8 Population-Averaged Models and Generalized Estimating Equations

8.1 Introduction

In this chapter, we consider ***population-averaged*** models for longitudinal data where

(i) the responses may be ***discrete*** ,

(ii) an appropriate model for the ***overall population mean response*** trajectory may be ***nonlinear*** in parameters; and/or

(iii) the aggregate ***variance*** of the response given covariates is possibly a ***function*** of the parameters in the mean model ***and*** additional, unknown ***variance parameters***.

These models extend those of Chapter 5 to incorporate these features and can also be viewed as a ***multivariate extension*** of the mean-variance models for univariate response in Chapter 7. The models are of course all applicable to ***continuous*** responses under conditions (ii) and (iii).

As we discussed in Section 2.7 , when the response is ***discrete*** , specification of a full ***multivariate distribution*** for $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is problematic and thus an infeasible basis for modeling and inference. As we noted in that section, unlike for the normal distribution, where the extension from univariate to multivariate is immediate, multivariate versions of discrete distributions have densities that depend in a complicated way on so-called ***higher-way associations*** among the elements of a data vector. Moreover, in contrast to the multivariate normal distribution, for which there are no ***restrictions*** on the nature of ***pairwise correlations*** among elements of a response vector, discrete multivariate responses do involve ***complicated restrictions*** on these. We demonstrated this in the particular case of ***binary response*** in Section 2.7.

These challenges led to a ***classic paper*** , Liang and Zeger (1986), in which the authors proposed a framework in which one specifies models only for the ***first two moments*** of the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$. In particular, one posits models for the ***mean response*** and ***aggregate variance of the response*** along with a "***working model*** " for the ***aggregate correlation*** among elements of a response vector. The working model is most certainly ***misspecified*** , but the hope is that it can capture the salient features of pairwise correlation structure and lead to a more efficient analysis than, say, erroneously assuming all observations are ***independent***.

Liang and Zeger (1986) popularized this approach, which is now considered fundamental. This original paper restricts to considering responses such as binary data, counts, and so on that from a univariate point of view could be modeled by the ***scaled exponential family*** , and to models for the mean and variance that are thus of the "***generalized linear model type***," with ***no unknown parameters*** in the variance model. However, this restriction is ***unnecessary*** ; accordingly, in this chapter, we consider more general nonlinear mean models and variance functions depending on possibly unknown parameters.

Liang and Zeger (1986) proposed that the model be fitted by solving a ***linear estimating equation*** for the mean parameters along with a suitable method for estimating the parameters in the ***working correlation model***. Subsequently, other authors proposed refinements and extensions of this approach. There are numerous references that cover the types of estimating equations we are about to discuss. Some of the key references are Prentice (1988), Zhao and Prentice (1990), Prentice and Zhao (1991), and Liang, Zeger, and Qaqish (1992). See also Chapter 9 of Vonesh and Chinchilli (1997) and Chapter 3 of Fitzmaurice et al. (2009) and the references therein.

In this chapter, we discuss this modeling framework and key inferential approaches. We also discuss an issue that we have so far downplayed, that of the difficulties that arise in modeling and interpretation when covariate information is ***time-dependent***.

8.2 Model specification

DATA, RESTATED: We first restate the form of the observed data for convenience. These data are

$$(\boldsymbol{Y}_i, \boldsymbol{z}_i, \boldsymbol{a}_i) = (\boldsymbol{Y}_i, \boldsymbol{x}_i), \quad , i = 1, \ldots, m,$$

independent across i, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$ (possibly different times for different individuals); $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$ comprising ***within-individual*** covariate information $\boldsymbol{u}_i$ and the t_{ij}; $\boldsymbol{a}_i$ is a vector of ***among-individual*** covariates; and $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$.

Before we state the basic model, we note an important consideration that we discuss in detail in Section 8.6. Up to now, we have assumed that the among-individual covariates $\boldsymbol{a}_i$ ***do not change over time*** , as is the case with individual characteristics such as gender, treatment group in a conventional randomized study, or characteristics such as age or weight ascertained once at ***baseline***.

If there are any within-individual covariates $\boldsymbol{u}_i$, we have assumed that these are either ***time-independent***, as in the case of a one-time drug dose D_i administered to subject i in a pharmacokinetic study, or are determined according to a ***fixed design***, as would be the case in a pharmacokinetic study where each participant is to receive repeated doses over several fixed ***dosing intervals***.

Of course, this ***need not always*** be the case. To fix ideas, recall the following example.

EXAMPLE 6: Maternal smoking and child respiratory health, continued. Recall from Chapter 1 the example of the Six Cities Study. In this part of the study, $m = 300$ mother-child pairs were to be examined once a year at the child's ages 9–12, so that the ***intended number*** of examinations for each pair i is $n = 4$, although some pairs are missing data at some examinations, so that $n_i \leq 4$ in general. At each examination, the outcome of interest is the ***binary response*** "***wheezing status***," coded as 0 (no) or 1 (yes), where 1 corresponds to respiratory problems. ***Maternal smoking status*** was recorded at each examination as a categorical variable indicating the current level of the mother's smoking: 0=none, 1=moderate, 2=heavy. The ***city*** in which each pair lived was also recorded.

The goals of this portion of the study were to determine how the ***wheezing response pattern*** in the population changes with age and how it might be associated with ***level of maternal smoking***. These are questions regarding ***population-averaged*** phenomena; for example, as is usually the case in studies of ***public health***, the second question is focused on the association between maternal smoking and child respiratory status overall in the ***population***.

As we discussed in Section 2.2, from this point of view, the ***covariate*** maternal smoking status would ordinarily be interpreted as an ***among-individual*** covariate, as it reflects how a child was ***treated*** over the period of the study and thus is relevant to the overall ***population-level*** question.

Thus, in this study, if mother-child pair i from city c_i (= 0 for Portage, = 1 for Kingston) was examined at ages t_{ij}, $j = 1, \ldots, n_i$, we have $\boldsymbol{a}_i = (c_i, s_{i1}, \ldots, s_{in_i})^T$, where s_{ij}, $j = 1, \ldots, n_i$, is such that $s_{ij} = 0, 1, 2$ according to mother i's smoking status at t_{ij}. There are no within-individual level covariates. We can thus write $\boldsymbol{x}_i$, the collection of all covariates on pair i, as

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in_i}^T)^T = \{(t_{i1}, c_i, s_{i1})^T, \ldots, (t_{in_i}, c_i, s_{in_i})^T\}^T, \tag{8.1}$$

where $\boldsymbol{x}_{ij} = (t_{ij}, c_i, s_{ij})^T$ records the age, city, and smoking status at the jth examination for pair i.

We contrast this situation with one in which ***among-individual*** covariates do not change over time.

EXAMPLE 5: Epileptic seizures and chemotherapy, continued. Recall from Chapter 1 the clinical trial conducted in $m = 59$ subjects suffering from simple or partial seizures, who were randomized to the anti-epileptic drug progabide or placebo. On each, a ***baseline measure*** of each subject's propensity for seizures was recorded, namely, the number of seizures suffered in the 8 weeks leading up to the start of the study. Also recorded was each subject's ***age*** at the start of assigned treatment. After initiation of assigned treatment, the ***number of seizures*** experienced by each subject in $n = 4$ consecutive two-week periods was recorded, so that the response is a ***count***. Seizure counts were recorded for all m subjects at all n periods, so there are no missing data.

Thus, in this study, the ***among-individual covariates*** are the assigned treatment $\delta_i = 0$ for placebo and 1 for progabide, baseline seizure count c_i, and age a_i at the start of the study, both of which are ***time-independent*** , and there are no within-individual covariates. Thus, letting $\boldsymbol{a}_i = (\delta_i, c_i, a_i)^T$, $i = 1, \ldots, m$, and letting $t_j = 1, 2, 3, 4$ indicate the observation period, $j = 1, \ldots, 4$. The we can write $\boldsymbol{x}_i$, the collection of all covariates on subject i, as

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in_i}^T)^T = \{(t_{i1}, \boldsymbol{a}_i^T)^T, \ldots, (t_{in_i}, \boldsymbol{a}_i^T)^T\}^T = \{(t_1, \boldsymbol{a}_i^T)^T, \ldots, (t_4, \boldsymbol{a}_i^T)^T\}^T. \tag{8.2}$$

In (8.2), $\boldsymbol{x}_{ij}$, the component of $\boldsymbol{x}_i$ associated with time j, involves among-individual covariates that ***do not change*** over time. Thus, the subscript j thus corresponds only to the time of the jth response measure. In constrast, in (8.1), $\boldsymbol{x}_{ij}$ involves among-individual covariate information that ***does change*** over time.

Although it is conventional to write the model for mean response in terms of $\boldsymbol{x}_{ij}$ as we do below, it is it is ***critical*** to appreciate exactly what one ***implicitly assumes*** when specifying such a model in a situation like (8.1), as we discuss momentarily and in detail in Section 8.6.

BASIC MODEL: We focus on the general PA mean-covariance model of the form

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta), \qquad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i). \tag{8.3}$$

- In model (8.3), with $\boldsymbol{x}_i$ partitioned as $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in_i}^T)^T$ as in (8.1) and (8.2), $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ is ordinarily taken to be the $(n_i \times 1)$ vector

$$\boldsymbol{f}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f(\boldsymbol{x}_{in_i}, \beta) \end{pmatrix} \quad (n_i \times 1) \tag{8.4}$$

  for some function $f(\boldsymbol{x}, \beta)$ that may be ***nonlinear*** in β $(p \times 1)$. We say more about this model below.

- The covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ $(n_i \times n_i)$ is taken to have the form

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\beta, \theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\beta, \theta, \boldsymbol{x}_i). \quad (8.5)$$

- In (8.5), $\boldsymbol{T}_i(\beta, \theta, \boldsymbol{x}_i)$ is the $(n_i \times n_i)$ diagonal matrix whose diagonal elements are the models for $\text{var}(Y_{ij}|\boldsymbol{x}_i)$, e.g., involving a variance function

$$\text{var}(Y_{ij}|\boldsymbol{x}_i) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_{ij}),$$

  so depending on possibly unknown variance parameters $\theta = (\sigma^2, \delta^T)^T$ $(r \times 1)$. In the case of popular models for responses that would from a univariate point of view follow a ***scaled exponential family*** distribution, the variance function g^2 ***does not depend*** on unknown parameters δ. Moreover, for some of these distributions, such as the Bernoulli for binary response and the Poisson for responses in the form of counts, the ***scale parameter*** σ^2 might also be taken to be known and $\sigma^2 = 1$. In the case of possible ***overdispersion***, as discussed in Section 7.2, an unknown scale parameter σ^2 would be incorporated.

  Considerations for specifying an appropriate ***variance function model*** are as in Section 7.2.

- In (8.5), the $(n_i \times n_i)$ matrix $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ is a correlation matrix that generally depends on $\boldsymbol{x}_i$ only through the within-individual times or other conditions t_{ij} at which observations in $\boldsymbol{Y}_i$ are taken. Here, α $(s \times 1)$ is a vector of unknown correlation parameters. Considerations for specifying $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ are discussed momentarily.

- The vector of variance and correlation parameters $\xi = (\theta^T, \alpha^T)^T$ $(r + s \times 1)$ may be ***entirely unknown***. Alternatively, it may be that only α is unknown in models where the form of the variance function is ***entirely specified***, as discussed above.

Model (8.3) can be viewed as a ***multivariate analog*** to the univariate mean-variance models discussed in Chapter 7. Thus, it should come as no surprise that inferential strategies for (8.3) exploit some of the same ideas as in the univariate case.

In particular, estimation of β and ξ is typically carried out by solution of ***linear and quadratic estimating equations*** that are similar in spirit to those discussed in Chapters 5, 6, and 7. Of necessity, these equations are ***more complicated***, as we will see in Sections 8.3 and 8.4, although the basic principles are the same. The term ***generalized estimating equations*** (GEEs), first coined by Liang and Zeger (1986), has come to refer broadly to the body of techniques for inference for β and ξ based on solution of such estimating equations.

- As suggested by the title of Liang and Zeger (1986), "Longitudinal data analysis using generalized linear models," and noted above, the original formulation was restricted to responses and models of the "***generalized linear model***" type. This explains why in much of the classical literature on GEEs there are ***no unknown variance parameters*** δ, and interest focuses exclusively on estimating ***correlation parameters*** α in a "working" correlation model as discussed below and possibly a scale parameter σ^2 in the case of ***overdispersion***.

- Our development allows the model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ to involve ***both*** unknown variance parameters θ and unknown correlation parameters α. We note simplifications that occur in the case where the variance function does not depend on any unknown parameters θ.

