13 Advanced topics

13.1 Introduction

In this chapter, we conclude with brief overviews of several advanced topics. Each of these topics could realistically be the subject of an entire course!

13.2 Generalized linear mixed models

The models considered in Chapter 12 were of the **population-averaged** type; that is, the focus was on explicit modeling of the mean $E(\boldsymbol{Y}_i)$ of a data vector. Of course, the elements of $E(\boldsymbol{Y}_i)$, $E(Y_{ij})$, represent the mean response at a particular time t_{ij} and possibly setting of covariates; i.e. the **average** over all possible values of Y_{ij} we might see under those conditions, the average being over all members of the **population**. The models used to represent $E(Y_{ij})$ as a function of t_{ij} and other covariates were of the generalized linear type, so were no longer **linear** functions of the parameter $\boldsymbol{\beta}$ characterizing mean response.

In Section 12.5, we discussed briefly the alternative strategy of **subject-specific** models for nonnormal data. Here, the idea is to model **individual trajectories**, where the "mean" at time t_{ij} over all observations we might see for a **specific individual** is represented again by a generalized linear model, but where the parameters are in turn allowed to depend on **random effects**. A general representation of such a model is as follows; recall that the **conditional expectation** of $\boldsymbol{Y}_i$ **given** a vector of random effects $\boldsymbol{b}_i$ unique to individual i may be thought of as the "mean" response for a particular individual. We have for an element of $\boldsymbol{Y}_i$ that, for a suitable function f,

$$E(Y_{ij} \mid \boldsymbol{b}_i) = f(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_i), \tag{13.1}$$

where the subject-specific parameter $\boldsymbol{\beta}_i$ may be represented as before, e.g. in the most general case,

$$\boldsymbol{\beta}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i. \tag{13.2}$$

Here, then $\boldsymbol{\beta}$ is the parameter that describes the "typical" value of $\boldsymbol{\beta}_i$s across all individuals with covariate matrix $\boldsymbol{A}_i$; e.g. all individual in a particular treatment group. $\boldsymbol{b}_i$ is a **random effect** assumed to come from a distribution with mean $\boldsymbol{0}$, almost always taken to be the **multivariate normal** distribution, so that

$$\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}).$$

It is further assumed that, at the level of the **individual**, the data in $\boldsymbol{Y}_i$ follow one of the distributions such as the binomial, Poisson, or gamma in the scaled exponential family. It is common to assume that observations on a given individual are taken far apart enough in time so that there is no correlation introduced by the way the data are collected (within an individual); in fact, the observations on a particular individual i, Y_{ij}, $j = 1, \ldots, n_i$, are assumed to be **independent** at the level of the individual. The variance of an observation **at the level of the individual** will thus depend on the mean of an observation at the individual level. Thus, we think of the variance associated with observations **within** a particular individual as being **conditional** on that individual's random effects, because the mean is conditional on them. Thus, we think of the variance within an individual as

$$\text{var}(Y_{ij} \,|\, \boldsymbol{b}_i) = \phi V\{f(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_i)\},$$

where ϕ may or may not be known depending on the nature of the data. For example, if the Y_{ij} are **counts**, then appropriate distribution; for example, if the Y_{ij} are **counts**, then it follows that

$$\text{var}(Y_{ij} \,|\, \boldsymbol{b}_i) = f(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_i).$$

The model defined in (13.1) and (13.2) with the stated properties is referred to in the statistical literature as a **generalized linear mixed model**, for obvious reasons. It is an alternative model to the population-averaged models in Chapter 12. Just as in the linear case, it may be more advantageous or natural to think of individual trajectories rather than the average response over the population; this model allows thinking this way.

