Chapter 1

Introduction

I begin with an extended example to motivate both the basic ideas behind generalized linear mixed models (GLMMs) and the variety of inferences possible within the context of such models. The example concerns chestnut trees and leaf blight.

1.1 Example: Chestnut leaf blight

The American chestnut tree was a predominant hardwood in the forests of the eastern United States, reaching 80–100 feet in height at maturity and providing timber and low-fat, high-protein nutrition for animals and humans in the form of chestnuts. In the early 1900s an imported fungal pathogen, which causes chestnut leaf blight, was introduced into the United States. The pathogen spread from infected trees in the New York City area and, by 1950, had killed over 3 billion trees and virtually eliminated the chestnut tree in the United States (Woods and Shanks, 1959). Economic losses in both timber and nut production have been estimated in the hundreds of billions of dollars. As well, there are the ecological impacts of eliminating a dominant species.

Attempts to restore this tree to the U.S. forests include development of blight resistant varieties of chestnuts and weakening of the fungus by infecting it with a virus which reduces the fungus' virulence (hypovirulence). I will describe the latter in more detail. The basic idea is to release isolates (genetically identical individuals) of chestnut blight fungus that are hypovirulent and let the viruses infect the natural populations of the fungus, thereby allowing chestnuts trees to survive.

Viruses spread between fungal individuals when they come in contact and fuse together. A major obstacle in spreading this virus and thus controlling the disease is that different isolates of the fungus cannot necessarily transfer the virus to one another.

Michael Milgroom (Cornell Plant Pathology) and his colleague, Paolo Cortesi (University of Milan) have described (Cortesi and Milgroom, 1998) the effect of genes that govern whether or not transmission of this virus is possible between isolates of the fungus. Mismatches on one or more of these so-called "incompatibility genes" decrease the probability of transmission.

To estimate the effects of these genes, they have paired numerous isolates which

differ on the first gene only, the second gene only, the first and the second gene, etc. For each combination of isolates they have averaged about 30 attempts and record a binary response of whether or not the attempt succeeded in transmitting the virus.

Questions of interest include whether pre-identified genes actually do have an influence on transmission of the virus (and if so, to what degree), whether there are other, as yet unidentified, genes which might affect transmission, and whether transmission is symmetric. By symmetry of transmission we mean the following: suppose the infected fungus is type b at the locus for the first gene and the non-infected isolate (which we are trying to infect) is type B. The two isolates are the same at the other five loci. Is the probability of transmission the same as when using a type B to try to infect a type b?

a. A model

Let us begin by defining basic notation and a starting point for a probit model. Let Y_i be a binary indicator which is 1 if a transmission attempt is successful and 0 otherwise. We make the following assumptions:

$$
\begin{aligned}
& Y_i \sim \text{indep. Bernoulli}(p_i), \\
& p_i = \Phi\left(\mu + \sum_s \beta_s \text{MCH}_{is} + \sum_s \gamma_s \text{ASY}_{is}\right),
\end{aligned}
\tag{1.1}
$$

where $\Phi(\cdot)$ is the standard normal c.d.f., $\text{MCH}_{is} = 1$ if there is a mismatch at locus s for pairing i and 0 otherwise, and $\text{ASY}_{is} = \frac{1}{2}$ if there is a mismatch at locus s in pairing i with a b donor, $-\frac{1}{2}$ if there is a mismatch at locus s pairing i with a B donor and 0 if there is no mismatch.

Under this coding of the covariates, β_s is the effect of a mismatch at gene s (averaged over the two types of donors and on the scale of $\Phi^{-1}[p]$) and γ_s is the difference in transmission probabilities between the donor types when there is a mismatch (b versus B).

i. Basic elements of the model

Several observations (some obvious) about the model are worth pointing out at this juncture since they foreshadow more detailed developments later.

- The model is nonlinear in the parameters.
- However, $\Phi^{-1}(p_i)$ is a linear model in the parameters. This is often called the *linear predictor* portion of the model.
- The inverse standard normal c.d.f. in the model serves as a *link* between the response and the predictors.
- Some form of inverse c.d.f. is natural to use as the "link function" since it expands the $(0, 1)$ range of the probabilities, p_i, to the whole real line.