WORKING CORRELATION MODEL: In accordance with the considerations given at the beginning of this chapter, $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ is often a ***working correlation model*** that is acknowledged in many circumstances to be ***incorrect*** but that is specified as a way to accommodate the expected ***aggregate pairwise correlations*** among elements of $\boldsymbol{Y}_i$.

- As for the linear population-averaged models in Chapter 5, the modeling framework (8.3) is feasible in situations where there are *n* ***intended times*** at which all individuals are to be observed. (We discuss implications of ***missing responses*** in this context in Section 8.7.)

- In this case, $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ is might taken to be ***completely unstructured***, so that $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ involves $n(n-1)/2$ unknown correlation parameters in addition to possibly unknown variance parameters θ. Although the large sample theory we discuss in Section 8.5 suggests that, analogous to that in Sections 5.5 and 7.4, the properties of the estimator for β ***do not depend*** on whether or not ξ is ***estimated or known***, in finite samples (m), estimation of a large number of correlation and variance parameters can lead to ***inefficient*** estimation of β.

- This framework can also be used in more general situations where the observation times t_{ij} are ***different for different individuals***, although the working correlation models that are practically feasible in this situation are ***limited***.

- Accordingly, it is popular to use working correlation models that involve a ***small number of correlation parameters*** α. If it is thought that ***within-individual sources of correlation*** due to time-ordered data collection are dominant, AR(1), exponential, or Gaussian correlation models might be used. Conversely, if ***among-individual sources of correlation*** dominate, a ***compound symmetric*** correlation model might be selected. This model is in principle feasible even if the observation times are ***different*** for different individuals.

- Recognizing that such working models are ***almost certainly misspecified*** , it is commonplace to use the appropriate form of the ***robust sandwich*** or ***empirical*** estimator for the covariance matrix of the approximate sampling distribution of the estimator for β to assess uncertainty, as we demonstrate in Section 8.5.

SPECIFICATION OF THE MEAN MODEL: Models for the $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ are specified in accordance with the particular type of response and the nature of the questions of interest. For example, if Y_{ij} is ***binary*** , then a natural model is a general ***logistic*** model as in (7.2), i.e.,

$$f(\boldsymbol{x}_{ij},\beta) = \frac{\exp\{h(\boldsymbol{x}_{ij})^T\beta\}}{1+\exp\{h(\boldsymbol{x}_{ij})^T\beta\}}, \quad \text{or equivalently} \quad \text{logit}\{f(\boldsymbol{x}_{ij},\beta)\} = \log\left\{\frac{f(\boldsymbol{x}_{ij},\beta)}{1-f(\boldsymbol{x}_{ij},\beta)}\right\} = h(\boldsymbol{x}_{ij})^T\beta, \tag{8.6}$$

where $h(\cdot)$ is a vector of functions of $\boldsymbol{x}_{ij}$. Similarly, for Y_{ij} in the form of counts, a ***loglinear model***

$$f(\boldsymbol{x}_{ij},\beta) = \exp\{h(\boldsymbol{x}_{ij})^T\beta\) \quad \text{or equivalently} \quad \log\{f(\boldsymbol{x}_{ij},\beta)\} = h(\boldsymbol{x}_{ij})^T\beta \tag{8.7}$$

would be an obvious choice. Other models, linear or nonlinear, can of course be posited as appropriate. E.g., for a ***continuous response*** for which the population mean trajectory approaches an ***asymptote*** as $t \to \infty$, a possible model is

$$f(\boldsymbol{x}_{ij},\beta) = \beta_1 + (\beta_2 - \beta_1)\exp\{-\exp(-\beta_3)t_{ij}\},$$

where the rate constant is parameterized to enforce positivity.

As noted in (8.4), with $\boldsymbol{x}_i$ partitioned as $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T,\ldots,\boldsymbol{x}_{in_i}^T)^T$, the jth element of $\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)$ is ordinarily taken to depend on $\boldsymbol{x}_i$ ***through*** $\boldsymbol{x}_{ij}$ ***only***. It is ***critical*** that the data analyst understand the implications of this, as we now discuss.

Specifically, as $\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)$ is a model for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, its jth component is a model for $E(Y_{ij}|\boldsymbol{x}_i)$. Thus, restricting the jth component of $\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)$, $f(\boldsymbol{x}_{ij},\beta)$ to depend on $\boldsymbol{x}_i$ ***only through*** $\boldsymbol{x}_{ij}$ implicitly embodies the assumption that

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{in_i}) = E(Y_{ij}|\boldsymbol{x}_{ij}). \tag{8.8}$$

It is ***critical*** to recognize that (8.8) is an ***assumption*** that ***may or may not*** hold.

- In a situation like that of ***EXAMPLE 5*** , the seizure trial, $\boldsymbol{x}_{ij}$ involves only the ***time-independent covariate*** $\boldsymbol{a}_i$, which includes treatment, baseline seizure count, and age at study entry. Here, then, $\boldsymbol{x}_{ij}$ varies with j ***only*** through the ***planned*** time periods t_j, so, effectively,

  $$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{a}_i)$$

  for all j, in which case (8.8) with $\boldsymbol{x}_{ij} = (t_{ij},\boldsymbol{a}_i^T)^T$ ***necessarily holds***.

- Similarly, consider a study where each individual is assigned to receive n different doses of a drug d_j on n separate occasions $t_1, \ldots, t_n$, and the response Y_{ij} is ascertained at each occasion for each individual i. Interest focuses on the ***relationship*** between population mean response and dose. Assuming ***no missing data***, $\boldsymbol{x}_{ij} = (t_j, d_j)$ for all i, and thus $\boldsymbol{x}_i = \{(t_1, d_1)^T, \ldots, (t_n, d_n)^T\}^T$. Here, then, the $\boldsymbol{x}_{ij}$, are ***fixed by design***, specified ***a priori*** in a way that is ***unrelated to*** the responses that they might elicit, rather than ***observed***.

  Thus, the relationship between dose and response is ***clear cut***. E.g., if the goal is to evaluate the relationship between mean response and ***current dose level***, assuming that the occasions are ***sufficiently far apart*** that effects of previous doses on the current response have "***washed out***," (8.8), so that the response at occasion j depends on all doses ***only*** through the dose given at occasion j, is reasonable. If we redefine $\boldsymbol{x}_{ij} = (t_j, d_1 + \cdots + d_j)$ and wish to evaluate the relationship between mean response and ***cumulative dose***, the assumption in (8.8) is also reasonable. The main issue is that, because the $\boldsymbol{x}_{ij}$ are ***fixed in advance***, their values are ***not impacted*** by the response. Contrast this situation with the following.

- Consider the setting of the Six Cities study. Recall that $\boldsymbol{x}_{ij} = (t_{ij}, c_i, s_{ij})^T$, where t_{ij} is the time variable (examination occasion corresponding to the age of the child), c_i is the (fixed) city, and s_{ij} is the mother's smoking status status at the jth examination for pair i. Here, smoking status is ***observed*** rather than dictated by design.

  Suppose that, at examination j for pair i, the child's wheezing status $Y_{ij} = 1$, and the mother is currently engaging in heavy smoking, $s_{ij} = 2$. After seeing her child's wheezing result, the mother decides to ***cut back*** on her ***future smoking***. In this case, her smoking status $s_{i,j+1}$ at the $(j+1)$th examination is ***associated with*** and thus ***not independent of***, the child's wheezing status Y_{ij} at the jth examination. Clearly, then, it ***cannot be true*** that

$$E(Y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{x}_{i,j+1}) = E(Y_{ij}|\boldsymbol{x}_{ij}), \tag{8.9}$$

  because, given the smoking status s_{ij} in $\boldsymbol{x}_{ij}$ at examination j, the mother's smoking status $s_{i,j+1}$ in $\boldsymbol{x}_{i,j+1}$ is ***not independent*** of Y_{ij}. Applying this reasoning more generally, it should be clear that (8.8) ***cannot hold*** under these circumstances.

- From a ***causal perspective***, under which we wish to attribute mother's smoking status s_{ij} at examination j as "***causing***" the child's wheezing status Y_{ij} at j, smoking status $s_{i,j+1}$ ***confounds*** the relationship between Y_{ij} and s_{ij}.

That is, if (8.9) holds, we ***cannot hope*** to isolate the effect of smoking status at examination j on wheezing at j. If we adopt a model in that implicitly assumes (8.8), the parameter β in that model ***does not characterize*** the ***causal mechanism*** of interest, and thus ***misleading inferences*** could result.

- In general, for ***time-dependent among-individual covariates*** that are ***not fixed by design***, as in the dose-response study above, ***great care*** must be taken in modeling relationships between response and covariates and interpretating the model. ***Blindly adopting*** a model for which conditional mean at time j depends only on covariates at time j as in (8.4) can lead to ***complex difficulties*** with ***interpretation***. We discuss this further in Section 8.6.

8.3 Linear estimating equations

LINEAR ESTIMATING EQUATION: Analogous to developments in Chapter 5 for linear PA models and Chapter 7 for the case of univariate response, we can derive an estimating equation for β by considering the situation where the covariance matrix $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***known***. Writing $\boldsymbol{V}_i$ to denote this known matrix and adopting a working assumption of ***normality*** for $\boldsymbol{Y}_i|\boldsymbol{x}_i$, we can differentiate the loglikelihood

$$\log L = -(1/2)\sum_{i=1}^{m}\left[\log|\boldsymbol{V}_i| + \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\right]$$

with respect to β to obtain the estimating equation

$$\sum_{i=1}^{m}\boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} = \boldsymbol{0}, \qquad \boldsymbol{X}_i(\beta) = \begin{pmatrix} f_\beta^T(\boldsymbol{x}_{i1},\beta) \\ \vdots \\ f_\beta^T(\boldsymbol{x}_{in_i},\beta) \end{pmatrix} \quad (n_i \times p). \tag{8.10}$$

This follows from the same matrix differentiation results used in Section 5.3. Alternatively, we can take a "***weighted least squares***" point of view to arrive at (8.10), analogous to (7.18).

When $\boldsymbol{V}_i$ is taken to depend on β but not on covariance parameters ξ, we can make a multivariate analogy to the ***maximum likelihood estimating equation*** for the ***scaled exponential family*** in (7.15) to arrive at an estimating equation of the form in (8.10), where now $\boldsymbol{V}_i$ also depends on β.

These considerations suggest that, when the model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ depends on ***both*** β and covariance parameters ξ, we estimate β by solving the ***linear estimating equation***

$$\sum_{i=1}^{m}\boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} = \boldsymbol{0}, \tag{8.11}$$

where an estimator for ξ is substituted.

Analogous to the streamlined notation used in Chapter 5, define $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_m^T)^T$,

$$\boldsymbol{f}(\beta) = \{\boldsymbol{f}_1^T(\boldsymbol{x}_1, \beta), \ldots, \boldsymbol{f}_m^T(\boldsymbol{x}_1, \beta)\}^T \quad (N \times 1), \quad \boldsymbol{X}(\beta) = \begin{pmatrix} \boldsymbol{X}_1(\beta) \\ \vdots \\ \boldsymbol{X}_m(\beta) \end{pmatrix} \quad (N \times p), \tag{8.12}$$

$$\boldsymbol{V}(\beta, \xi) = \text{block diag}\{\boldsymbol{V}_1(\beta, \xi, \boldsymbol{x}_1), \ldots, \boldsymbol{V}_m(\beta, \xi, \boldsymbol{x}_m)\}, \quad (N \times N).$$

Then we can write (8.11) compactly as

$$\boldsymbol{X}^T(\beta)\boldsymbol{V}^{-1}(\beta, \xi)\{\boldsymbol{Y} - \boldsymbol{f}(\beta)\} = \boldsymbol{0}. \tag{8.13}$$

- ***Even if*** the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***misspecified***, the estimating equation in (8.11) and equivalently (8.13) is ***unbiased*** as long as the model for the mean is correctly specified ***and*** satisfies conditions we discuss in Section 8.6.
- Given that the correlation model incorporated in $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***quite possibly misspecified***, as discussed above, this is ***critically important***.

Henceforth in this chapter, except when we discuss modeling considerations in Section 8.6, we assume that the model $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ is ***correctly specified***.

IMPLEMENTATION: By analogy to the univariate case, an obvious strategy for solving (8.11) is to use a multivariate version of the ***generalized least squares algorithm*** for solving (7.22) in Section 7.3. Specifically, start with an ***initial estimate*** $\widehat{\beta}^{(0)}$. A natural choice for $\widehat{\beta}^{(0)}$ is the ***OLS estimator*** treating all N elements of $\boldsymbol{Y}$ as if they were ***mutually independent*** with the ***same*** conditional variance; i.e., replacing $\boldsymbol{V}$ in (8.13) by a $(N \times N)$ identity matrix. That the OLS estimator is ***consistent*** for the true value of β follows from arguments in Section 8.5, as we discuss shortly. Then at iteration ℓ:

1. Holding β fixed at $\widehat{\beta}^{(\ell)}$, estimate ξ by $\widehat{\xi}^{(\ell)}$; we discuss approaches to doing this momentarily.
2. Substitute $\widehat{\xi}^{(\ell)}$ for ξ in $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ in (8.11). Then holding ξ fixed at $\widehat{\xi}^{(\ell)}$ solve the linear estimating equations (8.11) in β to obtain $\widehat{\beta}^{(\ell+1)}$. Set $\ell = \ell + 1$ and return to step 1.