However, as discussed in Section 12.5, it is **not** the case that this model and a population-averaged model constructed using the **same** function f lead to the **same** model for $E(Y_{ij})$, as was fortuitously true in the case of a **linear** model. Thus, whether one adopts a **population-averaged** or **subject-specific** approach will lead to **different** implied models for the mean response for the population! Technically, this is because, under the population-averaged model, we would take

$$E(Y_{ij}) = f(\boldsymbol{x}_{ij}'\boldsymbol{\beta}),$$

while under the subject-specific approach, we would take

$$E(Y_{ij} \,|\, \boldsymbol{b}_i) = f(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_i),$$

which implies upon averaging over the population that

$$E(Y_{ij}) = E\{f(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_i)\}.$$

Plugging in (13.2) for $\boldsymbol{\beta}_i$, we see that under the subject-specific approach, the implied model for mean over the population is

$$E(Y_{ij}) = E[f\{\boldsymbol{x}_{ij}'(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i)\}].$$

It is a mathematical fact that, because f is not a **linear function** of $\boldsymbol{b}_i$, taking this expectation is an operation that is likely to be impossible to do in closed form. It follows that it is simply not possible that

$$f(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) = E[f\{\boldsymbol{x}_{ij}'(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i)\}];$$

that is, the two types of model for mean response implied by each strategy are almost certainly not the same.

This has caused some debate about which strategy is more appropriate. For linear models, the debate is not as strong, because the mean response model turns out to be the same, the only difference being how one models the covariance. Here, instead, what is implied about the most prominent aspect, the **mean** over the population, is **not** the same. The debate has not been resolved and still rages in the statistical literature. In real applications, the following is typically true:

- For studies in public health, education, and so on, where the main goal of data analysis is to make proclamations about the **population**, the usual strategy has been to use population-averaged models. The rationale is that interest focuses on what happens **on the average** in a population, so why not just model that directly? For example, if a government health agency wishes to understand whether maternal smoking affects child respiratory health for the purposes of making public policy statements, it wants to make statements about what happens "on the average" in the whole population. For the purposes of making general policy, there is no real interest in **individual** children and their respiratory trajectories. Thus, the thinking is – "why complicate matters by assuming a subject-specific model when there is no interest in individual subjects?"

- On the other hand, in the context of a clinical trial, there may be interest in individual patients and understanding how they evolve over time. For example, in the epileptic seizure study in Chapter 12, researchers may think that the process of how epileptic seizures occur over time is something that happens "within" a subject, and they may wish to characterize that for individual subjects. As a result, it is more common to see generalized linear mixed models used in this kind of setting.

INFERENCE: One **major** complication in **implementing** the fitting of generalized linear mixed models is that it is no longer straightforward to write down the implied **likelihood** of a data vector. The actual form of this likelihood is quite complicated and will involve an **integral** with respect to the elements of $\boldsymbol{b}_i$. Rather than write down this mess, we note what the problem is by considering again something that is related to the full likelihood of a data vector – the mean vector. Here, the mean vector is

$$E(Y_{ij}) = E[f\{\boldsymbol{x}_{ij}'(\boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i)\}],$$

which is a calculation that we have already noted is generally not possible to do in **closed form**. This suggests that trying to derive the whole **likelihood** function in closed form would be equally difficult, which it is!

The result is that the function we would like to use as the basis of estimation and testing is not even something we can **write down**! A variety of approaches to dealing with this problem by way of **approximations** that might allow something "**close to**" the true likelihood function to be written down have been proposed. Discussion of these methods is beyond our scope; see the references in Diggle, Heagerty, Liang, and Zeger (2002) for an introduction to the statistical literature. One of these approximate approaches is implemented in a macro provided by SAS, `glimmix`. The procedure `proc nlmixed` fits these models directly. A new procedure, `proc glimmix`, is being developed. It is important that the user fully understand the basis of these approximate approaches before attempting to fit such models – the interpretation and fitting can be very difficult!

