Chapter 6

Early research

6.1 Introduction

Early approaches to the accommodation of correlation in non-normal data tended to focus on techniques that were easier to compute, which is not surprising given the computational hurdles identified in previous chapters. Their main drawback was that they tended to be ad hoc and could not easily accommodate more complicated situations.

6.2 Beta-binomial model

One of the earliest models for binary data was the beta-binomial model, which hypothesizes a mixture distribution directly on the probability scale. Since the random "factor" does not enter through a linear predictor, strictly speaking it is not a generalized linear mixed model. As motivation, consider a hypothetical teratology experiment in which lead is administered in the drinking water of pregnant rats, perhaps at different doses and including a control group with no lead. The response we record on all members of a litter is the absence of a birth defect in animal k from mom (and litter) j in treatment group i. Since the response is binary, the outcome must have a marginal Bernoulli distribution. We might hypothesize the following model:

$$
\begin{aligned}
Y_{ijk} &= 1 \text{ if animal } k \text{ from mom } j \text{ in treatment } i \\
&\qquad \text{has a birth defect and 0 otherwise,} \\
Y_{ijk}|p_{ij} &\sim \text{indep. Bernoulli}(p_{ij}), \\
p_{ij} &\sim \text{indep. Beta}(\alpha_i, \beta_i).
\end{aligned}
\tag{6.1}
$$

Under this model, Y_{ijk} has a marginal Bernoulli distribution with mean $\mu_i = \alpha_i/(\alpha_i + \beta_i)$. Data that share a value of p_{ij} are modeled as correlated, but observations from different litters are regarded as independent. Temporarily dropping the subscripts i and j, and using $B(\cdot, \cdot)$ to represent the beta function and $Y_\cdot = \sum_k Y_k$,

consider the joint distribution of the n observations within a litter:

$$
\begin{aligned}
f_{\mathbf{Y}} &= \int_0^1 \prod_k p^{Y_k}(1-p)^{(1-Y_k)} p^{\alpha-1}(1-p)^{\beta-1}/B(\alpha,\beta)dp \\
&= \int_0^1 p^{\alpha+Y_\cdot-1}(1-p)^{\beta+n-Y_\cdot-1}/B(\alpha,\beta)dp \\
&= \frac{B(\alpha+Y_\cdot,\beta+n-Y_\cdot)}{B(\alpha,\beta)}.
\end{aligned} \tag{6.2}
$$

Going back to the i,j,k notation and the full sample, the likelihood is the product across the independent litters:

$$
L = \prod_{i,j} \frac{B(\alpha_i+Y_{ij\cdot},\beta_i+n_{ij}-Y_{ij\cdot})}{B(\alpha_i,\beta_i)}. \tag{6.3}
$$

This model is adequate for the simple situation of nested litters within treatment groups but is much less amenable to extensions than, say, a logit model. First, we run into the difficulties identified in Section 5.7: how would we model the α_i and β_i as a function of more complicated predictors? Second, what about more complicated correlation structures?

One suggestion has been to build in the correlation structure using the beta-binomial mixing distribution on the p scale, but to hypothesize a logistic regression function for the fixed factors (Lee and Nelder, 1996). While this is certainly feasible, it starts losing the computational simplicity of the original beta-binomial model and seems less natural than mixed models on the linear predictor scale.

6.3 A Poisson–Gamma model

A similar model has been hypothesized for count data. As a motivating example, consider comparing a treatment program for alcoholics with regard to the number of hospitalizations in the year following enrollment in a treatment program. Hospitals in a large HMO are randomized to use either the new program or stay with the one currently in use. We record Y_{ijk}, the number of hospitalizations for patient k of hospital j in treatment group i. Hospitalization rates within a hospital will almost certainly be correlated due to the implementation of the program at that hospital, the staff at the hospital, and or the intangible factors associated with the hospital's patient population. We might hypothesize the following model:

$$
\begin{aligned}
Y_{ijk}|\mu_{ij} &\sim \text{indep. Poisson}(\mu_{ij}), \\
\mu_{ij} &\sim \text{indep. Gamma}(r_i,\lambda_i).
\end{aligned} \tag{6.4}
$$