- The model, which can be thought of as a model for $\Phi^{-1}(p_i)$ is quite different from transforming the actual data, Y_i, with $\Phi^{-1}(Y_i)$. Y_i takes on only the values 0 and 1, for which $\Phi^{-1}(Y_i)$ is $-\infty$ or ∞.

ii. Using the model

If were are interested in whether there is asymmetry of transmission, that is, if the probability of transmission depends on the donor type when there is a mismatch, we can formulate this hypothesis as

$$\mathrm{H}_0 : \gamma_s = 0 \qquad \forall s.$$

Since model (1.1) is simply a probit regression model, a standard method of testing this hypothesis is with a likelihood ratio test. The log likelihood for the model is straightforward to write down and numerically maximize. The model including the γ_s has a maximized log likelihood of -955.303 with 13 parameters (μ, the six β_s and the six γ_s), whereas the maximized log likelihood for the model restricting the γ_s to all be zero is -1116.639 with 7 parameters. Twice the difference of these maximized log likelihoods forms the usual test statistic, which, applying the usual large sample theory, is distributed approximately as a chi-square with 6 degrees of freedom.

This gives the likelihood ratio test as

$$\begin{aligned} -2\log l &= -2(-1116.639 + 955.03) = 323.218, \\ p\text{-value} &= \mathrm{P}\{\chi^2_6 > 323.218\} \approx 0, \end{aligned} \tag{1.2}$$

so we can soundly reject the null hypothesis and conclude there is asymmetric transmission.

b. Threshold model

A common model in genetics for describing the presence or absence of a trait is the threshold model (Falconer, 1965). This arises from assuming that a large number of genes each have a small and additive effect and when the cumulative effect exceeds a threshold the trait is present in an individual. As before, let $Y = 1$ if the trait is present, and 0 otherwise. Let $\mathbf{x}'\boldsymbol{\beta}$ represent either genetic or nongenetic fixed effects for an underlying latent variable and let ε be the genetic effect not captured in $\mathbf{x}'\boldsymbol{\beta}$. Since we are modeling an unobservable, latent trait, there is an overparameterization that can be resolved (Manski, 1988) by choosing the threshold to be zero and the error term to have variance 1. The model thus has Y equal to 1 when $\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0$.

Appealing to the central limit theorem gives an approximately normally distributed ε and the probit model:

$$\begin{aligned} \mathrm{P}\{Y = 1\} &= \mathrm{P}\{\mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0\} \\ &= \mathrm{P}\{-\varepsilon < \mathbf{x}'\boldsymbol{\beta}\} \\ &= \Phi(\mathbf{x}'\boldsymbol{\beta}). \end{aligned} \tag{1.3}$$

This genesis of the probit model gives a bit of justification for its use in this example over the more commonly used logit model.

c. Correlations in the chestnut blight example

Different isolates of the fungus were used in the experiments that were the same on the six incompatibility genes but differed with regard to other genes. This raises a serious complication in the model since all transmission attempts using the same isolate will share those "other" genes, which may affect the transmission probabilities. If so, all of those transmission attempts will be correlated, probably invalidating the calculations leading to (1.2). To begin, we might model the effects associated with each isolate as being selected from a normal distribution.

First we need to enlarge the notation somewhat to more clearly describe the possible correlation. Let $Y_{ijk} = i$th observation from an attempted infection from the jth isolate (the donor) to the kth isolate (the recipient). Further let $\mathbf{x}_{ijk}$ be the column vector of covariates (like MCH and ASY) for Y_{ijk}. A reasonable model using the threshold formulation might then be

$$\mathrm{P}\{Y_{ijk} = 1|\mathbf{u}\} = P\{\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1j} + u_{2k} + \varepsilon_{ijk} > 0\}, \tag{1.4}$$

where the u_{1j} represent the (random) effects of the donor isolate and the u_{2k} represent the (random) effects of the recipient isolate (which could well be different than the donor effects) and the vertical bar indicates that the probability is conditional on the value of the random effect vector $\mathbf{u}$.