A variation on step 2 is to substitute $\widehat{\beta}^{(\ell)}$ in $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ in (8.11) along with $\widehat{\xi}^{(\ell)}$, so that the "***weights***" are held fixed.

As in the univariate case, one would iterate between steps 1 and 2 until "***convergence***." As in that case, it is not necessarily true that this algorithm should ***converge***, as the procedure does not correspond to maximizing some ***objective function***.

ESTIMATION OF COVARIANCE PARAMETERS: By analogy to the univariate case, a natural approach to estimating ξ would be to solve a suitable ***quadratic estimating equation***, as we discuss shortly.

In the early papers on GEEs in the biostatistical literature by Liang and Zeger, estimation of ξ was advocated based on ***simple, moment-based approaches***. Actually, as this early work was in the context of "***generalized linear model-type***" problems, this involved estimation only of correlation parameters α, as in this setting the variance function ***does not depend*** on unknown parameters (except perhaps a scale parameter σ^2 in the case of overdispersion or distributions like the gamma).

For example, for the ***exponential correlation model*** , reparameterized here as

$$\text{corr}(Y_{ij}, Y_{ij'}|\boldsymbol{x}_i) = \alpha^{|t_{ij}-t_{ij'}|},$$

where Y_{ij} and $Y_{ij'}$ are observed at times t_{ij} and $t_{ij'}$, and $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ depends on the scalar parameter α. Assume that $\text{var}(Y_{ij}|\boldsymbol{x}_{ij}) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_{ij})$ with δ ***known***. Given an estimate $\widehat{\beta}^{(\ell)}$ at the ℓth iteration of the above algorithm, the weighted residuals

$$wr_{ij} = \{Y_{ij} - f(\boldsymbol{x}_{ij}, \widehat{\beta}^{(\ell)})\}/g(\widehat{\beta}^{(\ell)}, \delta, \boldsymbol{x}_{ij})$$

have ***approximate*** mean 0 and satisfy

$$E(wr_{ij}wr_{ij'}|\boldsymbol{x}_i) \approx \sigma^2\alpha^{|t_{ij}-t_{ij'}|}.$$

Taking logarithms of both sides of this expression yields the approximate relationship

$$\log(wr_{ij}wr_{ij'}) \approx \log\sigma^2 + |t_{ij} - t_{ij'}|\log\alpha;$$

thus, the suggestion was to form ***all pairs of lagged residuals*** for each i, ***pool*** them together, and estimate $\log\alpha$ by ***simple linear regression*** of the $\log(wr_{ij}wr_{ij'})$ on the $|t_{ij} - t_{ij'}|$. The resulting estimator may be exponentiated to yield an estimator for α.

Similar ***moment-based methods*** were proposed by Liang and Zeger for other correlation models.

In subsequent publications, it was proposed ***instead*** to estimate α, and more generally ξ, by solving a ***suitable estimating equation*** derived analogously to those used in Chapters 5 and 7 by starting with the loglikelihood under the assumption $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is normally distributed, namely

$$\log L = -(1/2)\sum_{i=1}^{m}\left[\log|\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)| + \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\right]. \tag{8.14}$$

Using ***matrix differentiation results*** in Appendix A as we did in Section 5.3, treating β as fixed and taking the partial derivatives of (8.14) with respect to each element ξ_k, $k = 1,\ldots,r+s$, of ξ, we obtain the $(r+s\times 1)$ estimating equations

$$(1/2)\sum_{i=1}^{m}\Big(\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}$$
$$-\operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\right]\Big) = 0,\quad k = 1,\ldots,r+s. \tag{8.15}$$

By analogy to the linear and univariate cases, if $\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)$ is ***correctly specified***, then using the result for the expectation of a quadratic form in Appendix A, it is straightforward (verify) that (assuming expectation is under the parameter values β and ξ)

$$\begin{aligned}E\Big[\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\mid\boldsymbol{x}_i\Big]&\\ = \operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\right]&\\ = \operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\right],&\end{aligned}$$

from whence it follows that (8.15) is an ***unbiased estimating equation***.

These considerations lead to the $(p+r+s\times 1)$ system of estimating equations

$$\sum_{i=1}^{m}\boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} = \boldsymbol{0}, \tag{8.16}$$

$$(1/2)\sum_{i=1}^{m}\Big(\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}$$
$$-\operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\right]\Big) = 0,\quad k = 1,\ldots,r+s. \tag{8.17}$$

The multiplicative factor of (1/2) in the equation for ξ in (8.17) could be disregarded, but we maintain it for now, as it proves important for the developments of Section 8.9, the motivation for which we now discuss.

ALTERNATIVE FORM OF THE QUADRATIC ESTIMATING EQUATION FOR ξ: In a series of papers in the late 1980s/early 1990s, Prentice (1988), Zhao and Prentice (1990), and Prentice and Zhao (1991), an ***alternative way*** of representing estimating equations such as (8.17) was popularized and became the standard way to write such equations.

Recall from (7.38) that we recognized that the ***system of estimating equations*** (7.37) could be seen to be of the ***generic form*** (7.38). In the current context, this generic form is

$$\sum_{i=1}^{m} \mathcal{D}_i^T(\eta)\mathcal{V}_i^{-1}(\eta)\{\boldsymbol{s}_i(\eta) - \boldsymbol{m}_i(\eta)\} = \boldsymbol{0}, \tag{8.18}$$

where η is a $(k \times 1)$ vector of parameters; $\boldsymbol{s}_i(\eta)$ is a $(v \times 1)$ vector of functions of Y_i, $\boldsymbol{x}_i$, and η;

$$\boldsymbol{m}_i(\eta) = E\{\boldsymbol{s}_i(\eta)|\boldsymbol{x}_i\} \quad (v \times 1), \qquad \mathcal{V}_i(\eta) = \text{var}\{\boldsymbol{s}_i(\eta)|\boldsymbol{x}_i\} \quad (v \times v), \qquad \mathcal{D}_i(\eta) = \partial/\partial\eta^T \boldsymbol{m}_i(\eta) \quad (v \times k).$$

It is possible to express (8.17) in the form (8.18). To demonstrate this directly analytically is ***quite involved*** and is presented in Section 8.9.

Instead, the approach was to develop directly a set of equations of the form (8.18) for estimating ξ. This formulation leads to an estimating equation for ξ of the form discussed in the papers cited above.

We consider β fixed for now and consider estimation of ξ only; we combine the equations for ξ we now derive with the ***linear equation*** (8.11) for β at the end of the following argument, and we consider combining with ***quadratic estimating equations*** for β in the next section.

If we are interested in estimating the elements of ξ, which describe an ***entire covariance structure*** (variances and correlations, so variances and covariances), we must consider the ***variances*** of each element of a response vector and ***all pairwise associations*** among elements of a response vector. Of course, if there are ***no unknown variance parameters*** θ, we need only consider associations.

Let $v_{ijk}(\beta, \xi)$ be the (j, k) element of $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ (which is of course equal to the (k, j) element by symmetry). Then if $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is specified correctly,

$$E\left[\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}^2|\boldsymbol{x}_i\right] = v_{ijj}(\beta, \xi),$$

$$E\left[\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}\{Y_{ik} - f(\boldsymbol{x}_{ik}, \beta)\}|\boldsymbol{x}_i\right] = v_{ijk}(\beta, \xi).$$

The idea is to identify $\boldsymbol{s}_i(\eta)$ in (8.18) as containing ***all quadratic and distinct cross-product terms*** of the form $\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}^2$ and $\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}\{Y_{ik} - f(\boldsymbol{x}_{ik}, \beta)\}$ for $j, k = 1, \ldots, n_i$, so that $\boldsymbol{m}_i(\eta)$ contains the expectations of these given above, which comprise the distinct elements of $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$.

- By ***symmetry***, for $\boldsymbol{Y}_i$ ($n_i \times 1$), there are n_i ***quadratic terms*** (one for each entry of $\boldsymbol{Y}_i$) and $n_i(n_i-1)/2$ ***distinct crossproducts*** (i.e., number of covariances), for a total of $n_i(n_i+1)/2$ distinct terms.

- Explicitly, the $n_i(n_i+1)/2$ distinct terms are the n_i ***squared deviations*** $\{Y_{i1}-f(\boldsymbol{x}_{i1},\beta)\}^2, \ldots, \{Y_{in_i}-f(\boldsymbol{x}_{in_i},\beta)\}^2$ and the $n_i(n_i-1)/2$ ***crossproduct terms***

$$\{Y_{i1}-f(\boldsymbol{x}_{i1},\beta)\}\{Y_{i2}-f(\boldsymbol{x}_{i2},\beta)\},\ \{Y_{i1}-f(\boldsymbol{x}_{i1},\beta)\}\{Y_{i3}-f(\boldsymbol{x}_{i3},\beta)\}, \ldots,$$

$$\{Y_{i,n_i-1}-f(\boldsymbol{x}_{i,n_i-1},\beta)\}\{Y_{in_i}-f(\boldsymbol{x}_{in_i},\beta)\}.$$

  Thus, $\boldsymbol{s}_i(\eta)$ in (8.18) should be of length $n_i(n_i+1)/2$, assuming that the model contains unknown variance parameters. If it does not, then only the $n_i(n_i-1)/2$ crossproduct terms are required.

- Of course, as the quadratic estimating equation (8.17) depends on the quadratic form in $\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}$, it ***also depends*** on these squared deviations and crossproducts.

To formalize, define

$$u_{ijk}(\beta)=\{Y_{ij}-f(\boldsymbol{x}_{ij},\beta)\}\{Y_{ik}-f(\boldsymbol{x}_{ik},\beta)\}. \tag{8.19}$$

Then we can collect the ***distinct*** $u_{ijk}(\beta)$ defined in (8.19) in a vector of length $n_i(n_i+1)/2$ (with unknown variance parameters θ) or $n_i(n_i-1)/2$ (with no unknown variance parameters) in some order. In the former case, suppressing the dependence of the u_{ijk} on β for brevity, let

$$\boldsymbol{u}_i(\beta)=(u_{i11},u_{i12},u_{i13},\ldots,u_{i22},u_{i23},\ldots,u_{in_i-1n_i-1},u_{in_i-1n_i-1},u_{in_in_i})^T.$$

- Here, we have used the ordering imposed by defining

$$\boldsymbol{u}_i(\beta)=\text{vech}\left[\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\right],$$

  where vech($\cdot$) is defined in Appendix A. If there are ***no unknown variance parameters*** in ξ, $\boldsymbol{u}_i(\beta)$ would be defined by ***deleting*** the squared components.

We can define a ***corresponding vector***

$$\boldsymbol{v}_i(\beta,\xi)=\{v_{i11}(\beta,\xi),v_{i12}(\beta,\xi),v_{i13}(\beta,\xi),\ldots,v_{i22}(\beta,\xi),v_{i23}(\beta,\xi),\ldots,v_{in_in_i}(\beta,\xi)\}^T;$$

i.e., $\boldsymbol{v}_i(\beta,\xi)=\text{vech}\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}$. Clearly,

$$E\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\}=\boldsymbol{v}_i(\beta,\xi).$$

If $\boldsymbol{u}_i(\beta)$ contains no squared components (i.e., no unknown variance parameters), then the corresponding elements of $\boldsymbol{v}_i$ would be deleted.

Thus, identifying $\boldsymbol{s}_i(\eta) = \boldsymbol{u}_i(\beta)$, we have $\boldsymbol{m}_i(\eta) = \boldsymbol{v}_i(\beta, \xi)$.

It is important to recognize in reading the literature that there are ***variations*** on the construction we describe here. For example, some authors instead base the equations on $\boldsymbol{s}_i(\eta) = \text{vech}(\boldsymbol{Y}_i \boldsymbol{Y}_i^T)$, and/or may "***stack***" things in a different order.

With $\boldsymbol{m}_i(\eta) = \boldsymbol{v}_i(\beta, \xi)$, to identify $\mathcal{D}_i(\eta)$ in (8.18), define

$$\boldsymbol{E}_i(\beta, \xi) = \partial / \partial \xi \; \boldsymbol{v}_i(\beta, \xi).$$

$\boldsymbol{E}_i(\beta, \xi)$ has $n_i(n_i+1)/2$ or $n_i(n_i-1)/2$ rows, depending on the form of $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$, and $(r+s)$ columns.

