Benjamin Bolker, a contributor to `R` and `lme4` in particular. It simulates mixed effects Poisson data with a compound symmetry covariance for the random effects.

```
## function to simulate data:
##     ng groups, nr replicates (or individuals) per group, 1 observation / individual
##     fsd = SD of group random effects, indsd = SD of individual random effects
##     b = fixed effect coefficients, of size 2: intercept and slope
simfun = function(ng=4, nr=10, fsd=1, indsd=0.5, b=c(1,2)) {
  ntot = nr*ng
  b.reff = rnorm(ng,   sd=fsd)
  b.rind = rnorm(ntot, sd=indsd)
  x  = runif(ntot) # predictor
  dd = data.frame(x, f=factor(rep(LETTERS[1:ng], each=nr)), obs=factor(1:ntot))
  dd$eta0 = as.vector(model.matrix(~x,data=dd) %*% b)
  dd$bi   = b.rind
  dd$eta  = with(dd, eta0 + b.reff[f] + b.rind) # log E(Yi) = Xbeta + group RE + indiv RE bi
  dd$mu   = exp(dd$eta)
  dd$y    = with(dd, rpois(ntot, lambda=mu))
  dd
}

## try it
library(lme4)
set.seed(12345)
dd = simfun()
(m0 <- glmer(y ~ x + (1|f), family="poisson", data=dd))
(m1 <- glmer(y ~ x + (1|f) + (1|obs), family="poisson", data=dd))

## function to extract the variance components and fixed effects
cfun = function(d) {
  m = glmer(y ~ x + (1|f) + (1|obs), family="poisson", data=d)
  c(sqrt(unlist(VarCorr(m))), fixef(m))
}

## repeat many times
rr = replicate(100, cfun(simfun()))
```

(a) Simulate a single data set under each of the 3 scenarios below. Plot the data: show the response, the predictor `x`, and use colors and a smooth line for each group (factor `f`). Next, fit this data set using `glm`, a Poisson response, and two separate fixed-effect models: one model with `x` as single predictor (in addition to an intercept of course), and one using both `x` and `f` as predictors but no interaction (as if the factor `f` represents blocks). Do the data exhibit overdispersion? Do this under each of the following 3 simulation scenarios:

i. no random effects at all: set `fsd` and `indsd` to 0;

ii. with group random effects, causing group differences (set `indsd` to 0, keep `fsd` to its default value)

iii. with group random effects and individual random effects.

(b) For this question, use the data set that you simulated for (a)(iii), with both group random effects and individual random effects. Fit this data set using `glmer`, as shown above to build models `m0` and `m1`. Use a likelihood ratio test (with comparison to a chi-square distribution) to test the effect of `x` on the response `y`. How do the results compare when using `m1` versus `m0`?

(c) Use a likelihood ratio test with parametric bootstrap to test the effect of `x`, using model `m1`. How do the results compare to those in (b)?

(d) How (un)biased and variable are the parameter estimates across the 100 simulations?