This yields

$$\mathrm{P}\{Y_{ijk} = 1|\mathbf{u}\} = \Phi\left(\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1j} + u_{2k}\right) \tag{1.5}$$

Several remarks are in order about this model:

- The model is conditionally specified by assuming that, if the genetics were known, we would incorporate them like we did the fixed effects.
- In fact, since they *are* genetic effects, just like MCH and ASY (which went into the fixed effects portion of the model), it is quite reasonable to have them enter the model in exactly the same fashion as do those effects. Said another way, if we believe that MCH and ASY should enter the linear predictor then so should the random effects.
- Because of the conditional specification, almost all calculations are most naturally performed by starting from the conditional distribution and then calculating any needed marginal quantities of interest.

1.2 Consequences of introducing random factors

What are the consequences of introducing the random effects $\mathbf{u}$ into (1.5)? We consider now the marginal mean and the correlation structure.

a. On the mean

The marginal mean is easiest to calculate using the equivalence of the probit model to the threshold model in (1.4). Using the threshold representation and iterated

expectation we have

$$
\begin{aligned}
\mathrm{E}[Y_{ijk}] &= \mathrm{E}\left[\mathrm{E}[Y_{ijk}|\mathbf{u}]\right] \\
&= \mathrm{E}\left[\mathrm{P}\{Y_{ijk}=1|\mathbf{u}\}\right] \\
&= \mathrm{E}\left[\mathrm{P}\{\mathbf{x}'_{ijk}\boldsymbol{\beta}+u_{1j}+u_{2k}+\varepsilon_{ijk}>0|\mathbf{u}\}\right] \\
&= \mathrm{P}\{\mathbf{x}'_{ijk}\boldsymbol{\beta}+u_{1j}+u_{2k}+\varepsilon_{ijk}>0\},
\end{aligned} \tag{1.6}
$$

this last equality holding because the expected value of the conditional probability is the unconditional probability. Rearranging (1.6) we obtain

$$
\begin{aligned}
\mathrm{E}[Y_{ijk}] &= \mathrm{P}\{-(u_{1j}+u_{2k}+\varepsilon_{ijk}) < \mathbf{x}'_{ijk}\boldsymbol{\beta}\} \\
&= \mathrm{P}\{W < \mathbf{x}'_{ijk}\boldsymbol{\beta}\},
\end{aligned} \tag{1.7}
$$

where W is defined as $-(u_{1j}+u_{2k}+\varepsilon_{ijk})$. Recall that we have taken $\varepsilon_{ijk} \sim \mathcal{N}(0,1)$ and, if we make the assumptions that the u_{1j} and u_{2k} are independently and normally distributed with means zero and variances of σ_1^2 and σ_2^2, (respectively) then we have $W \sim \mathcal{N}(0, 1+\sigma_1^2+\sigma_2^2)$. Finally this gives us the marginal mean as

$$
\begin{aligned}
\mathrm{E}[Y_{ijk}] &= \mathrm{P}\{W < \mathbf{x}'_{ijk}\boldsymbol{\beta}\} \\
&= \mathrm{P}\left\{W/\sqrt{1+\sigma_1^2+\sigma_2^2} < \mathbf{x}'_{ijk}\boldsymbol{\beta}/\sqrt{1+\sigma_1^2+\sigma_2^2}\right\} \\
&= \mathrm{P}\left\{Z < \mathbf{x}'_{ijk}\boldsymbol{\beta}/\sqrt{1+\sigma_1^2+\sigma_2^2}\right\} \\
&= \Phi\left(\mathbf{x}'_{ijk}\boldsymbol{\beta}/\sqrt{1+\sigma_1^2+\sigma_2^2}\right) \\
&\equiv \Phi\left(\mathbf{x}'_{ijk}\boldsymbol{\beta}^*\right),
\end{aligned} \tag{1.8}
$$

where we have defined $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\sqrt{1+\sigma_1^2+\sigma_2^2}$.