To identify $\mathcal{V}_i(\eta)$ in (8.18), we must find, suppressing dependence of $\boldsymbol{u}_i(\beta)$ and its elements on β for brevity,

$$\text{var}(\boldsymbol{u}_i | \boldsymbol{x}_i) = \boldsymbol{Z}_i(\beta, \xi),$$

say. To specify this matrix, we must be ***willing to make assumptions*** about quantities of the general form

$$\text{cov}(u_{ijk}, u_{i\ell p} | \boldsymbol{x}_i) = E(u_{ijk} u_{i\ell p} | \boldsymbol{x}_i) - E(u_{ijk} | \boldsymbol{x}_i) E(u_{i\ell p} | \boldsymbol{x}_i). \tag{8.20}$$

Clearly, this is ***quite complex***. To demonstrate, consider the particular model for $\text{var}(\boldsymbol{Y}_i | \boldsymbol{x}_i)$ such that

$$\text{var}(Y_{ij} | \boldsymbol{x}_i) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_{ij}),$$

with correlation matrix $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$. Define

$$\epsilon_{ij} = \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \} / \{ \sigma g(\beta, \delta, \boldsymbol{x}_{ij}) \}.$$

Then, of course $E(\epsilon_{ij} | \boldsymbol{x}_i) = 0$, and $\text{var}(\epsilon_{ij} | \boldsymbol{x}_i) = 1$, where expectation here and subsequently is under the parameter values β and ξ. Clearly, the elements of $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$ are ***correlated***, with correlation matrix equal to $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ (assuming as we are that this matrix is correctly specified for the purposes of formulating this approach).

Under this model, for $j, k = 1, \ldots, n_i$,

$$u_{ijj} = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_{ij}) \epsilon_{ij}^2, \quad u_{ijk} = \sigma^2 g(\beta, \delta, \boldsymbol{x}_{ij}) g(\beta, \theta, \boldsymbol{x}_{ik}) \epsilon_{ij} \epsilon_{ik}.$$

Thus,

$$v_{ijj} = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}) E(\epsilon_{ij}^2 | \boldsymbol{x}_i) = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}),$$

$$v_{ijk} = \sigma^2 g(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}) g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_{ik}) E(\epsilon_{ij}\epsilon_{ik} | \boldsymbol{x}_i) = \sigma^2 g(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}) g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_{ik}) \text{corr}(\epsilon_{ij}, \epsilon_{ik} | \boldsymbol{x}_i),$$

where $\text{corr}(\epsilon_{ij}, \epsilon_{ik})$ is the (j, k) element of the correlation matrix $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$.

In general, using the shorthand notation

$$g_{ij} = g(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{x}_{ij}),$$

we may rewrite (8.20) in terms of the ϵ_{ij} as

$$\text{cov}(u_{ijk}, u_{i\ell p} | \boldsymbol{x}_i) = \sigma^4 g_{ij} g_{ik} g_{i\ell} g_{ip} \{ E(\epsilon_{ij}\epsilon_{ik}\epsilon_{i\ell}\epsilon_{ip} | \boldsymbol{x}_i) - E(\epsilon_{ij}\epsilon_{ik} | \boldsymbol{x}_i) E(\epsilon_{i\ell}\epsilon_{ip} | \boldsymbol{x}_i) \}. \tag{8.21}$$

The representation (8.21) highlights how ***complex*** specification of $\boldsymbol{Z}_i(\beta, \xi)$ is; in particular, some special cases of (8.21) are

$$\text{cov}(u_{ijj}, u_{ijj} | \boldsymbol{x}_i) = \text{var}(u_{ijj} | \boldsymbol{x}_i) = \sigma^4 g_{ij}^4 \text{var}(\epsilon_{ij}^2 | \boldsymbol{x}_i),$$

$$\text{cov}(u_{ijk}, u_{ijk} | \boldsymbol{x}_i) = \text{var}(u_{ijk}) = \sigma^4 g_{ij}^2 g_{ik}^2 [\, E(\epsilon_{ij}^2 \epsilon_{ik}^2 | \boldsymbol{x}_i) - \{ E(\epsilon_{ij}\epsilon_{ik} | \boldsymbol{x}_i) \}^2 \,],$$

$$\text{cov}(u_{ijj}, u_{i\ell\ell} | \boldsymbol{x}_i) = \sigma^2 g_{ij}^2 g_{i\ell}^2 \{ E(\epsilon_{ij}^2 \epsilon_{i\ell}^2 | \boldsymbol{x}_i) - 1 \},$$

$$\text{cov}(u_{ijk}, u_{ijp} | \boldsymbol{x}_i) = \sigma^4 g_{ij}^2 g_{ik} g_{ip} \{ E(\epsilon_{ij}^2 \epsilon_{ik}\epsilon_{ip} | \boldsymbol{x}_i) - E(\epsilon_{ij}\epsilon_{ik} | \boldsymbol{x}_i) E(\epsilon_{ij}\epsilon_{ip} | \boldsymbol{x}_i) \},$$

$$\text{cov}(u_{ijj}, u_{ijp} | \boldsymbol{x}_i) = \sigma^2 g_{ij}^3 g_{ip} \{ E(\epsilon_{ij}^3 \epsilon_{ip} | \boldsymbol{x}_i) - E(\epsilon_{ij}\epsilon_{ip} | \boldsymbol{x}_i) \}.$$

Of course, (8.21) represents the general case for $j \neq k \neq \ell \neq p$.

The result is that, to specify the "***covariance matrix***" $\mathcal{V}_i(\eta) = \boldsymbol{Z}_i(\beta, \xi)$ in (8.18), we must be prepared to make assumptions about ***numerous higher moments*** involving the elements of $\boldsymbol{Y}_i$, or equivalently, $\boldsymbol{\epsilon}_i$, up to ***four-way associations***, i.e., $E(\epsilon_{ij}\epsilon_{ik}\epsilon_{i\ell}\epsilon_{ip} | \boldsymbol{x}_i)$.

The diligent student will verify that, if $n_i = 1$, so that $\boldsymbol{Y}_i$ is a ***scalar***, this reduces to ***quadratic estimating equation*** (7.21) for θ in the univariate case, where all we have are ***variance parameters***.

Putting aside the troublesome issue of specifying $\boldsymbol{Z}_i(\beta, \xi)$, the preceding developments suggest the following approach. Given some assumption on $\boldsymbol{Z}_i(\beta, \xi)$ (thus, some assumption on ***higher moments*** of the ϵ_{ij}), to estimate ξ (treating β fixed for now), one would solve an ***estimating equation*** of the form

$$\sum_{i=1}^{m} \boldsymbol{E}_i^T(\beta, \xi) \boldsymbol{Z}_i^{-1}(\beta, \xi) \{ \boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta, \xi) \} = \boldsymbol{0}. \tag{8.22}$$

- The estimating equation (8.22) is clearly ***unbiased*** , ***regardless*** of the choice of $\boldsymbol{Z}_i(\beta, \xi)$.
- The discussion at the end of Chapter 7 suggests that the ***optimal estimating equation*** for ξ of form (8.22) is that found by specifying $\boldsymbol{Z}_i(\beta, \xi)$ ***correctly*** , so that, in truth, $\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{Z}_i(\beta, \xi)$.
- In step 1 of the iterative algorithm, at the ℓth iteration, ξ could be estimated by replacing β in (8.22) by $\widehat{\beta}^{(\ell)}$ everywhere, including in $\boldsymbol{u}_i(\beta)$, and solving in ξ.

Two issues must be resolved:

- Clearly, specification of $\boldsymbol{Z}_i(\beta, \xi)$ is ***challenging*** , and the chance one could correctly specify all the relevant moments is ***slim to none***. Thus, a ***realistic strategy*** for specifying $\boldsymbol{Z}_i(\beta, \xi)$ in practice is required.
- Both (8.22) and (8.17) depend on $\boldsymbol{Y}_i$ through the elements of $\boldsymbol{u}_i(\beta)$, the latter equation through the ***quadratic form*** in $\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}$. Because the second equation was derived from the loglikelihood assuming that the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is ***normally distributed*** , intuition suggests that choosing $\boldsymbol{Z}_i(\beta, \xi)$ to be the "covariance matrix" for $\boldsymbol{u}_i(\beta)$ that would be obtained under this condition should lead to an equation (8.22) that is ***equivalent*** to (8.17).

We consider each of these issues in turn.

SPECIFICATION OF $\boldsymbol{Z}_i(\beta, \xi)$: Because correct specification is a considerable challenge, the suggestion is to make a "***working assumption***" for $\boldsymbol{Z}_i(\beta, \xi)$, similar in spirit to that made for the correlation matrix of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ in the ***linear estimating equation*** for β. Popular working assumptions are

- ***Independence working assumption.*** Take $\boldsymbol{Z}_i(\beta, \xi)$ to be the covariance matrix for $\boldsymbol{u}_i|\boldsymbol{x}_i$ that would be obtained if the elements Y_{ij} of $\boldsymbol{Y}_i$ were assumed to be ***mutually independent*** across j.
- ***Gaussian working assumption.*** Take $\boldsymbol{Z}_i(\beta, \xi)$ to be the covariance matrix for $\boldsymbol{u}_i|\boldsymbol{x}_i$ that would be obtained by assuming that the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is ***normal*** with the first two moments ***correctly specified*** according to the assumed mean-covariance model (8.3). Equivalently, assume that $\epsilon_i|\boldsymbol{x}_i$ is normal with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$.

It can be shown under this condition that

$$\text{cov}(u_{ijk}, u_{i\ell p}|\boldsymbol{x}_i) = v_{ij\ell}v_{ikp} + v_{ijp}v_{ik\ell} = \sigma^4 g_{ij}g_{ik}g_{i\ell}g_{ip}\{E(\epsilon_{ij}\epsilon_{i\ell}|\boldsymbol{x}_i)E(\epsilon_{ik}\epsilon_{ip}|\boldsymbol{x}_i) + E(\epsilon_{ij}\epsilon_{ip}|\boldsymbol{x}_i)E(\epsilon_{ik}\epsilon_{i\ell}|\boldsymbol{x}_i)\}. \tag{8.23}$$

The entries of $\boldsymbol{Z}_i(\beta, \xi)$ can then be determined from the ***simplified*** relationship (8.23).

Note that, as $E(\epsilon_{ij}\epsilon_{ik}|\boldsymbol{x}_i)$ is equal to the conditional correlation between ϵ_{ij} and ϵ_{ik}, from (8.23) all of the needed entries of $\boldsymbol{Z}_i(\beta, \xi)$ depend only on the assumed correlation model $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$. Moreover, (8.23) also yields

$$\text{var}(u_{ijj}|\boldsymbol{x}_i) = 2\sigma^4 g_{ij}^4,$$

where the "2" agrees with the fourth moment properties of the normal.

Although these choices may indeed be misspecifications, the hope is that they will produce estimators "***closer to***" being "***optimal***" than simply ***ignoring*** the pattern of association among the elements of $\boldsymbol{u}_i$ altogether.

RELATIONSHIP BETWEEN (8.22) ***AND*** (8.17): The estimating equation in (8.17), derived ***directly*** from the assumed normal loglikelihood for $\boldsymbol{Y}_i|\boldsymbol{x}_i$, namely,

$$(1/2)\sum_{i=1}^{m}\Big(\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}^T \boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\{\partial/\partial\xi_k\, \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}$$
$$- \text{tr}\left[\boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)\}\right]\Big) = 0, \quad k = 1, \ldots, r+s, \tag{8.24}$$

is quadratic, as it depends on a quadratic form in $\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}$. A quadratic form may of course be written as a ***linear combination*** of squared deviations and crossproducts; e.g., for square matrix $\boldsymbol{A}_i$,

$$\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\}^T \boldsymbol{A}_i^{-1}\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\} = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\{Y_{ij} - f(\boldsymbol{x}_{ij}, \beta)\}\{Y_{ik} - f(\boldsymbol{x}_{ik}, \beta)\}a^{ijk},$$

where a^{ijk} is the (j, k) element of $\boldsymbol{A}_i^{-1}$. The quadratic form in the summand of (8.24) is thus a ***linear combination*** of the elements of $\boldsymbol{u}_i(\beta)$.

- Of course, the summands of estimating equation in (8.22),

$$\sum_{i=1}^{m}\boldsymbol{E}_i^T(\beta, \xi)\boldsymbol{Z}_i^{-1}(\beta, \xi)\{\boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta, \xi)\} = \boldsymbol{0}, \tag{8.25}$$

  are ***also linear combinations*** of the elements of $\boldsymbol{u}_i(\beta)$.