Y_{ijk} follows a count distribution model with mean equal to $\mathrm{E}[\mu_{ij}] = r_i/\lambda_i$. However, it does *not* have a marginal Poisson distribution since it is "over-dispersed" compared to a Poisson distribution. That is, its variance is greater than its mean. By an argument similar to the beta-binomial model, the likelihood is given by

$$
L = \prod_{i,j} \left(\frac{\lambda_i}{\lambda_i+r_i}\right)^{r_i} \frac{\Gamma(Y_{ij\cdot}+r_i)}{\Gamma(r_i)\prod_k Y_{ijk}!}. \tag{6.5}
$$

This model, like the beta-binomial model of the previous section is adequate for simple situations, but does not extend easily, for example, to crossed random effects.

6.4 Marginal models

Another early approach was to directly specify the joint marginal distribution of the variables. Since this is often a difficult undertaking for non-normally distributed variates, it is sometimes the case that only certain aspects of the joint distribution are specified. This direct specification bypasses the supposition of random effects or a mixed model and so does not really involve generalized linear mixed models. In some cases, correspondence between such models and random effects models can be drawn. For example, the marginal distribution resulting from the supposition of a random effects model.

a. A model for binary data

In constructing a marginal model we will often be interested in explicitly modeling the marginal means and accommodating associations among the observations. As an illustration, consider modeling a multivariate binary vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$. If joint density will be proportional to a function of the Y_i, crossproducts of Y_iY_j, etc.,

$$f_{\mathbf{Y}} \propto \exp\left\{\sum_i u_i y_i + \sum_{i<j} u_{ij} y_i y_j + \cdots + u_{12\cdots m} y_1 y_2 \cdots y_m\right\}, \tag{6.6}$$

where, for example,

$$u_{12} = \log\left(\mathrm{OR}[y_1, y_2 | y_j = 0; j > 3]\right), \tag{6.7}$$

and OR represents the odds ratio.

This is not, however, a convenient parameterization for describing the marginal means of the Y_i and transformation back and forth from the above parameters and the marginal means and odds ratios (e.g.) can lead to difficulties in restrictions on the parameters. See Liang et al. (1992), Zhao and Prentice (1990) and Ekholm et al. (1995) for more details.

6.5 Conditional inference

An approach to modeling a factor that is very different from treating it as a random effect is to treat the effects as nuisance parameters and use a conditioning argument to remove them from the likelihood. The classic example is that of a matched pairs binary logistic regression.

As motivation consider a study in which we are interested in whether clopidagrel (an anti-platelet drug) in addition to aspirin will reduce the incidence of stroke following a transient ischemic attack (a disorder caused by temporary disruption of the blood supply to the brain) as compared to aspirin alone. Patients are matched

in pairs according to the age, sex, the severity of the attack (how long it lasts and how severe the symptoms were) and are randomized within a pair to either aspirin alone or aspirin plus colpidagrel. The outcome of interest is whether they have a stroke within 90 days following the transient ischemic attack. It is expected that observations within a matched pair will be similar to one another, that is, correlated. Let Y_{ij} be 1 if person $j = 1, 2$ within pair i has a stroke within 90 days and be 0 otherwise. We will assume that $j = 1$ identifies the aspirin alone patient. A model for this situation would be

$$\begin{aligned} Y_{ij}|\alpha_i &\sim \text{indep. Bernoulli}(p_{ij}), \\ \text{logit}(p_{ij}) &= \alpha_i + \beta x_{ij}, \end{aligned} \tag{6.8}$$

where x_{ij} is 0 for $j = 1$ and 1 for $j = 2$ (the treatment indicator), and the α_i are the pair effects, incorporated in order to accommodate the correlation within a pair.