This result has important ramifications. First, unlike linear mixed models, in which the conditional (on the random effects) and marginal means are the same, here they are different. Hence it is quite important to distinguish in interpretations whether we are discussing the marginal or conditional mean. This lack of equality of the two means is a characteristic of the nonlinear model, and is not due to the non-normality of the distribution for the data. Interestingly, the form of the mean is a probit model either conditionally or marginally. This, it turns out, is a special situation that does not hold in general, even for closely related models such as the logit model. Finally the regression coefficients for the marginal mean are *attenuated* as compared to the conditional mean. More precisely they are smaller by the factor $\sqrt{1+\sigma_1^2+\sigma_2^2}$.

b. On the variance-covariance structure

One of the main reasons for considering models with random effects is that it is a convenient way to specify a correlated data model. I illustrate this by calculating the correlation of two observations sharing both the same donor and same recipient isolate (calculations of correlations of observations sharing only one of the random

TABLE 1.1.
Correlations for a probit model

	ρ				
μ	0	0.3	0.5	0.7	0.9
0	0.00	0.19	0.33	0.49	0.72
1	0.00	0.16	0.31	0.47	0.71
2	0.00	0.10	0.23	0.43	0.70
3	0.00	0.04	0.14	0.36	0.68

effects are similar). First I calculate $\mathrm{E}[Y_{ijk}Y_{ljk}]$. Again this is straightforward using iterated expectations:

$$\begin{aligned}\mathrm{E}[Y_{ijk}Y_{ljk}] &= \mathrm{E}\left[\mathrm{E}[Y_{ijk}Y_{ljk}|\mathbf{u}]\right] \\ &= \mathrm{E}\left[\mathrm{E}[Y_{ijk}|\mathbf{u}]\mathrm{E}[Y_{ljk}|\mathbf{u}]\right] \\ &= \mathrm{E}\left[\mathrm{P}\{Y_{ijk}=1|\mathbf{u}\}\mathrm{P}\{Y_{ljk}=1|\mathbf{u}\}\right] \\ &= \mathrm{E}\left[\Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta}+u_{1i}+u_{2j})\Phi(\mathbf{x}'_{ljk}\boldsymbol{\beta}+u_{1i}+u_{2j})\right].\end{aligned} \tag{1.9}$$

If we now let $W = u_{1i} + u_{2j}$ so that $W \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$, we can complete the calculation as

$$\begin{aligned}\mathrm{E}[Y_{ijk}Y_{ljk}] &= \mathrm{E}\left[\Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta}+W)\Phi(\mathbf{x}'_{ljk}\boldsymbol{\beta}+W)\right] \\ &= \int_{-\infty}^{\infty}\Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta}+\tau z)\Phi(\mathbf{x}'_{ljk}\boldsymbol{\beta}+\tau z)\phi(z)dz,\end{aligned} \tag{1.10}$$

where $\tau = \sqrt{\sigma_1^2+\sigma_2^2}$ and $\phi(\cdot)$ is the standard normal p.d.f.

Thus, sharing of the random effects introduces a positive association among the observations. This is perhaps easiest to think about using the threshold representation (1.3). The threshold model is a linear mixed model on the probit scale and we can calculate the usual correlation (Searle et al., 1992) of two observations sharing both random effects as $\rho = (\sigma_1^2+\sigma_2^2)/(1+\sigma_1^2+\sigma_2^2)$, the 1 appearing because the error term in the threshold model has unit variance. How does this translate into an association for the binary variables? Although correlation is not the most natural measure of association for binary variables, it is an easily understood measure. Using (1.8), (1.10) and the usual formula for a correlation, Table 1.1 gives the correlations for various values of ρ and $\mu = \mathbf{x}'_{ijk}\boldsymbol{\beta} = \mathbf{x}'_{ljk}\boldsymbol{\beta}$ (assumed to be equal for simplicity).

There is a fairly close correspondence between the value of ρ and the induced correlation and μ has a modest influence on the correlation as well.

c. On the likelihood

Calculation of the likelihood also proceeds by first writing down the conditional density of $\mathbf{Y}$ given $\mathbf{u}$ and then calculating the marginal density. The elements of $\mathbf{Y}$

are assumed to be independent given $\mathbf{u}$ so the conditional density is given by

$$f_Y(\mathbf{y}|\mathbf{u}) = \prod_{i,j,k} \Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j})^{y_{ijk}} \times [1 - \Phi(\mathbf{x}'_{ijk}\boldsymbol{\beta} + u_{1i} + u_{2j}]^{1-y_{ijk}}, \tag{1.11}$$

so that the likelihood is given by

$$f_Y(\mathbf{y}) = \int f_Y(\mathbf{y}|\mathbf{u}) f_u(\mathbf{u}) d\mathbf{u}. \tag{1.12}$$

Unfortunately, the dimension of $\mathbf{u}$ for the chestnut blight example is over 200. Further, the data do not separate into independent clusters, so the dimension of the integral required to calculate the likelihood cannot easily be reduced. This is one of the central problems with likelihood analysis of generalized linear mixed models.