- If the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ ***really is*** normal, and we choose $\boldsymbol{Z}_i(\beta, \xi)$ according to the ***Gaussian working assumption***, then, as discussed at the end of Chapter 7, (8.25) must be the (asymptotically) "***optimal***" estimating equation of this (quadratic) form, as the "weight matrix" $\boldsymbol{Z}_i^{-1}(\beta, \xi)$ is correctly specified.

- The quadratic equation (8.24) is also the ***normal theory ML estimating equation*** for ξ. Thus, it is also (asymptotically) "optimal" if the data ***really are*** normally distributed.
- Both equations ***cannot*** be the "optimal" quadratic equation and be different; thus, intuition suggests that estimating equations (8.24) and (8.25) must be ***the same***.

It is possible to show this equivalence analytically; the argument is carried out in detail in Section 8.9.

STACKED EQUATIONS: Using a (quadratic) equation of the form (8.25) to estimate ξ, along with the linear equation for β, it is clear that the iterative two-step scheme given earlier solves the $p + r + s$-dimensional system of equations

$$\sum_{i=1}^{m} \begin{pmatrix} \boldsymbol{X}_i^T(\beta) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{E}_i^T(\beta,\xi) \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_i(\beta,\xi) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta) \\ \boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta,\xi) \end{pmatrix} = \mathbf{0}. \tag{8.26}$$

Compare these equations to the univariate equations (7.25) (which can be seen to incorporate the ***univariate version*** of the "Gaussian working assumption" to yield the "$2\sigma^4 g_j^4$" term). It is straightforward to show that (verify), with $n_i = 1$ for all i and the Gaussian working assumption, (8.26) reduces to (7.25).

The equations (8.26) can be written as

$$\sum_{i=1}^{m} \mathcal{D}_i^T(\eta)\mathcal{V}_i^{-1}(\eta)\{\boldsymbol{s}_i(\eta) - \boldsymbol{m}_i(\eta)\} = \mathbf{0}, \tag{8.27}$$

where

$$\mathcal{D}_i(\eta) = \begin{pmatrix} \boldsymbol{X}_i(\beta) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{E}_i(\beta,\xi) \end{pmatrix} \quad \{n_i + n_i(n_i+1)/2 \times p + r + s\},$$

$$\mathcal{V}_i(\eta) = \begin{pmatrix} \boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_i(\beta,\xi) \end{pmatrix} \quad \{n_i + n_i(n_i+1)/2 \times n_i + n_i(n_i+1)/2\},$$

$$\boldsymbol{s}_i(\eta) = \begin{pmatrix} \boldsymbol{Y}_i \\ \boldsymbol{u}_i(\beta) \end{pmatrix}, \quad \boldsymbol{m}_i(\eta) = \begin{pmatrix} \boldsymbol{f}_i(\boldsymbol{x}_i,\beta) \\ \boldsymbol{v}_i(\beta,\xi) \end{pmatrix} \quad \{n_i + n_i(n_i+1)/2 \times 1\}.$$

With the equations written in the form (8.27), it is clear that the two-step algorithm for solving them is natural, and that each of steps 1 and 2 can be implemented in principle via a ***Gauss-Newton*** updating scheme, which is common in practice. In particular

- Step 1, with β held fixed [last $r + s$ rows of (8.27)], can be implemented by a Gauss-Newton algorithm iterated to convergence to obtain the next iterate of ξ.
- Step 2, with ξ held fixed [first p rows of (8.27)], can be implemented by a Gauss-Newton algorithm iterated to convergence to obtain the next iterate of β.

8.4 Quadratic estimating equations

As in the univariate case, it is natural in the longitudinal context to consider the potential ***increase efficiency*** for estimation of β by extracting information about β from the covariance matrix $\boldsymbol{V}_i(\beta, \xi)$. As in that case, we can deduce a ***quadratic estimating equation*** for β by appealing to the ***normal loglikelihood***.

Differentiating the normal loglikelihood (8.14) with respect to β, it is straightforward using the same matrix differentiation operations as before to obtain the resulting equation

$$\begin{aligned}\sum_{i=1}^{m} \Big\{ & \boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} \\ & + \Big(\Big(\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T \boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\partial/\partial\beta_\ell \boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\} \\ & - \text{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\partial/\partial\beta_\ell\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\right]\Big)\Big) \Big\} = \boldsymbol{0} \quad (p\times 1), \end{aligned} \tag{8.28}$$

where the double parentheses indicate p terms of the form inside them stacked for $\ell = 1, \ldots, p$. Of course, underlying estimating equation (8.28) is the assumption of normality. This equation would be solved ***jointly*** with the quadratic equation (8.24) for ξ to obtain the MLEs for β and ξ under the assumption of normality.

Alternatively, following the ***same reasoning*** as in the previous section for estimation of ξ, we can formulate a general quadratic equation. Defining $\boldsymbol{v}_i(\beta,\xi)$ and $\boldsymbol{u}_i(\beta)$ as before, and letting

$$\boldsymbol{B}_i(\beta,\xi) = \partial/\partial\beta\, \boldsymbol{v}_i(\beta,\xi) \quad \{n_i(n_i+1)/2 \times p\},$$

so that $\boldsymbol{B}_i(\beta,\xi)$ is the "gradient matrix" of $E\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{v}_i(\beta,\xi)$, the obvious joint estimating equations for β and ξ are

$$\sum_{i=1}^{m} \begin{pmatrix} \boldsymbol{X}_i^T(\beta) & \boldsymbol{B}_i^T(\beta,\xi) \\ \boldsymbol{0} & \boldsymbol{E}_i^T(\beta,\xi) \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_i(\beta,\xi) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i,\beta) \\ \boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta,\xi) \end{pmatrix} = \boldsymbol{0}. \tag{8.29}$$

- Because of the presence of the "gradient matrix" $\boldsymbol{B}_i(\beta,\xi)$ in the first p rows of (8.29), this leads to an estimating equation for β that is ***quadratic***.
- If $\boldsymbol{Z}_i(\beta,\xi)$ were chosen according to the ***Gaussian working assumption***, then, by the same reasoning as in the previous section, the equation corresponding to the first p rows of (8.29) should be the "optimal" such equation if ***normality really holds***, and, thus, intuitively, should be ***identical*** to the normal theory ML equation (8.28). This can be shown by arguments analogous to those in Section 8.9.

In (8.28), the ***off-block-diagonal elements*** of the "covariance matrix"

$$\begin{pmatrix} \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_i(\beta, \xi) \end{pmatrix}$$

are all equal to zero. ***Under normality*** of $\boldsymbol{Y}_i|\boldsymbol{x}_i$, it is indeed the case that

$$\text{cov}\{\boldsymbol{Y}_i, \boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{0}.,$$

which is consistent with the view that (8.29) with the ***Gaussian working assumption*** is the optimal joint equation if ***normality really holds***, as in this case $\mathcal{V}_i(\beta, \xi)$ is exactly the covariance matrix of the "response" $\boldsymbol{s}_i(\eta) = \{\boldsymbol{Y}_i^T, \boldsymbol{u}_i^T(\beta)\}^T$.

By analogy to the univariate case discussed at the end of Chapter 7, if the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is not believed to be normal, we could ***in principle*** arrive at a more general equation by specifying the covariance matrix of the "response" $\boldsymbol{s}_i(\eta) = \{\boldsymbol{Y}_i^T, \boldsymbol{u}_i^T(\beta)\}^T$ to embody corresponding assumptions about the moments of $\boldsymbol{s}_i(\eta)$. If we specify

$$\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{Z}_i(\beta, \xi) \quad \text{and} \quad \text{cov}\{\boldsymbol{Y}_i, \boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = \boldsymbol{C}_i(\beta, \xi),$$

say, we obtain

$$\sum_{i=1}^{m} \begin{pmatrix} \boldsymbol{X}_i^T(\beta) & \boldsymbol{B}_i^T(\beta, \xi) \\ \boldsymbol{0} & \boldsymbol{E}_i^T(\beta, \xi) \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) & \boldsymbol{C}_i(\beta, \xi) \\ \boldsymbol{C}_i^T(\beta, \xi) & \boldsymbol{Z}_i(\beta, \xi) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta) \\ \boldsymbol{u}_i(\beta) - \boldsymbol{v}_i(\beta, \xi) \end{pmatrix} = \boldsymbol{0}. \tag{8.30}$$

- In the ***unlikely*** event that the assumptions for $\boldsymbol{Z}_i(\beta, \xi)$ and $\boldsymbol{C}_i(\beta, \xi)$ are ***correct***, we would expect to have constructed the "***optimal***" such quadratic equation.

- This entails making the moment assumptions on $\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\}$ in $\boldsymbol{Z}_i(\beta, \boldsymbol{x}_i)$, but ***also*** for $\boldsymbol{C}_i(\beta, \xi)$, which involves specifying quantities of the form

  $$\text{cov}(Y_{ij}, u_{ik\ell}|\boldsymbol{x}_i) = \sigma^3 g_{ij} g_{ik} g_{i\ell} E(\epsilon_{ij}\epsilon_{ik}\epsilon_{i\ell}|\boldsymbol{x}_i).$$

  That is, we need to be willing to specify ***not only*** the ***skewness*** $E(\epsilon_{ij}^3|\boldsymbol{x}_i)$ as in the univariate case, but ***also*** the "three-way associations."

- The chance that we can specify the matrices $\boldsymbol{Z}_i(\beta, \xi)$ and $\boldsymbol{C}_i(\beta, \xi)$ ***completely correctly*** in practice is ***slim***. Indeed, specifying $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ correctly in itself is ***difficult enough***.

Putting this issue aside for the moment, clearly, solution of estimating equations of the general form (8.30) can be ***implemented*** by a Gauss-Newton updating scheme, redefining $\mathcal{D}_i(\eta)$ and $\mathcal{V}_i(\eta)$ in (8.27) in the obvious way as

$$\mathcal{V}_i(\eta) = \begin{pmatrix} \boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) & \boldsymbol{C}_i(\beta, \xi) \\ \boldsymbol{C}_i^T(\beta, \xi) & \boldsymbol{Z}_i(\beta, \xi) \end{pmatrix}. \tag{8.31}$$

This must be carried out via a ***single*** such updating algorithm of dimension $p+r+s$, as it is no longer possible to ***decouple*** the equations for β (first p rows) and ξ (last $r+s$ rows) and to use separate, lower-dimensional updating. Thus, solution of (8.30) by this approach is ***more complex*** numerically.

WORKING ASSUMPTIONS: As above, in practice, one would make a "***working assumption***" about the entire matrix $\mathcal{V}_i(\eta)$ in (8.31). This of course involves an assumption on $\boldsymbol{Z}_i(\beta, \xi)$ as before, along with one on $\boldsymbol{C}_i(\beta, \xi)$. Popular working assumptions for $\mathcal{V}_i(\eta)$ are as follows, analogous to the previous discussion.

- ***Independence working assumption.*** Taking the elements of $\boldsymbol{Y}_i$ to be ***mutually independent*** leads to the choice for $\boldsymbol{Z}_i(\beta, \xi)$ discussed previously and $\boldsymbol{C}_i(\beta, \xi) = \boldsymbol{0}$.
- ***Gaussian working assumption.*** Taking of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ to be normal leads to the choice for $\boldsymbol{Z}_i(\beta, \xi)$ given in (8.23) and $\boldsymbol{C}_i(\beta, \xi) = \boldsymbol{0}$.

TERMINOLOGY: Equations like those in (8.26) and (8.30) have been referred to as ***generalized estimating equations*** of specific types:

- Equations of the form (8.26), which involve solving a ***linear*** estimating equation for β jointly with a quadratic one for ξ, have been called ***GEE-1***.
- Equations of the form (8.30), which involve solving a ***quadratic*** estimating equation for β jointly with a quadratic one for ξ, have been called ***GEE-2***.
- This terminology was evidently coined in a paper by Liang, Zeger, and Qaqish (1992).
- The unqualified acronym "GEE" is used popularly both to refer to the general approach of specifying estimating equations for mean-covariance models of the form (8.3) ***and*** to the particular case where the ***linear*** equation for β in (8.11) is solved with the elements of ξ estimated by simple ***moment-based*** estimators.

- "GEE" is often also taken to imply that "***working assumptions***" are involved, including those on the correlation matrix of $\boldsymbol{Y}_i|\boldsymbol{x}_i$, that are likely to be ***incorrect*** , so that the ***robust sandwich covariance matrix*** , discussed in the next section, should be used by default when approximating the sampling distribution of the estimator for β.

REMARKS: The ***same*** issues involving ***trade-offs*** between linear and quadratic estimating equations for β that we discussed for the univariate case extend to the longitudinal, multivariate setting.