13.3 Nonlinear mixed effects models

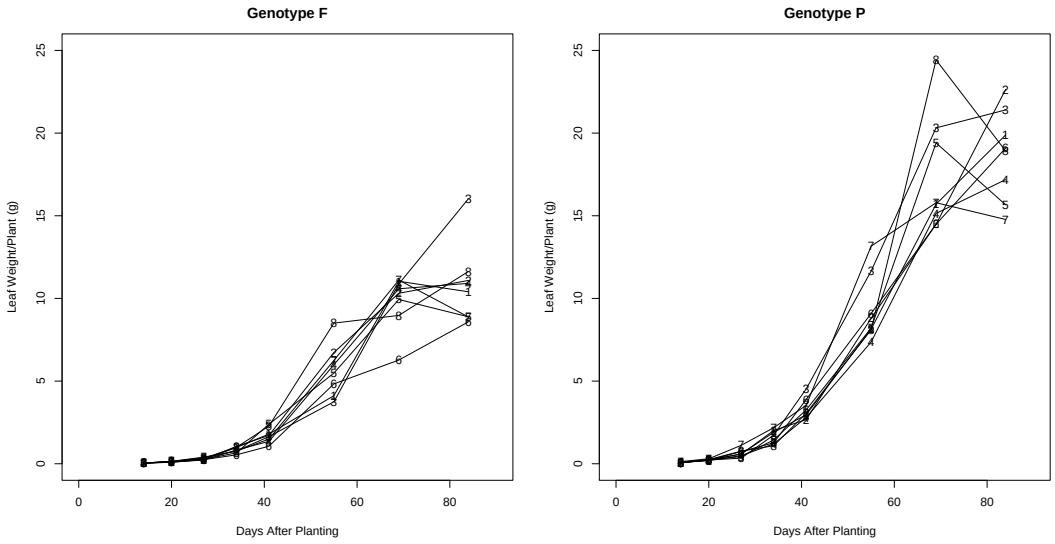
A more complicated version of generalized linear mixed models is possible. In many applications, a suitable model for individual trajectories is dictated by **theoretical concerns**. Recall, for example, the soybean growth data introduced in Chapter 1; the plot is reproduced here as Figure 1. A common model for the process of **growth** is the so-called **logistic growth function**; this function is of a similar form as the logistic regression model discussed previously, but the interpretation is different.

If one assumes that the **rate of change** of the growth value ("size" or "weight", for example) of the organism (here, plants in a soybean plot) relative to the size of the organism at any time declines in a linear fashion with increasing growth, it may be shown that the growth value at any particular time t may be represented by a function of the form

$$f(t, \boldsymbol{\beta}) = \frac{\beta_1}{1 + \beta_2 \exp(-\beta_3 t)}, \tag{13.3}$$

where $\beta_1, \beta_2, \beta_3 > 0$.

Figure 1: *Average leaf weight/plant profiles for 8 plots planted with Forrest and 8 plots planted with PI #416937 in 1989.*



Here, the value β_1 corresponds to the "asymptote" of growth; that is, the value that growth seems to "level out" at as time grows large. The parameter β_3 is sometimes called a "growth-rate" parameter, because it characterizes how the growth increases as a function of time by decreasing the denominator of (13.3). A scientist may have specific interest in these features.

It is natural in a setting like this to think that each soybean plot evolves over time according to a "growth process" "unique" to that plot. If the model (13.3) is a reasonable way to represent the process a particular plot might undergo, then it is natural to think of representing the situation of **several** such plots by allowing each plot to have its **own** logistic growth model, with its **own** parameters that characterize how large it ultimately gets and its "growth-rate." More formally, if Y_{ij} is the measurement on the growth value at time t_{ij} for the ith plot, we might think of the mean at the **individual plot** level as being represented by (13.3) with plot-specific values for $\beta_1, \beta_2, \beta_3$; that is

$$E(Y_{ij}\,|\boldsymbol{b}_i) = \frac{\beta_{1i}}{1+\beta_{2i}\exp(-\beta_{3i}t_{ij})}, \quad \boldsymbol{\beta}_i = \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \\ \beta_{3i} \end{pmatrix} = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \tag{13.4}$$

where $\boldsymbol{b}_i$ are random effects and $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are suitable matrices allowing covariate information (e.g. genotype) and other considerations to be represented.

This seems like a natural way to think, and it is indeed the way scientists feel comfortable thinking when trying to formally represent the data. Of course, the model (13.4) and more general versions of it (e.g. other functions f) is a **subject-specific** model. Thus, for many applications in the biological sciences, there is a "theoretical" basis for preferring the subject-specific modeling approach.

This model looks very similar to the general form of a generalized linear mixed model, with one important exception. The function f in (13.3) is **not** a function of a **single argument**, so that t_{ij} and the parameter enter the model only in terms of a **linear predictor**. Rather, the way time and parameters enter this model is more complicated. The result is that we have a model one might think of as being even "**more**" **nonlinear**. Indeed, it is the case in biological and physical sciences that theoretical models that may be derived from scientific principles are typically **nonlinear** in this more complicated way.