What remedies might there be to deal with this computational intractibility? Possibilities include:

- Find some way to calculate or approximate the 200+ dimensional integral,
- Consider models which do not give such difficult likelihoods, or
- Change the estimation technique to something that is not so difficult to compute.

All of these approaches have been tried and we revisit them in later chapters, especially Chapters 6, 7 and 8.

1.3 Testing for other genetic effects

For now we will bypass the computational details and consider another question of interest for this problem. Namely, whether genes other than the six identified are contributing to the transmission probability.

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 genes are affecting transmission, then all observations associated with a particular isolate (as a recipient or as a donor, depending on what is affected) will be correlated. This will show up as a nonzero random effect for donors and/or recipients. The null hypothesis of no other genetic effects can thus be stated as

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = 0.$$

This null hypothesis is interesting for two reasons. First, the hypothesis is naturally phrased in terms of the variances, not the means, and second, the genetic theory requires both to be zero simultaneously to prove the point.

Suppose we reject H_0. How could we go about finding the genes that control incompatibility? We might look at the isolates that have the most extreme values of u_{1i} or u_{2j} to see if we could identify the newly found genes affecting transmission. To do so we would want to calculate predicted values of the u_{1i} or u_{2j}.

One way to do this is to would be to form the "best" predictor:

$$\text{best predictor of } u_{1j} \equiv \text{BP}(u_{1j}) = \text{E}[u_{1j}|\mathbf{Y}]. \tag{1.13}$$

This uses the standard result that the minimum mean square error predictor based on $\mathbf{Y}$ is the conditional expected value of u_{1j} given $\mathbf{Y}$ (Searle et al., 1992).

There are (at least) two problems with the use of (1.13) in practice. First, it depends on unknown parameters and so estimates would need to be inserted. Second, the expectation is with respect to the conditional density of $\mathbf{u}$ given $\mathbf{Y}$, which is given by

$$f_{u|\mathbf{Y}} = f_{\mathbf{Y},\mathbf{u}}/f_{\mathbf{Y}}. \tag{1.14}$$

That is, it depends on the hard-to-calculate likelihood, $f_{\mathbf{Y}}$, and raises the same sort of difficulties in calculation as we have seen for maximum likelihood.

1.4 Summary

In this chapter I have introduced most of the basic ideas of the remainder of the monograph: the use of generalized linear models, the incorporation of correlation via random effects, the richness of inferential goals accommodated by these models, and the computational difficulties of likelihood inference.

The probit model is a member of the class of generalized linear models. These models are nonlinear, but of a restricted form. Namely, the model for the link function applied to the *mean* of the data is a linear model in the parameters. For the chestnut blight example it is natural to incorporate both the major genetic effects of the six genes as well as cumulative effects of remaining genes in the linear predictor.

This incorporation is an easy and natural way to model or accommodate correlation in the context of a nonlinear model for non-normal data. It generates a rich class of correlated data models, which otherwise is fairly difficult to specify—there just are not readily available multivariate distributions for non-normally distributed data.

Inferences for this model can be of the usual variety, that is, modeling the effect of predictors on the mean, in which case the random effects and correlation are "nuisance" features of the model. For this example, however, both estimation and testing of the variances of the random effects, as well as prediction of the realized values of the random effects are of interest, as described in Section 1.3.

In the remainder of the monograph I explore these issues in more detail. Chapters 2 and 3 briefly review linear mixed and generalized linear models and Chapter 4 defines and introduces GLMMs. Chapter 5 illustrates the breadth of inferential goals possible with GLMMs. Finally, Chapters 6 through 9 cover the difficult aspects of fitting these models to data; this is where much of the current research interest lies.