- The linear equation is clearly unbiased ***even if*** the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***misspecified***. Such misspecification is ***more likely*** in the multivariate case, as the analyst must model ***not only*** variance but also correlation structure. The latter is more difficult, so the focus is on possible ***incorrect modeling*** of the correlation matrix $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$.
- The quadratic equation for β will be unbiased ***as long as*** the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is correctly specified. Thus, one must be confident that this is the case to reap the benefits of ***possible increased efficiency***. Even in this case, the ***optimal*** quadratic equation ***also*** requires correct specification of $\boldsymbol{Z}_i(\beta, \xi)$ and $\boldsymbol{C}_i(\beta, \xi)$. This is almost certainly ***not the case*** in practice, so even though the estimating equation is unbiased with $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ correct, whether or not it leads to a more efficient estimator for β is ***no longer clear*** , analogous to the results at the end of Chapter 7.
- The quadratic equations will ***not be unbiased*** if $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is not correctly specified so could result in ***inconsistent*** estimation of β. Thus, in practice, it is generally agreed that "GEE-1" estimation based on (8.26) is the "safer" choice for routine use in fitting population-averaged models. Accordingly, we discuss large sample inference ***only*** for the estimator for β obtained by solving the ***linear estimating equation***.

8.5 Large sample inference

As we did for the ***linear population-averaged models*** discussed in Chapter 5, we derive an approximate sampling distribution for $\widehat{\beta}$ solving (8.11), ***linear estimating equation***

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta)\boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\} = \boldsymbol{0} \tag{8.32}$$

with an estimator $\widehat{\xi}$ ***substituted***.

The estimator $\widehat{\xi}$ can be a ***moment-based*** estimator or that found by solving the ***quadratic estimating equation*** for ξ under some ***working assumption***, all of which satisfy the conditions below.

As in those arguments, we adopt a large sample framework in which the number of individuals $m \to \infty$ with the n_i treated as ***fixed***. As we noted in Section 5.6, it is ***not appropriate*** to regard the n_i as fixed when data vectors of intended length n are of different lengths as the result of some ***missingness mechanism***. We discuss the implications of ***missing data*** for inference using GEEs in Section 8.7.

COVARIANCE MODEL POSSIBLY INCORRECTLY SPECIFIED: As in Section 5.5 for a linear PA model with covariance matrix not depending on β and in the univariate setting in Section 7.4, consider the general situation in which, although the mean model $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ is ***correctly specified*** , so that the ***true mean*** is

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0),$$

where β_0 is the true value of β, the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ ***need not be*** correctly specified. Thus, analogous to these previous arguments, letting the ***true covariance matrix*** be

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_{i0},$$

there is ***not necessarily*** a value ξ_0 such that $\boldsymbol{V}_i(\beta_0, \xi_0, \boldsymbol{x}_i)$.

Assume that the estimator $\widehat{\xi}$ is such that $\widehat{\xi} \xrightarrow{p} \xi^*$ for some ξ^* and $m^{1/2}(\widehat{\xi} - \xi^*) = O_p(1)$, and let

$$\boldsymbol{V}_i^* = \boldsymbol{V}_i(\beta_0, \xi^*, \boldsymbol{x}_i).$$

As in Section 5.5, even with the ***covariance model misspecified*** , (8.32) is still an ***unbiased estimating equation*** , as clearly

$$E\Big[\boldsymbol{X}_i^T(\beta_0)\boldsymbol{V}_i^{-1}(\beta_0, \xi^*, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\}|\boldsymbol{x}_i\Big] = \boldsymbol{0}.$$

The argument we now present is a ***generalization*** of those in Sections 5.5 and 7.4. Expanding (8.32) in a Taylor series in $(\widehat{\beta}^T, \widehat{\xi}^T)^T$ about $(\beta_0^T, \xi^{*T})^T$, we have

$$\begin{aligned} \boldsymbol{0} &= m^{-1/2}\sum_{i=1}^{m} \boldsymbol{X}_i^T(\widehat{\beta})\boldsymbol{V}_i^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \widehat{\beta})\} \\ &\approx \boldsymbol{C}_m^* + (\boldsymbol{A}_{m1}^* + \boldsymbol{A}_{m2}^* + \boldsymbol{A}_{m3}^*)\, m^{1/2}(\widehat{\beta} - \beta_0) + \boldsymbol{E}_m^*\, m^{1/2}(\widehat{\xi} - \xi^*). \end{aligned} \tag{8.33}$$

Let $\boldsymbol{X}_i = \boldsymbol{X}_i(\beta_0)$, and, as in (8.12), $\boldsymbol{X} = \boldsymbol{X}(\beta_0)$. Also define

$$\boldsymbol{V}_0 = \text{block diag}(\boldsymbol{V}_{01}, \ldots, \boldsymbol{V}_{0m}), \qquad \boldsymbol{V}^* = \text{block diag}(\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_m^*).$$

It is straightforward that

$$\boldsymbol{C}_m = m^{-1/2} \sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta_0) \boldsymbol{V}_i^{*-1} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\} \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}^*),$$

$$\boldsymbol{B}^* = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*-1} \boldsymbol{V}_{0i} \boldsymbol{V}^{*-1} \boldsymbol{X}_i = \lim_{m\to\infty} m^{-1} \boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{V}_0 \boldsymbol{V}^{*-1} \boldsymbol{X},$$

$$\boldsymbol{A}_{m2}^* = -m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta_0) \boldsymbol{V}_i^{*-1} \boldsymbol{X}_i(\beta_0) \xrightarrow{p} \boldsymbol{A}^* = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*-1} \boldsymbol{X}_i = \lim_{m\to\infty} m^{-1} \boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{X},$$

$$\boldsymbol{A}_{m1}^* = m^{-1} \sum_{i=1}^{m} \{\partial/\partial\beta \, \boldsymbol{X}_i^T(\beta_0)\} \boldsymbol{V}_i^{*-1} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\} \xrightarrow{p} \boldsymbol{0},$$

$$\boldsymbol{A}_{m3}^* = m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta_0) \{\partial/\partial\beta \boldsymbol{V}_i^{-1}(\beta_0, \xi^*, \boldsymbol{x}_i)\} \boldsymbol{V}_i^{*-1} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\} \xrightarrow{p} \boldsymbol{0},$$

$$\boldsymbol{E}_m^* = m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta_0) \{\partial/\partial\xi \boldsymbol{V}_i^{-1}(\beta_0, \xi^*, \boldsymbol{x}_i)\} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta_0)\} \xrightarrow{p} \boldsymbol{0}.$$

Combining, we obtain the (***not surprising***) result

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{*-1} \boldsymbol{B}^* \boldsymbol{A}^{*-1}), \tag{8.34}$$

which is ***identical*** in form to the result (5.74) in the linear PA model case, but with $\boldsymbol{X}$ and $\boldsymbol{V}^{*-1}$ depending on β_0.

- From (8.33), because $\boldsymbol{A}_{m3}^*$ and $\boldsymbol{E}_m^* \xrightarrow{p} \boldsymbol{0}$, there is ***no effect*** of estimating ξ in the incorrect model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$, ***nor*** is there an effect of the fact that this model ***depends on*** β, which must be estimated to form "weights."

- As before, when the model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***correctly specified***, $\xi^* = \xi_0$, the value such that $\boldsymbol{V}_{0i} = \boldsymbol{V}_i(\beta_0, \xi_0, \boldsymbol{x}_i)$, and $\boldsymbol{V}_i^* = \boldsymbol{V}_{0i}$, so that (8.34) reduces to

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}), \quad \boldsymbol{A} = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_{0i}^{-1} \boldsymbol{X}_i = \lim_{m\to\infty} m^{-1} \boldsymbol{X}^T \boldsymbol{V}_0^{-1} \boldsymbol{X}. \tag{8.35}$$

OPTIMAL LINEAR ESTIMATING EQUATION: Analogous to (5.75) and (5.76), (8.35) and (8.34) yield the approximate sampling distributions

$$\widehat{\beta}_C \dot{\sim} \mathcal{N}\{\beta_0, (\boldsymbol{X}^T \boldsymbol{V}_0^{-1} \boldsymbol{X})^{-1}\} \tag{8.36}$$

when $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***correctly specified***, where *C* indicates "correct;" and, when $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ is ***incorrectly specified***,

$$\widehat{\beta}_{IC} \dot{\sim} \mathcal{N}\{\beta_0, (\boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{V}_0 \boldsymbol{V}^{*-1} \boldsymbol{X}) (\boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{X})^{-1}\}, \tag{8.37}$$

where *IC* indicates "incorrect."

- The comparison of the approximate covariance matrices in (8.36) and (8.37) is ***identical*** to that in Chapter 5. Thus, we conclude immediately that, for "large" m, the components of $\widehat{\boldsymbol{\beta}}_{IC}$ are ***inefficient relative to*** the corresponding components of $\widehat{\boldsymbol{\beta}}_{C}$.

- It follows that using a ***correct covariance model*** leads to the ***optimal linear estimating equation*** of this type, extending the result in Chapter 5 to ***nonlinear models*** and covariance models that ***depend on*** $\boldsymbol{\beta}$.

- In fact, it is possible to obtain an ***even more general*** result. Consider the linear estimating equation of the form

$$\sum_{i=1}^{m} \boldsymbol{\mathcal{A}}_i^T(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{0}, \tag{8.38}$$

  where $\boldsymbol{\mathcal{A}}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, are arbitrary $(n_i \times p)$ matrices depending on $\boldsymbol{\beta}$, some additional parameter $\boldsymbol{\gamma}$, and possibly $\boldsymbol{x}_i$. Clearly, (8.38) is an ***unbiased estimating equation*** for $\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{\beta}}_G \xrightarrow{p} \boldsymbol{\beta}_0$ be the solution to (8.38), where $\boldsymbol{\gamma}$ has been replaced by some $\widehat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*$ that is bounded in probability. It is straightforward, expanding (8.38) evaluated at $(\widehat{\boldsymbol{\beta}}_G^T, \widehat{\boldsymbol{\gamma}}^T)^T$ about $(\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}^{*T})^T$ and letting $\boldsymbol{\mathcal{A}}_i = \boldsymbol{\mathcal{A}}_i(\boldsymbol{\beta}_0, \boldsymbol{\gamma}^*, \boldsymbol{x}_i)$ and $\boldsymbol{\mathcal{A}} = (\boldsymbol{\mathcal{A}}_1^T, \ldots, \boldsymbol{\mathcal{A}}_m^T)^T$ $(N \times p)$, to deduce that

$$\widehat{\boldsymbol{\beta}}_G \dot{\sim} \mathcal{N}\left\{\boldsymbol{\beta}_0, (\boldsymbol{\mathcal{A}}^T\boldsymbol{X})^{-1}(\boldsymbol{\mathcal{A}}^T\boldsymbol{V}_0\boldsymbol{\mathcal{A}})(\boldsymbol{X}^T\boldsymbol{\mathcal{A}})^{-1}\right\}. \tag{8.39}$$

  By an argument similar to that in Section 5.5, it can be shown that the matrix difference

$$(\boldsymbol{\mathcal{A}}^T\boldsymbol{X})^{-1}(\boldsymbol{\mathcal{A}}^T\boldsymbol{V}_0\boldsymbol{\mathcal{A}})(\boldsymbol{X}^T\boldsymbol{\mathcal{A}})^{-1} - (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}$$

  is ***nonnegative definite*** (try it).

- It follows from (8.36), (8.39), and this result that, among ***all linear estimating equations*** for $\boldsymbol{\beta}$ of ***arbitrary form*** (8.38), the equation (8.11) with covariance model ***correctly specified*** is ***optimal***. This is an ***even more general*** result than that above.

- This result can be derived formally from a ***geometric perspective*** by appealing to ***semiparametric theory*** ; the gory details are presented in Chapter 4 of Tsiatis (2006).

- In fact, this result suggests that general estimating equations of the form (8.27), namely,

$$\sum_{i=1}^{m} \boldsymbol{\mathcal{D}}_i^T(\boldsymbol{\eta})\boldsymbol{\mathcal{V}}_i^{-1}(\boldsymbol{\eta})\{\boldsymbol{s}_i(\boldsymbol{\eta}) - \boldsymbol{m}_i(\boldsymbol{\eta})\} = \boldsymbol{0},$$

  with the "covariance matrix" $\boldsymbol{\mathcal{V}}_i^{-1}(\boldsymbol{\eta})$ ***correctly specified*** are ***optimal*** among all equations linear in $\{\boldsymbol{s}_i(\boldsymbol{\eta}) - \boldsymbol{m}_i(\boldsymbol{\eta})\}$.

ROBUST COVARIANCE MATRIX: As we have discussed, because the covariance model $\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i)$ and in particular the correlation model $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ are likely to be ***misspecified***, the latter being a "***working model***," it is standard to use (8.37) as the basis for the approximate sampling distribution for the estimator $\widehat{\beta}$.