In a generalized linear mixed model we would go on to assume that the pair effects followed a distribution. In the conditional approach we instead treat them as fixed, unknown parameters. We begin by exploring the possibility of estimating the α_i as well as β by maximum likelihood.

a. Matched pairs: maximum likelihood

If there are N pairs, the likelihood for (6.8) is

$$\begin{aligned} L &= \prod_{i,j} \frac{\exp\{\alpha_i y_{ij} + \beta x_{ij} y_{ij}\}}{1 + \exp\{\alpha_i + \beta x_{ij}\}} \\ &= \prod_{i=1}^{N} \frac{\exp\{\alpha_i y_{i\cdot} + \beta y_{i2}\}}{(1 + \exp\{\alpha_i\})(1 + \exp\{\alpha_i + \beta\})} \\ &= \exp\left(\sum_i \alpha_i y_{i\cdot} + \beta y_{\cdot 2}\right)/d, \end{aligned} \tag{6.9}$$

where $d = \prod_i (1 + \exp\{\alpha_i\})(1 + exp\{\alpha_i + \beta\})$, and again I use the dot notation to signify a sum over the missing subscript.

Consider the derivative of the log of (6.9) with respect to α_i when $y_{i1} = y_{i2} = 0$:

$$\begin{aligned} \frac{\partial \log L}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i} \log\left(1 + e^{\alpha_i}\right)^{-1} \left(1 + e^{\alpha_i + \beta}\right)^{-1} \\ &= -\frac{e^{\alpha_i}}{1 + e^{\alpha_i}} - \frac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}} \\ &< 0. \end{aligned} \tag{6.10}$$

Since the derivative is everywhere decreasing as a function of α_i, the maximum likelihood estimate is $\hat{\alpha}_i = -\infty$. Similarly, when $y_{i1} = y_{i2} = 1$, the maximum likelihood estimate is given by $\hat{\alpha}_i = +\infty$. Inserting these values into the likelihood gives

$$L = \prod\nolimits'_{i=1}{}^{N'} \frac{e^{\alpha_i y_{i\cdot} + \beta y_{i2}}}{(1 + e^{\alpha_i})(1 + e^{\alpha_i + \beta})}, \tag{6.11}$$

where N' is the number of "discordant" pairs (i.e., $y_{i1} \neq y_{i2}$) and the prime denotes a product only over the discordant pairs. This gives a log likelihood of

$$\log L = \sum_i{}' \alpha_i + \beta y_{i2} - \log\left(1 + e^{\alpha_i}\right) - \log\left(1 + e^{\alpha_i + \beta}\right), \tag{6.12}$$

where again the prime denotes a summation only over discordant pairs. Differentiating this with respect to α_i gives a solution (for α_i) of $\beta/2$. Finally, plugging those values of α_i in and maximizing with respect to β gives an estimate of

$$\hat{\beta} = 2 \log \frac{N_{01}}{N_{10}}, \tag{6.13}$$

where N_{10} is the number of pairs with $y_{i1} = 0$ and $y_{i2} = 1$ and N_{01} is the number of pairs with $y_{i1} = 1$ and $y_{i2} = 0$. This is exactly twice what we might guess would be a sensible answer, since $\log(N_{01}/N_{10})$ converges in probability to β.

There are thus two unattractive features of maximum likelihood for this problem. First, it fails to give a reasonable estimator of β. This situation, in which the number of parameters grows proportionally with the sample size, is a well-known situation in which maximum likelihood fails (Neyman and Scott, 1948). Second, it estimates extreme values for the α_i.

There are two approaches to resolving these difficulties. The first has already been considered, which is to declare the α_i to be random effects. Reasonable distributional assumptions on the α_i prevent them from being $\pm\infty$ and replaces the growing (with sample size) number of α_i with a fixed number of parameters describing the parameters of the random effects distribution. The second approach is through conditional maximum likelihood, which I now explore.

b. Matched pairs: conditional likelihood

The basic idea behind conditional likelihood is to identify the sufficient statistics associated with the nuisance parameters (in this case the α_i) and work with the conditional distribution given those sufficient statistics. By definition, this conditional likelihood will not involve the nuisance parameters.

