STAT 506 Midterm
2015 Solutions

1. Red Deer on 24 game farms in Spain were tested and found to have *E. cervi* parasites in their gut (1) or not (0). The binary response is measured on each deer individually. Possible explanatory variables are sex and centered length (a proxy variable for age – older animals have had more time to pick up the parasite). Three models were fit in R.

   (a) Write out the model we are fitting in `deer.fit0` above. Use Greek letters for all parameters. Explain all relevant distributions. (5 pts)

   $$logit(p_i) = \beta_0 + \beta_1 x_i + \beta_2 g_i + \beta_3 g_i x_i + \tau_{j[i]}$$

   where β_0 is the intercept for males on Farm = 1
   where β_1 is the slope on centered length (x) for males
   where β_2 is the intercept adjustment for females ($g = 1$)
   where β_3 is the slope adjustment for females
   where τ_j is the adjustment for Farm j
   The relevant distribution is that parasite presence in deer i is y_i which has a $Binomial(1, p_i)$ (or Bernoulli(p_i)) distribution.

   (b) How does `deer.fit1` differ from the model in (a)? Again use Greek letters for parameters and describe all relevant distributions. (4 pts)

   $$\text{logit}(p_i) = \beta_0 + \beta_1 x_i + \beta_2 g_i + \beta_3 g_i x_i + b_{0,j[i]}$$

   The β's are now "population average effects" and τ_j becomes $b_{0,j}$ and is given a $N(0, \sigma_0^2)$ distribution for the random farm intercept. Variance σ_0^2 describes variation in intercept from farm to farm.

   (c) How does `deer.fit2` differ from the model `deer.lme1`? Again use Greek letters for parameters and describe all relevant distributions. (4 pts)

   We add a random slope for each farm.

   $$logit(p_i) = \beta_0 + \beta_1 x_i + \beta_2 g_i + \beta_3 g_i x_i + b_{0j[i]} + b_{1j[i]} x_i$$

   where

   $$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \Psi) \textit{ and } \Psi = \begin{pmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}$$

   (d) Explain how these estimates fit into your described model. (4 pts)

   $$\widehat{\Psi} = \begin{pmatrix} 1.5456^2 & 0.65(1.5456)(0.0293) \\ 0.65(1.5456)(0.0293) & 0.0293^2 \end{pmatrix}$$

   (e) We compare the last two models with the `anova` command. Explain exactly what this comparison is testing (null and alternative hypotheses) and what you conclude from the results shown. (5 pts)

$H_0: \ \sigma_1^2 = 0$ and $\rho = 0$.
$H_a: \ \sigma_1^2 > 0$ (and ρ doesn't really matter – it could be 0 or not)
With the small p-value, we reject H_0 and conclude that there is fairly strong evidence of a random slope from farm to farm. I didn't ask this, but setting a variance term to 0 is on the boundary of the parameter space. The χ_2^2 distribution used is too conservative. A better approach would be to use a 50/50 mixture of χ_1^2 and χ_2^2.

(f) For the test shown just above, does R compare REML based or ML based estimation methods? Explain when ML is the right method. (4 pts)

REML (which is the right method for comparing models with the same fixed effects and different random effects). We use ML when comparing models with the same random effects and different fixed effects. BTW, `lmer` *will automatically refit to give ML if you try to makeone of these comparisons on REML fit models.*

(g) Based on the summary below, give and interpret the model's estimated response for the average clength female deer ($\overline{x}_F = -9.7$cm) and the average clength male deer ($\overline{x}_M = 8.2$cm). (5 pts)

For the "population average" male of centered length 8.2 cm, our logit scale prediction is $1.088 + 0.043 \times 8.2 = 1.44$. *For the "population average" female of centered length -9.7 cm, our logit scale prediction is* $1.088 + 0.558 + (0.043 + 0.031) \times -9.7 = 0.928$. *Converting to probability, we get estimated probabilities of parasite: 0.809 for the male, and 0.717 for the female. No "farm" random effects are included in these estimates.*

(h) How would you find logit scale standard errors for the predictions you just computed? What further information is needed? (5 pts)

We need the estimated variance-covariance matrix for $\widehat{\beta}$, as from the `vcov` *function in R. Then we could compute $SE(\mathbf{x}\widehat{\beta}) = \sqrt{\mathbf{x}\widehat{\mathbf{V}}(\widehat{\beta})\mathbf{x}^T}$ where $\mathbf{x}$ is the row of covariates for the average male and average female deer in turn.*

(i) In model `deer.fit0` we had `+ Farm` and in model `deer.fit1` we used `+ (1|Farm)`. What difference does that make in terms of our inference about Farms? (5 pts)

In the fixed effects model we can compare one farm to another, but basically the 24 farms are considered the only ones of interest. In the model with random effects we get an estimate of σ_0^2 which describes the farm to farm variance across a population of game farms in Spain. We can predict individual farm effects, but they are not of primary interest.

(j) If the game farms were chosen at random from a larger population of such farms, and if all deer at each sampled farm were tested for the parasite, what is the scope of inference for this study? Explain your thinking. (5 pts)

We could then extend our inference back to the population of farms from which these were chosen. It would be inference of association, not causation because this was an observational study.

2. For the Red Deer data, consider the coefficient estimates for clength across the 24 farms in a model with no Sex predictor.

   (a) How would you fit the "unpooled" estimate of the clength effect? (2 pts)

   Fit separate glm's for each farm. It would also work to fit a model with clength, Farm, and clength by Farm interaction.

   (b) How would you fit the "complete pooling" estimate? (2 pts)

   Fit one glm model using clength but ignoring farm.

   (c) How would you fit the "partial pooling" estimate? (2 pts)

   Fit a mixed model (glmer) with intercept and clength fixed effects and intercept and clength random effects for each farm.

   (d) Recall that when building models for radon exposure in Minnesota, some county estimates "shrank" much more than others.

   i. Explain what shrinkage means. (3 pts)

   We call it "shrinkage" when group level estimates fit in the unpooled way are pulled toward the pooled estimate to obtain the "partially pooled" mixed effects estimates.

   ii. What attribute of game farm data could make the clength estimate for one farm shrink much more than the estimate for another? (5 pts)

   Sample size. Farms with fewer deer will see their estimates shrink toward the pooled estimate much more than will farms with lots of deer.