In particular, analogous to the argument in Section 5.5,

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_R), \qquad \widehat{\boldsymbol{\Sigma}}_R = \widehat{\boldsymbol{A}}_m^{*-1} \widehat{\boldsymbol{B}}_m^* \widehat{\boldsymbol{A}}_m^{*-1} \tag{8.40}$$

$$\widehat{\boldsymbol{A}}_m^* = \sum_{i=1}^{m} \boldsymbol{X}_i^T(\widehat{\beta}) \boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i) \boldsymbol{X}_i(\widehat{\beta}),$$

$$\widehat{\boldsymbol{B}}_m^* = \sum_{i=1}^{m} \boldsymbol{X}_i^T(\widehat{\beta}) \boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i) \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \widehat{\beta}\} \{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \widehat{\beta}\}^T \boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i) \boldsymbol{X}_i(\widehat{\beta}).$$

In (8.40), $\widehat{\boldsymbol{\Sigma}}_R$ is the ***robust sandwich*** or ***empirical*** covariance matrix, and it is straightforward to demonstrate that $m^{-1}\widehat{\boldsymbol{\Sigma}}_R$ is a consistent estimator for the true sampling covariance matrix in (8.34).

Under the assumption that the covariance model is ***correctly specified***, one would instead use (8.36) as the basis for the approximate sampling distribution of $\widehat{\beta}$, namely,

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_M), \qquad \widehat{\boldsymbol{\Sigma}}_M = \left\{ \sum_{i=1}^{m} \boldsymbol{X}_i^T(\widehat{\beta}) \boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\xi}, \boldsymbol{x}_i) \boldsymbol{X}_i(\widehat{\beta}) \right\}^{-1} = \widehat{\boldsymbol{A}}_m^{*-1}, \tag{8.41}$$

where $\widehat{\boldsymbol{\Sigma}}_M$ is the so-called ***model-based*** covariance matrix.

- Not surprisingly, software for solving GEEs typically uses the robust covariance matrix in (8.40) ***by default***, and the user must request explicitly the model-based analysis.

8.6 Modeling issues

As the preceding sections demonstrate, the mechanics of specifying and fitting a general population-averaged mean-covariance model of the form (8.3) and carrying out approximate large-sample inference for β defining the assumed mean response model are ***messy but straightforward***. ***However***, there are ***more abstract issues*** that can render inferences and practical interpretation suspect. In this and the next section, we discuss these in some detail.

TIME-DEPENDENT AMONG-INDIVIDUAL COVARIATES: As we indicated in Section 8.2, specification of a sensible mean model when some covariates change over time can involve critical ***conceptual challenges***. We now take a more formal look at this issue.

Recall from our initial discussion of the basic mean model that, ***ordinarily*** , the model $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ satisfies

$$\boldsymbol{f}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f(\boldsymbol{x}_{in_i}, \beta) \end{pmatrix} \quad (n_i \times 1). \tag{8.42}$$

As in (8.8), (8.42) ***implicitly incorporates*** a rather ***strong assumption*** , namely, that

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}) = E(Y_{ij}|\boldsymbol{x}_{ij}). \tag{8.43}$$

We noted earlier that, when the $\boldsymbol{x}_{ij}$ are ***time-independent*** , or when they are time-dependent but ***fixed by design*** or depend on j only through t_{ij}, there is no conceptual difficulty with (8.43) because the values of the $\boldsymbol{x}_{ij}$ are determined in a way that is ***unrelated*** to the longitudinal response. In contrast, if the $\boldsymbol{x}_{ij}$ are ***observed*** and not under the control of the investigator, there is the possibility that their values are ***impacted*** by those of the longitudinal response in ways that can ***distort*** the relationship between mean response and covariates and lead to difficulties in ***interpretation***.

This can be formalized as follows.

CONVENTION: For this discussion, assume that there are n intended observation times t_j, $j = 1, \ldots, n$, for each of which $\boldsymbol{x}_{ij}$ comprises ***time-independent*** covariates, such as gender, age at study entry, and so on, that ***do not vary*** with time, along with other ***time-dependent*** covariates that ***do vary*** over the observation times, such as dose in a designed experiment or smoking status in the Six Cities study. We do not make this explicit in the notation; however, it is worth noting that it is implicit in the following developments that the presence of the time-independent covariates in each of $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}$ means that all expressions are ***conditional*** on the time-independent covariates.

Recall from Chapter 2 that we conceptualize that each individual i has an associated ***stochastic response process*** $\mathcal{Y}_i(t)$ in ***continuous time*** , where we suppress dependence on within-individual covariates $\boldsymbol{u}_i$ for brevity. With time-dependent covariates, we can also conceptualize an individual ***covariate process*** $\boldsymbol{x}_i(t)$, say.

As in previous chapters, we assume that time-dependent covariates are observed ***without measurement error*** , so that in principle it is possible to observe $\boldsymbol{x}_i(t)$ at any t ***without error***.

For the following developments, we adopt the ***convention*** that, for observation time j, Y_{ij} is the response ascertained at t_j, and we let $\boldsymbol{x}_{ij}$ involve values of $\boldsymbol{x}_i(t)$ at $t < t_j$, so up to a time ***immediately prior to*** t_j. That, is we adopt the convention that the ***temporal ordering*** of the data is

$$\boldsymbol{x}_{i1}, Y_{i1}, \boldsymbol{x}_{i2}, Y_{i2}, \ldots, \boldsymbol{x}_{i,n-1}, Y_{i,n-1}, \boldsymbol{x}_{in}, Y_{in}. \tag{8.44}$$

- In the dosing study above, where the doses d_j are ***fixed by design***, this convention coincides with the expectation that the response at t_j is ascertained ***after*** the dose d_j is administered.

- In the Six Cities study, where $\boldsymbol{x}_{ij}$ includes mother i's smoking status s_{ij} at age t_j, smoking status s_{ij} is naturally regarded as being ***already established*** at age t_j and thus ***preceding*** the child's wheezing response at t_j.

- In studies where t_1 corresponds to ***baseline***, (8.44) implies that $\boldsymbol{x}_{i1}$ comprises covariates whose values are ascertained ***immediately prior*** to the first measure of the response, Y_{i1}, including the ***randomized treatment***. Previously, in a ***randomized study***, we have used the index $j = 1$ to refer to baseline and taken Y_{i1} to be the response ascertained ***prior to initiation treatment***.

  For purposes of the discussion here, we adhere to the temporal ordering (8.44), so that, with randomized treatment included in $\boldsymbol{x}_{i1}$, Y_{i1} represents the first measure of the response ***after*** initiation of treatment. If a value of the response is ascertained ***prior to treatment***, we take this to be at the same time $\boldsymbol{x}_{i1}$ is recorded and can include this in $\boldsymbol{x}_{i1}$ in the developments below.

  In an ***observational longitudinal study*** such as the Six Cities study, the temporal ordering (8.44) is natural.

  We make further remarks on handling of such ***baseline responses*** in Section 8.8.

With these conventions, we can formalize the issues raised above.

EXOGENOUS COVARIATE: A covariate process is said to be ***exogenous*** with respect to a response/outcome process if, given all previous values of the covariate and response, the covariate at time j is ***independent*** of all preceding responses. Formally, in our context and obvious notation, for all individuals i and $j = 2, \ldots, n$, an exogenous covariate process satisfies

$$p(\boldsymbol{x}_{ij} | y_{i1}, \ldots, y_{i,j-1}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}) = p(\boldsymbol{x}_{ij} | \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}). \tag{8.45}$$

- Such covariates have also been referred to as ***external*** in the survival analysis literature.

- Practically speaking, (8.45) states that the values taken on by the covariate at time j depend only on previous covariate values and ***not*** on values taken on by the response prior to j.

- Clearly, $\boldsymbol{x}_{ij} = (t_j, d_j)$ in the fixed dose design study is ***exogenous***, as the doses are fixed in advance so are determined ***completely independently*** of responses they might elicit.

- In the Six Cities study, the covariate $\boldsymbol{x}_{ij} = (t_{ij}, c_i, s_{ij})^T$ is obviously ***not exogenous***. As discussed earlier, if a mother who has past smoking history embodied in $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}$ alters her future smoking behavior as a result of observing the current wheezing status of her child, then $\boldsymbol{x}_{ij}$ cannot be independent of $Y_{i,j-1}$. More generally, she might decide to alter her future smoking as a result of knowing her past smoking history and observing part or all of the ***entire past trajectory*** of wheezing of her child, so that $\boldsymbol{x}_{ij}$ is not independent of $Y_{i1}, \ldots, Y_{i,j-1}$.

- A covariate process that is not exogenous is referred to as ***endogenous*** (or ***internal***).

- In some settings, ***care*** must be taken in evaluating whether or not a covariate process is exogenous. For example, in studies of ***environmental health***, Y_{ij} might be some ***health outcome*** for study participant i, and $\boldsymbol{x}_{ij}$ might include ***air pollution levels*** recorded at a monitoring site near to i's residence when the health outcome is measured (and thus assumed to be in place ***prior*** to the observed health outcome). Here, future pollution levels are likely related to past pollution levels but are not impacted by, and are thus ***clearly independent*** of, the previous health outcomes of study participants.

  However, suppose ***instead*** that $\boldsymbol{x}_{ij}$ includes a measure of i's ***personal exposure*** to air pollution, e.g., a summary measure based on the level at the monitoring station and i's time spent outdoors in the past month. If i decides to ***limit*** his/her future personal exposure by spending less time outdoors because of his/her current and past trajectory of health outcomes, then the covariate process is ***not exogenous***.

- In fact, regarding ***time*** itself as a ***covariate***, it should be clear that observation times in a study that are ***predetermined in advance***, such as the visit times in the epileptic seizure study, are ***exogenous***. ***However*** consider an ***observational study***, such as one based on electronic health records. If individuals have different observation times t_{ij} at which responses and other information is recorded, it is plausible that an individual's future ***realized observation times*** might be ***related*** to his/her past responses and observation times, as s/he might require ***more frequent visits*** to his/her healthcare provider if s/he has poorer responses.

- From (8.45), a strategy to check the validity of the ***assumption of exogeneity*** in practice is to develop ***regression models*** for $j = 2, \ldots, n$ for $\boldsymbol{x}_{ij}$ as a function of $Y_{i1}, \ldots, Y_{i,j-1}$ and $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}$. If, given $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,j-1}$, there is no evidence of dependence on $Y_{i1}, \ldots, Y_{i,j-1}$, this would be taken as support for the contention that the covariate process is exogenous.

KEY RESULT: If a covariate process is exogenous, it can be shown that

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ij}), \qquad j = 1, \ldots, n-1, \tag{8.46}$$

where of course (8.46) is trivially true when $j = n$.

This result can be demonstrated formally by repeated application of the exogeneity condition (8.45). In particular, for $j < n$,

$$\begin{aligned} p(y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}) &= \frac{p(y_{ij}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in})}{p(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in})} = \frac{p(\boldsymbol{x}_{in}|y_{ij}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1})p(y_{ij}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1})}{p(\boldsymbol{x}_{in}|\boldsymbol{x}_{i1} \ldots, \boldsymbol{x}_{i,n-1})p(\boldsymbol{x}_{i1} \ldots, \boldsymbol{x}_{i,n-1})} \\ &= \frac{p(\boldsymbol{x}_{in}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1})p(y_{ij}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1})}{p(\boldsymbol{x}_{in}|\boldsymbol{x}_{i1} \ldots, \boldsymbol{x}_{i,n-1})p(\boldsymbol{x}_{i1} \ldots, \boldsymbol{x}_{i,n-1})} = p(y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1}), \end{aligned} \tag{8.47}$$

where the first equality in (8.47) follows because (8.45) implies that Y_{ij} for $j < n$ is independent of $\boldsymbol{x}_{in}$ given $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n-1}$. By applying these same steps to the result in (8.47) repeatedly, (8.46) follows.

Of course, (8.46) is ***not as strong*** a condition as (8.43), that is,

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}) = E(Y_{ij}|\boldsymbol{x}_{ij}). \tag{8.48}$$

However, which of the assumptions (8.48) or (8.45) is required depends on the precise questions of interest.

- If, as in our examples and implicit in the above arguments, $\boldsymbol{x}_{ij}$ is defined as involving ***only*** values of time-dependent covariates recorded at time j, then

$$E(Y_{ij}|\boldsymbol{x}_{ij}) \tag{8.49}$$

  is the ***marginal population mean*** at time j, representing the ***marginal relationship*** between ***current*** response and ***current*** covariate values. In some contexts, investigators might well be interested in simply characterizing the ***association*** between current values of response and covariates, the so-called ***cross-sectional association***, in which case they might posit a marginal model (8.49) ***directly*** as a way of ***empirically representing*** this relationship.