From the form of the likelihood in (6.9) it is clear that the sufficient statistic is $(S_1, S_2, \ldots, S_m, T) = (Y_{1\cdot}, Y_{2\cdot}, \ldots, Y_{m\cdot}, Y_{\cdot 2})$. Since the distribution is discrete, to find the distribution of $\mathbf{S}$ we merely have to sum over the appropriate values of $\mathbf{Y}$:

$$\begin{aligned} f_{\mathbf{S},T}(\mathbf{s}, t) &= \sum_{\mathbf{y}: s_i = y_{i\cdot}, t = y_{\cdot 2}} f_{\mathbf{Y}}(\mathbf{y}) \\ &= C(\mathbf{s}, t) \frac{e^{\sum_i \alpha_i s_i + \beta t}}{d}, \end{aligned} \tag{6.14}$$

where $C(\mathbf{s}, t)$ represents the number of combinations of $\mathbf{y}$ that satisfy the constraints and d was defined below (6.9).

From this it is straightforward to get the marginal distribution of $\mathbf{S}$:

$$\begin{aligned} f_{\mathbf{S}}(\mathbf{s}) &= \sum_z f_{\mathbf{S},T}(\mathbf{s}, z) \\ &= \sum_z C(\mathbf{s}, z) \frac{e^{\sum_i \alpha_i s_i + \beta z}}{d}, \end{aligned} \tag{6.15}$$

Hence the conditional distribution of T given $\mathbf{S}$ is also straightforward:

$$\begin{aligned} f_{T|\mathbf{S}}(t|\mathbf{s}) &= f_{\mathbf{S},T}(\mathbf{s},t)/f_{\mathbf{S}}(\mathbf{s}) \\ &= \frac{C(\mathbf{s},t)e^{\beta t}}{\sum_z C(\mathbf{s},z)e^{\beta z}}. \end{aligned} \tag{6.16}$$

As required by theory, none of the α_i remain in this conditional distribution and hence the conditional likelihood can be used to estimate β or form tests or confidence intervals.

For the matched pairs situation, the coefficient $C(\mathbf{s},t)$ is straightforward to evaluate. Conditional on $S_i = 0$ we know that $Y_{i1} = Y_{i2} = 0$ and conditional on $S_i = 2$ we know that $Y_{i1} = Y_{i2} = 1$. It is only the case in which $S_i = 1$ that any randomness remains. The result is a bit easier to state if we define r as the number of successes in the discordant pairs. Of course, basing a conditional test on r is equivalent to basing a test on t. With this it is not hard to show that

$$\begin{aligned} C(\mathbf{s},t) = &\text{ number of ways the successes in the } N_{10} \text{ and } N_{01} \\ &\text{ pairs can be distributed} \\ = &\binom{N_{10}+N_{01}}{r} = \binom{N'}{r}. \end{aligned} \tag{6.17}$$

This can be used, for example, to test $H_0 : \beta = 0$. Under the null hypothesis, the conditional distribution of r given $\mathbf{S}$ is

$$f_{R|\mathbf{S}}(r|\mathbf{s}) = \frac{\binom{N'}{r}}{\sum_z \binom{N'}{z}}, \tag{6.18}$$

use of which leads to McNemar's test.

This is a very effective use of conditional likelihood: it reduces a difficult-to-deal with likelihood with $m+1$ parameters to a simple combinatorial problem.

c. Between pairs: conditional likelihood

Conditional likelihood is not always so effective. Suppose we change the situation slightly so that $m/2$ subjects are allocated to the aspirin group and $m/2$ to the clopidagrel plus aspirin group. Number the subjects so that those with $i \leq m/2$ represent the aspirin group and those with $i > m/2$ are in the combination group. For each subject, we record, over two time periods, whether or not there is a stroke. We could build a model as follows:

$$\begin{aligned} Y_{ij} &= 1 \text{ if person } i \text{ has a stroke in time period } j \\ &\quad \text{and is 0 otherwise,} \\ Y_{ij}|\alpha_i &\sim \text{indep. Bernoulli}(p_{ij}), \\ \text{logit}(p_{ij}) &= \alpha_i + \beta x_{ij}, \end{aligned} \tag{6.19}$$

where x_{ij} is the treatment indicator function (i.e., it is equal to 0 if $i \leq m/2$ and is 1 otherwise). This model looks remarkably similar to (6.8). Let us consider the conditional likelihood approach.

