Chapter 5

Modeling and inference using GLMMs

5.1 Introduction

In this chapter I continue the prescription of Section 4.4 and present a number of examples and consider the inferential goals.

5.2 Chestnut blight (gene effects)

Recall the model we developed in the first chapter (1.1) for the chestnut blight example, now modified to include random effects:

$$
\begin{aligned}
&Y_i = 1 \qquad \text{if the virus is transmitted and 0 otherwise,} \\
&Y_i|\mathbf{u} \sim \text{indep. Bernoulli}(p_i), \\
&p_i = \Phi\left(\mu + \sum_s \beta_s \text{MCH}_{is} + \sum_s \gamma_s \text{ASY}_{is} + \mathbf{z}'_{d,i}\mathbf{u}_1 + \mathbf{z}'_{r,i}\mathbf{u}_2\right),
\end{aligned} \tag{5.1}
$$

where $\mathbf{z}'_{d,i}$ and $\mathbf{z}'_{r,i}$ are the ith rows of the model matrices for the donor and recipient random effects, respectively, and we assume

$$
\begin{aligned}
&\mathbf{u}_1 \sim \mathcal{N}(0, \mathbf{I}\sigma_d^2) \text{ independent of} \\
&\mathbf{u}_2 \sim \mathcal{N}(0, \mathbf{I}\sigma_r^2).
\end{aligned} \tag{5.2}
$$

One inferential goal might be to test if gene 4 had an effect. To do so we could fit the model described by (5.1) and (5.2) and evaluate the log likelihood. We would next fit the same model, but with β_4 and γ_4 set equal to zero and compare the value of the log likelihood. A large sample likelihood ratio test could be used to test $H_0 : \beta_4 = \gamma_4 = 0$. The inferential goal in this case is to form a hypothesis test of parameters from the linear predictor.

TABLE 5.1.
The progabide/seizure data set

	Number of seizures per					
Patient	Baseline	Period 1	Period 2	Period 3	Period 4	Trt
1	11	5	3	3	3	0
2	11	3	5	3	3	0
3	6	2	4	0	5	0
4	8	4	4	1	4	0
.	.	.	.	.	.	.
57	13	0	0	0	0	1
58	12	1	4	3	2	1

5.3 Progabide and seizures

A well studied example is the Progabide and seizures data set, reproduced in, for example, Diggle et al. (1994). Epileptics were randomly allocated to a placebo group or an drug (Progabide) group. The number of seizures was recorded for a baseline period of 8 weeks and during consecutive two-week periods for 4 periods after beginning treatment. The main question of interest is whether the drug is effective at reducing the number of seizures. Table 5.1 gives a portion of the data. Our outcome variable would be the number of seizures for individual i at time t for $t = 1, 2, 3, 4, 5$. For a distributional assumption we might entertain the Poisson and we would want to incorporate the following predictors: period, treatment, period $\times$ treatment (all fixed) and random effects for individuals. We will need to accommodate the fact that the baseline period is 8 weeks long, while the observation periods are 2 weeks long:

$$
\begin{aligned}
Y_{it} &= \text{ number of seizures in period } t \text{ for subject } i, \\
Y_{it}|\boldsymbol{\lambda} &\sim \text{indep. Poisson}(\lambda_{it}), \\
\ln(\lambda_{it}) &= \mu + s_i + \beta_1 \text{POST}_{it} + \beta_2 \text{TRT}_i + \beta_3 \text{POST} \times \text{TRT}_{it}, \\
&\quad + \ln(\text{TIM}_{it}), \\
s_i &\sim \text{ i.i.d. } \mathcal{N}(0, \sigma^2_{\text{subj}}),
\end{aligned}
\tag{5.3}
$$

where s_i are the random subject effects, POST_{it} is 1 if the observation is post-baseline and 0 otherwise (none of the post-baseline periods were found to be different), TRT_i is 1 if the subject is in the treatment group and zero otherwise, and TIM_{it} is 8 for the baseline period and 2 for all the other periods.

The role of the $\ln(\text{TIM}_{it})$ term on the right-hand-side of the equation (called an offset) is to be able to accommodate the differing time periods and model the number of seizures per week. Taken over to the left-hand-side of the equation, it is clear we are modeling $\ln(\lambda_{it}/\text{TIM}_{it})$ (i.e., the log of the rate per unit time) as a function of the remaining variables.