INFERENCE: The same issues that make model fitting difficult in the generalized linear mixed model case apply here as well – it is not generally possible to write down the likelihood of a data vector in closed form. Again, approximations are often used. A full account of these models in biological and physical applications may be found in Davidian and Giltinan (1995). There is a SAS macro, `nlinmix`, that implements approximate methods to accomplish this fitting; however, as above, it should only be used by those who have a full understanding of the model framework and the approximations used.

13.4 Issues associated with missing data

As we have mentioned, a common issue with longitudinal data, particularly when the units are **humans**, is that data may be **missing**. That is, although we may intend to collect data according to some experimental plan in which all units are seen at the same n times, it is quite often the case that things do not end up this way. The obvious consequence is that the resulting data may not be **balanced** as was originally intended. However, the fact that the data are not balanced is the least of the problems – all of the modern methods we have discussed can handle this issue with ease! The **real** problems are more insidious and were not in fact truly appreciated until quite recently.

As we have discussed, data may be "missing" for different reasons:

1. Mistakes, screw-ups, etc. – for example, a sample is dropped or contaminated, so that a measurement may not be taken.

2. Issues related to the thing being studied (more in a moment).

Missingness of the first type is mainly an annoyance, unless it happens a lot. Missingness of the second type can be a problem; previously in the course we have noted that if missingness happens in this way, then intuition suggests that the very fact that data are missing may have information about the issues under study! The fear is that if we treat the "missingness" as if it has no information, by simply attributing the fact that data vectors are of different length by chance, and this is not really true, the inference we draw may be **misleading**. We are now more formal about this.

TERMINOLOGY: In the literature on missing data, a certain terminology has been developed to characterize different ways missingness happens. This terminology seems somewhat arcane, but it is in widespread use. A statistical reference book that introduces this terminology is Little and Rubin (2002); the recent and current statistical literature always has papers about missing data, too. In reading further about the consequences of missing data, it is useful to be familiar with this terminology.

MISSING COMPLETELY AT RANDOM: In the first type of example, where, say, a sample is dropped and ruined, the fact that the associated observation is thus missing has nothing to do with what is being studied. If the sample is from a patient in a study to compare two treatments, the fact that it was dropped has nothing to do with the treatments and their effect, but rather (most likely) with the clumsiness of the person handling the sample! In the event that missingness is in **no way** related to the issues under study, it is referred to as occurring **completely at random**, or **MCAR**.

The consequence of MCAR is simply that we get less data than we'd hoped. Thus, concerns about sample size may be an issue – we may not be able to have the **power** to detect differences that we'd hoped. If a lot of observations are missing, obviously power will be much less than we had bargained for, and the ability of a study to detect a desired difference or estimate a particular quantity with a desired degree of precision will be compromised. If the problem isn't too bad, then power may not be too seriously affected. However, we don't have to worry about the inferences being misleading. Luckily, because the reason for the missingness has nothing to do with the issues under study, we can assume that the observation and the individual it came from are **similar** to all the others in the study, so that what's left is legitimately viewed as a fair representation of the response of interest in the population of interest. What's left might just be smaller than we hoped.

MISSING AT RANDOM: In the second type of example, we may have a situation where a patient is a participant in a longitudinal study to evaluate a blood pressure medication. The patient's blood pressure at the outset may have been very high, which is why he was recruited into the study. The study plan dictates that the patient be randomized to receive one of two study treatments and return monthly to the hospital to have his blood pressure recorded. For ethical reasons, however, a patient may be **withdrawn** from the study; e.g.

- In many such studies, the study plan dictates that if a patient's measured blood pressure on any visit goes above a certain "danger" level, the patient **must** be removed from the study and have his treatment options be decided based solely on his condition (rather than continue on his randomized treatment, which in some cases may be a placebo). This protects patients in the event they are assigned to a medication that does not work for them.
- The patient's personal physician may review the measurements taken over his previous monthly visits and make a judgment that the patient would be better off being removed from the study treatment. This, of course, would mean that the patient would be removed from the study.

In each of these cases, the patient will have data that are **missing** after a certain point because he is no longer a participant. The **reason** the data will be missing in this way is a **direct result** of observation of his **previous** response values!

Formally, in the event that missingness results because of the values of responses and other variables **already seen** for a unit, the missingness is said to be **at random**, abbreviated **MAR**.