The density of $\mathbf{Y}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \exp\left\{\sum_i \alpha_i y_{i\cdot} + \beta \sum_{ij} x_{ij} y_{ij}\right\} / d, \tag{6.20}$$

where $d = \prod_{i,j}(1 + \exp\{\alpha_i + \beta x_{ij}\})$. Now note that

$$\begin{aligned} \sum_{i,j} x_{ij} y_{ij} &= \sum_{i,j} I_{\{i>m/s\}} y_{ij} \\ &= \sum_i I_{\{i>m/s\}} y_{i\cdot} \\ &= \sum_{i>m/2} y_{i\cdot}\,. \end{aligned} \tag{6.21}$$

Putting this back into the density gives

$$f_{\mathbf{Y}}(\mathbf{y}) = \exp\left\{\sum_i \alpha_i y_{i\cdot} + \beta \sum_{i>m/2} y_{i\cdot}\right\} / d. \tag{6.22}$$

Now the sufficient statistic is $(S_1, S_2, \ldots, S_m) = (Y_{1\cdot}, Y_{2\cdot}, \ldots, Y_{m\cdot})$ with density

$$\begin{aligned} f_{\mathbf{S}}(\mathbf{s}) &= \sum_{\mathbf{y}: s_i = y_{i\cdot}} f_{\mathbf{Y}}(\mathbf{y}) \\ &= C(\mathbf{s}) \frac{e^{\sum_i \alpha_i s_i + \beta \sum_{i>m/2} y_{i\cdot}}}{d}, \end{aligned} \tag{6.23}$$

which gives a conditional distribution of $\mathbf{Y}$ given $\mathbf{S}$ of

$$f_{\mathbf{Y}|\mathbf{S}}(\mathbf{y}|\mathbf{s}) = \frac{1}{C(\mathbf{s})}. \tag{6.24}$$

In words, the conditional distribution of the entire sample given $\mathbf{S}$ contains no information about the parameter of interest, β, and hence is useless for making inferences. In this situation (where the desired inferences are between pairs) the conditioning argument removes all the information of interest.

6.6 Summary

In summary, the conditional approach is one that works very well for a restricted set of situations, namely those in which the data are balanced and (at least the majority of) the information for the parameter(s) of interest comes from comparisons within levels of the nuisance parameters. As such they are very useful practical tools.

However, compared to random effects models and other approaches, they can be arbitrarily inefficient, in the extreme case (as exhibited in the previous section) containing no information about the parameters of interest. Neuhaus and Lesperance (1996) compare the efficiency trade-offs in a binary data setting.

6.7 Further notes

The models of Sections 6.2 and 6.3 take advantage of the conjugacy of the distributions involved, namely the binomial with the beta and the Poisson with the gamma. This could be extended to other distributions in a obvious way and is one of the avenues explored in Lee and Nelder (1996). Another variation on this theme can be found in Conaway (1990). As demonstrated in this chapter, the use of the conditional distribution in the matched pairs case leads to inferences which essentially discard all the concordant pairs. This raises the issue as to whether information can be recovered from the concordant pairs and whether a random effects analysis might be more efficient. There is a nice discussion in Cox and Snell (1989); see also Liang and Zeger (1988). Verbeke et al. (2001) investigate situations in which it may be advantageous to assume that one factor is random while a different factor is treated as a nuisance parameter and a conditional likelihood derived to eliminate it from consideration. In some cases, the conditional and random effects approaches generate the same estimators (Lindsay et al., 1991; Neuhaus et al., 1994). Fay et al. (1998) explore a compromise between a conditional and GEE approach.