We are are mainly interested in β_3 since that measures the differential change over baseline between the control and treatment groups (in case there is a placebo

effect, we would like to show that the drop in the treatment group is larger). So our primary hypothesis is $H_0 : \beta_3 = 0$ versus the alternative $H_A : \beta_3 < 0$.

5.4 Chestnut blight (isolate effects)

I return to the chestnut blight example and the model given by (5.1) and (5.2). If there are no other genes affecting the transmission of the virus, then all isolates with a given set of fixed effects will behave the same. On the other hand, if other genetic factors are at work, isolates with the same values of the pre-identified genes will behave differently. To test for the presence of other genetic factors, we will be interested in testing the null hypothesis

$$H_0 : \sigma_d^2 = \sigma_r^2 = 0, \tag{5.4}$$

which we might attempt using a likelihood ratio test. Namely, we would compare the maximized value of the log likelihood for the full model and the model with both the donor and recipient variances restricted to zero. A p-value would be defined by

$$p\text{-value} = Pr\{W \geq 2\Delta \log L\}, \tag{5.5}$$

where W is the (log) likelihood ratio statistic and $2\Delta \log L$ is the observed value of the log likelihood ratio. Large p-values would support the null hypothesis of no other genetic effects. A difficulty is that the large sample distribution of W is hard to deal with since it is a mixture of independent χ^2s with degrees of freedom "0," 1 and 2.

5.5 Potomac River Fever

Potomac River Fever (equine *monocytic ehrlichiosis*) is a blood-borne rickettsial disease whose transmission mechanism is unknown. Both arthropod (e.g., blackfly) and direct oral transmission have been suspected but not verified. Identification of risk factors of horses in New York State might give clues to the spread of this disease and help with reducing its frequency. In Atwill et al. (1996) 511 farms were studied, each with several social groups of horses, for a total of 2,587 horses. The outcome was whether or not a horse tested positive for the presence of the disease, so the outcome is binary. We might construct a model as follows:

$$\begin{aligned} Y_{ijk} &= 1 \text{ if horse } k \text{ in social group } j \text{ on farm } i \\ &\qquad \text{tests positive and 0 otherwise,} \\ Y_{ijk}|\mathbf{s},\mathbf{f} &\sim \text{indep. Bernoulli}(p_{ijk}), \\ \text{logit}(p_{ijk}) &= \mathbf{x}'_{ijk}\boldsymbol{\beta} + s_{j(i)} + f_i, \\ \mathbf{s} &\sim \mathcal{N}(0, \mathbf{I}\sigma^2_{\text{group}}) \text{ independent of} \\ \mathbf{f} &\sim \mathcal{N}(0, \mathbf{I}\sigma^2_{\text{farm}}), \end{aligned} \tag{5.6}$$

where $\mathbf{x}'_{ijk}\boldsymbol{\beta}$ is the fixed effects portion of the model, $s_{j(i)}$ are the social group effects, and f_i are the farm effects. To assess the primary question of mode of

transmission, the fixed effects included predictors such as frequency with which the stall was cleaned, frequency with which fly spray was applied, distance to water and a number of others. In the analysis, none of the predictors related to possible modes of transmission was statistically significant.

Next I consider the random effects. The estimated variances of the random effects were $\hat{\sigma}^2_{\text{group}} = 0$ and $\hat{\sigma}^2_{\text{farm}} = 1.26$. So the difference in loglikelihood for testing $\hat{\sigma}^2_{\text{group}} = 0$ is zero and hence not statistically significant. After dropping it from the model, a test of $\hat{\sigma}^2_{\text{farm}} = 0$, *was* statistically significant at the 0.05 level. The lack of significance of the social group variance component suggests that the transmission mechanism is not directly from horse to horse, but instead is related to farm location or management practices at the farm level. Here, as in the previous example, inference focused on testing for the presence of a non-degenerate random effects distribution.