- The reason for this name is that missingness still happens as the result of observation of **random** quantities (the response observed so far), but is no longer necessarily just an annoyance. Because observations on any given patient are subject to (within-patient) variation, it could be that the patient registered above the "danger" level just by chance due to measurement error, and, in reality, his "true" blood pressure is really not high enough to remove him from the study.
- On the other hand, his blood pressure may have registered above the "danger" level because his true pressure really is high.

We have to be concerned that the latter situation is true; if this is the case, then we fear that the data end up seeing are not truly representative of the population; data values from patients who may have registered "high" at some point, whether by chance or not, are not seen.

It turns out that, as long as one uses **maximum likelihood** methods and the assumptions underlying them are correct, estimation of quantities of interest will not be compromised. However, implementation of such methods becomes more complicated, and specialized techniques may be necessary. Thus, some acknowledgment of the problem is required. In the case of GEE methods, things are worse – because these methods are **not** based on a likelihood, it is possible that the estimates themselves will be unreliable; in particular, they can end up being **biased**. Thus, if MAR is suspected, the user must be aware that the usual analyses may be flawed. Fancy methods to "correct" the problem are becoming more popular; these are beyond our scope here.

NONIGNORABLE NONRESPONSE: A more profound case of the second type of missingness is as follows. We discussed earlier in the course the case of patients in a study to evaluate AIDS treatments. Suppose patients are to come to the clinic at scheduled intervals and measurements of **viral load**, a measure of roughly "how much" HIV virus is in the system, are to be made. Patients with "high" viral load tend to be sicker than those with "low" viral load. Viral load is thus likely to be seen increasing over time for patients who are sicker. Moreover, the faster the rate of increase, the more rapidly patients seem to deteriorate.

Suppose that a particular patient fails to come in for his scheduled clinic visits because his disease has progressed to the point where he is too sick to come to the clinic ever again. If we think in terms of a the patient's individual **trajectory** of viral load, a patient who is too sick to come in probably also has a viral load trajectory that is increasing, and may be increasing more quickly than those for other patients who have not become so sick. Thus, if we think formally of a **random coefficient** model to describe viral load as a function of time, e.g.

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij},$$

say, then it may be that the fact that a patient is too sick to come in is reflected in the fact that his individual slope β_{1i} is large and positive.

Now, if the treatment is supposed to be targeting the disease, obviously the fact that this patient is too sick to return (yielding missing data) is caught up with the treatment. If we think of the random coefficient model, the fact that data for this patient end up being missing is a consequence of the fact that his slope β_{1i}, which is supposedly influenced by the treatment, is too large and positive. The patient has missing data not just because of data already seen, but in a sense because of his underlying characteristics (represented through his slope) that will carry him through the **rest of time**, even beyond the current time. Thus, missingness in this example is even more profound than missingness that results from values of data already seen; here, missingness is related to **all** data, observed or not, that we might see for this patient, because those data would all be the consequence of the patient's very steep slope!

This kind of missingness, which is caused by an underlying phenomenon that cannot be observed and operates throughout time, is known as **nonignorable nonresponse**, or **NINR**. Unlike the MAR situation, as the name indicates, if missingness happens this way, then a patient has missing data not just by chance, but because of an underlying characteristic of that patient that may be influenced by the treatment. Thus, we will have a completely unrealistic picture of the population of individuals from the available data, because we will only have incomplete information from part of it. The result can be that estimates of quantities of interest (like the difference in typical slope between two treatments) can be flawed (biased), because information from people who are the sickest is underrepresented.

"Correcting" the problem can be difficult, if not impossible, because the missingness is a consequence of something we **cannot see**! If NINR is suspected, it may not be possible to obtain reliable inferences without making assumptions about things like random effects that cannot be observed. This is a serious drawback, and one that is not always appreciated.

A full treatment of the consequences of missing data and how to handle the issues in the longitudinal context would fill an entire course. The foregoing discussion is meant simply to highlight some of the basic issues.

The book by Verbeke and Molenberghs (2000) devotes considerable attention to issues associated with missing data in the particular context of the **linear mixed effects model**. The book by Fitzmaurice, Laird, and Ware (2004) also offers more extensive introductory discussion.