5.6 Neotropical migrants

In the Breeding Bird Survey, counts of number of birds "sighted" has been made each June at thousands of locations across the U.S. and Canada. Many of the locations have been surveyed since the mid 1960s. In James et al. (1996) responses were summarized by estimating whether the trend in population size was positive within each of 37 physiographic strata. So the outcome for this analysis was whether or not species i in stratum j was estimated to have increased. A model for this situation might be:

$$
\begin{aligned}
Y_{ij} &= 1 \text{ if the estimated population trend is upward} \\
&\qquad \text{for species } i \text{ in stratum } j \text{ and 0 otherwise,} \\
Y_{ij}|\mathbf{s} &\sim \text{indep. Bernoulli}(p_{ijk}), \\
\text{logit}(p_{ij}) &= \mu + s_i + sp_j, \\
\mathbf{s} &\sim \mathcal{N}(0, \mathbf{I}\sigma^2_{\text{strata}}),
\end{aligned}
\tag{5.7}
$$

with the s_i being the stratum effects and the sp_j being the species effects. The stratum effects serve to build in a correlation for all data collected on a species in a stratum; a good idea since the data are collected all on the same day and by the same observers.

The analysis concerned a subset of 26 of the species, called neo-tropical migrants. These birds overwinter in Central and South America and come back to the United States to breed in the summer. The primary question was whether destruction of overwintering habitat was causing neo-tropical migrant bird populations to decline on a continent-wide basis.

Since not all the species occur in each of the 37 physiographic strata (in fact about 2/3 of the data are "missing") fitting the two-way model of (5.7) can be thought of as a "smoothing" method. That is, an attempt to understand the overall increase or decline in species populations, irrespective of the locations at which they are found. Interestingly, no overall declines were found, with about half the species showing increases and about half showing decreases. So the evidence was *not* supportive of continent-wide declines. There were some interesting conclusions related to the random effects however.

In testing the hypothesis of no stratum effects ($H_0 : \sigma^2_{\text{strata}} = 0$) the likelihood ratio test was equal to 11.88, which, when compared to a 50:50 mixture of a χ^2_0 and χ^2_1 is highly statistically significant. More interesting was looking at the empirical best predicted values of the stratum effects, namely $\hat{\mathrm{E}}[s_j|\mathbf{Y}]$, where the hat signifies that the ML estimates of the mean, the species effect and the stratum variance have been inserted. When the strata effects were assessed with respect to their geographic location the most negative ones (i.e., locations with the largest probabilities of decline) were all located in high-altitude locations in the eastern United States. This was highly suggestive of pollution effects taking place in the ecologically sensitive, high altitude locations.

5.7 Photosynthesis in corn relatives

Consider an experiment in which two species of corn relatives (an annual and perennial) were compared with respect to photosynthetic physiology. Seeds from two populations of each species were collected and grown in the greenhouse. The experimental design was a randomized complete block design with four blocks and three seeds from each population in each block (for a total of 12 seeds per block). After 24 days, photosynthesis was recorded at nine different light levels from full sunlight to darkness on one individual from each population in each block (N=16). Measurements on the same 16 plants were repeated after 48 days. From these data, photosynthesis versus irradiance (PAR) response curves reflecting the change in photosynthetic rate with light level were derived.

The traits of interest are the maximum photosynthetic rate, dark respiration, the light compensation point, and the quantum yield. The maximum photosynthetic rate measures the maximum amount of carbon dioxide the plants are able to assimilate in full sunlight, the dark respiration indicates how much carbon dioxide they respire in the dark, the light compensation point is the light level at which photosynthesis overcomes respiration and carbon assimilation becomes positive, and quantum yield is the efficiency of carbon assimilation at low light levels, or the slope of the light response curve as it crosses the light compensation point.

The main question of interest is to compare the two species with respect to their photosynthetic traits. To do so, a equation was hypothesized for the relationship between photosynthetic rate and light of the following form:

$$\mathrm{E}[\mathrm{PHOTO}] = \alpha + \beta(1 - e^{-\gamma \mathrm{LIGHT}}). \tag{5.8}$$

Clearly, we have not yet incorporated the effects of species, block, seed or time of measurement. Presumably, species and block should be random factors, making a mixed model.

This example was introduced to make two points. First, this is *not* a generalized linear mixed model. There is no (link) function of E[PHOTO] in (5.8) that will make it a linear model in the parameters α, β and γ. Second, it is not at all clear how the effects of species, block, seed or time of measurement should enter the model. At the most complicated extreme, we might have to build a separate model (of unknown form) for each of the parameters in (5.8). In a generalized linear mixed model, the modeling is much simpler since all the effects are assumed to enter through the

linear predictor. This is a significant advantage in cases where the more restrictive form of a generalized linear mixed model is adequate to describe the process being modeled.