Here, the investigators are ***not interested*** in making a ***causal interpretation***. ***Nonetheless***, as we demonstrate momentarily, unless they are willing to assume (8.48), ***great care*** must be taken in fitting the directly specified model (8.49)

- More often, however, (whether they admit it or not) investigators ***are*** interested in making ***causal interpretations***. From this perspective, when investigators adopt a marginal model, they wish to conclude that the current value of the covariate alone "***causes***" the response. As we have discussed in the context of the Six Cities study, such an interpretation ***cannot be made*** from fitting a marginal model because of the ***confounding*** that is likely present. That is, future smoking status is not independent of current response given past responses and smoking behavior. Thus, for instance, children whose wheezing status is poor might be more likely to be exposed to less smoking in the future than those not suffering from respiratory problems. Clearly, both of the assumptions (8.48) or (8.45) would be suspect for this study.

- In some situations, investigators might be interested in the relationship between ***cumulative exposure*** and response. Consider the air pollution example above, where interest focuses on the relationship between air pollution levels at nearby monitoring sites and health outcomes. Here, it is reasonable to assume that (8.45) holds, but (8.48) as stated is probably not true, as longitudinal health outcomes are likely impacted by the ***cumulative history*** of exposure to pollution.

  However, if we define a new covariate $\boldsymbol{x}_{ij}^* = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{ij}^T)^T$, then under (8.45), (8.48) holds with $\boldsymbol{x}_i^*$ and $\boldsymbol{x}_{ij}^*$ replacing $\boldsymbol{x}_i$ and $\boldsymbol{x}_{ij}$. As we demonstrate momentarily, this will facilitate valid inferences based on fitting a model based on $\boldsymbol{x}_{ij}^*$.

- Of course, the ***critical issue*** is whether or not (8.45) is a realistic assumption. If it is, as in the pollution example, then it is possible to draw ***causal interpretations***.

  If it is ***not***, then it is ***not possible*** to make causal interpretations by simply fitting models of the type discussed in this course. A specialized statistical framework for ***causal inference*** in the presence of ***time-dependent confounding*** is required of the type pioneered by Robins (1994), Robins, Greenland, and Hu (1999), and Robins, Hernán, and Brumback (2000); Vansteelandt and Joffe (2014) provide a comprehensive review and cite numerous references. This is the subject of an entire course.

UNBIASEDNESS OF THE LINEAR ESTIMATING EQUATION, REVISITED: As we noted previously, ***under the assumption*** that the model in (8.4),

$$\boldsymbol{f}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f(\boldsymbol{x}_{in_i}, \beta) \end{pmatrix} \tag{8.50}$$

is ***correctly specified***, the ***linear estimating equation*** (8.32), namely

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T(\beta) \boldsymbol{V}_i^{-1}(\beta, \xi, \boldsymbol{x}_i)\{\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{x}_i, \beta)\} = \boldsymbol{0}, \tag{8.51}$$

is an ***unbiased estimating equation***, so that we expect $\widehat{\beta}$ obtained by solving (8.51) to be a ***consistent estimator*** for the true value β_0 of β.

Implicit in (8.50) is that the assumption (8.48),

$$E(Y_{ij}|\boldsymbol{x}_i) = E(Y_{ij}|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}) = E(Y_{ij}|\boldsymbol{x}_{ij}),$$

holds. We now take a closer look at the implications of this.

In a ***world-famous*** paper, Pepe and Anderson (1994) made the following simple but critically important observation. Assume, as would almost always be the case, that the ***working*** covariance model (8.5) for var$(\boldsymbol{Y}_i|\boldsymbol{x}_i)$,

$$\boldsymbol{V}_i(\beta, \xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\beta, \theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\beta, \theta, \boldsymbol{x}_i),$$

is such that $\boldsymbol{T}_i(\beta, \theta, \boldsymbol{x}_i) = \sigma^2 \text{diag}\{g^2(\beta, \delta, \boldsymbol{x}_{i1}), \ldots, g^2(\beta, \delta, \boldsymbol{x}_{in_i})\}$, as when the variance function depends on the mean response; and the ***working correlation model*** $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ possibly depends on $\boldsymbol{x}_i$, usually through the times t_{ij}.

Under these conditions, the estimating equation (8.51) can be written as, ignoring the multiplicative scale parameter σ^2, (verify)

$$\sum_{i=1}^{m} \left(\frac{f_\beta(\boldsymbol{x}_{i1}, \beta)}{g^2(\beta, \delta, \boldsymbol{x}_{i1})} \cdots \frac{f_\beta(\boldsymbol{x}_{in_i}, \beta)}{g^2(\beta, \delta, \boldsymbol{x}_{in_i})} \right) \begin{pmatrix} \Gamma^{i11} & \cdots & \Gamma^{i1n_i} \\ \vdots & \vdots & \vdots \\ \Gamma^{in_i1} & \cdots & \Gamma^{in_in_i} \end{pmatrix} \begin{pmatrix} Y_{i1} - f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ Y_{in_i} - f(\boldsymbol{x}_{in_i}, \beta) \end{pmatrix}, \tag{8.52}$$

where Γ^{ijk} is the (j, k) element of the ***inverse*** of the ***working correlation matrix***, and dependence of this matrix on $\boldsymbol{x}_i$ through the t_{ij} is suppressed.

Using shorthand notation

$$f_{\beta ij} = f_\beta(\boldsymbol{x}_{ij}, \beta), \qquad g_{ij}^{-2} = g^{-2}(\beta, \delta, \boldsymbol{x}_{ij}),$$

it is straightforward to rewrite (8.52) as

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} f_{\beta ik} g_{ik}^{-2} \Gamma^{ikj} \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \}. \tag{8.53}$$

Taking the ***conditional expectation*** of a summand in (8.53) given $\boldsymbol{x}_i$ yields

$$\begin{aligned} E\left[f_{\beta ik} g_{ik}^{-2} \Gamma^{ikj} \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \} | \boldsymbol{x}_i \right] &= f_{\beta ik} g_{ik}^{-2} \Gamma^{ikj} E\left[\{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \} | \boldsymbol{x}_i \right] \\ &= f_{\beta ik} g_{ik}^{-2} \Gamma^{ikj} \{ E(Y_{ij} | \boldsymbol{x}_i) - f(\boldsymbol{x}_{ij}, \beta) \}. \end{aligned} \tag{8.54}$$

It follows that (8.54) will be equal to zero ***only if***

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

that is, if $E(Y_{ij}|\boldsymbol{x}_i)$ depends on $\boldsymbol{x}_i$ ***only through*** $\boldsymbol{x}_{ij}$, as in (8.48).

This result has ***important but often unappreciated*** implications. If the analyst assumes a model of the form (8.50), s/he must be willing to assume that (8.48),

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

holds. If this ***does not hold*** , then the estimator for β can be ***inconsistent*** for the true value β_0.

Thus, if one assumes a ***marginal population mean model*** as in (8.49) directly, one must also be willing to assume (8.48) to ensure valid inferences.

Pepe and Anderson (1994) noted an additional result. If one adopts an ***independence working assumption*** for var($\boldsymbol{Y}_i|\boldsymbol{x}_i$), so takes the ***working correlation matrix*** to be a $(n_i \times n_i)$ ***identity matrix*** , it is straightforward that (8.53) reduces to

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} f_{\beta ij} g_{ij}^{-2} \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \}. \tag{8.55}$$

Note that, for a summand of (8.55),

$$E\left[f_{\beta ij} g_{ij}^{-2} \{ Y_{ij} - f(\boldsymbol{x}_{ij}, \beta) \} | \boldsymbol{x}_{ij} \right] = f_{\beta ij} g_{ij}^{-2} \{ E(Y_{ij} | \boldsymbol{x}_{ij}) - f(\boldsymbol{x}_{ij}, \beta) \}. \tag{8.56}$$

It follows from (8.56) that, as long as $f(\boldsymbol{x}_{ij}, \beta)$ is a ***correctly specified*** model for $E(Y_{ij}|\boldsymbol{x}_{ij})$, the estimating equation (8.51) is ***unbiased***.

RESULT: These results suggest that, if one is ***not willing*** to assume (8.48), the ***working correlation matrix*** should be taken to be an identity matrix; that is, a ***independence working assumption*** should be made to ensure consistent inference. This is critical if one is interested in ***marginal inference*** as discussed above.

- Unfortunately, as noted above, appreciation for this result is not widespread, and this advice is ***rarely*** followed in practice.

8.7 Missing data

As discussed in Section 5.6, a key challenge in longitudinal data analysis is ***missing data***, in particular ***dropout*** of individuals over the course of the observation period. In this section, we review briefly the implications of such dropout for the ***validity of inferences*** in population-averaged, possibly nonlinear models for general response types fitted using ***GEE methods***. A more detailed treatment of this issue is presented Chapter 5 of the instructor's notes for the course "Statistical Methods for Analysis With Missing Data."

DATA STRUCTURE: As we did previously, we consider the situation where the ***full data*** on the response ***intended*** to be collected on each individual i are at n prespecified times $t_1, \ldots, t_n$. As in (5.93), define the full data response vector as

$$\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{in})^T. \tag{8.57}$$

The responses ***actually observed*** on i,, $\boldsymbol{Y}_i$, are a subset of the components of $\mathcal{Z}_i$.

The ***missing data indicators*** are, as in (5.94

$$R_{ij} = \begin{cases} 1 & \text{if } Z_{ij} \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases}$$

$j = 1, \ldots, n$, and

$$\mathcal{R}_i = (R_{i1}, \ldots, R_{in})^T, \tag{8.58}$$

where we denote possible values of $\mathcal{R}_i$ (vectors of 0s and 1s) by $\boldsymbol{r}$. In the particular case of ***dropout***, assuming all individuals are observed at t_1, which we take here to be ***baseline***, there are n possible ***missingness patterns r***, given by

$$\boldsymbol{r}^{(1)} = (1, 0, \ldots, 0), \quad \boldsymbol{r}^{(2)} = (1, 1, 0, \ldots, 0), \quad \ldots, \quad = \boldsymbol{r}^{(n)} = (1, 1, \ldots, 1). \tag{8.59}$$

Dropout is of course a ***monotone missingness pattern***.

We henceforth assume that ***all individuals*** are observed at baseline, so that $R_{i1} = 1$ for all i.

- It is clear that, under a ***dropout mechanism***, if $R_{ij} = 1$, then it must be the case that $R_{i,j-1} = \cdots = R_{i2} = R_{i1} = 1$ (convince yourself).
- If $R_{ij} = 1$, then clearly $Z_{i1} = Y_{i1}, \ldots, Z_{ij} = Y_{ij}$; that is, the responses ***actually observed*** are those ***intended*** through time j.

At each time t_j, $j = 1, \ldots, n$, it is also intended to collect ***covariates*** $\boldsymbol{x}_{ij}$, where $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \ldots, \boldsymbol{x}_{in}^T)^T$. Interest focuses on the ***relationship*** between population mean response and $\boldsymbol{x}_i$.

- As discussed in the last section, ***conceptual challenges*** arise in the case of ***endogenous covariates*** $\boldsymbol{x}_{ij}$. Accordingly, the developments we present in this section are restricted to the case where the $\boldsymbol{x}_{ij}$ are ***exogenous covariates***.
- Under ***exogeneity***, the value of $\boldsymbol{x}_i$ is observed/known to the data analyst ***throughout the study period***, regardless of whether or not i drops out.
- This would be the case in a study in which all covariates are ***time-independent***, recorded only at ***baseline*** (t_1).
- This would also hold in a situation like the air pollution monitoring example in the previous section, where the covariate (pollution) process evolves ***independently*** of the responses of any study participant.
- We thus take in this section $\boldsymbol{x}_{ij} = (\boldsymbol{a}_i, t_j)$, where $\boldsymbol{a}_i$ is a vector of ***exogenous covariates*** whose values are known to the data analyst at all n intended time points, regardless of dropout. The simplest case is where $\boldsymbol{a}_i$ are ***baseline covariates***.

It may also be the case that ***additional information*** $\boldsymbol{v}_{ij}$ is recorded on each i at each t_j.

- We emphasize that there is ***no interest*** in the $\boldsymbol{v}_{ij}$ insofar as the mean response goes. As we discuss shortly, the $\boldsymbol{v}_{ij}$ are of interest only for their potential utility for describing the ***missingness/dropout mechanism***.
- The $\boldsymbol{v}_{ij}$ are ***observed*** along with $Z_{ij} = Y_{ij}$ as long as i has ***not yet dropped out***; that is, as long as $R_{ij} = 1$.

SUMMARY: Combining, the scenario we consider in this section can be summarized as follows. For each individual i, the ***full data*** on responses in (8.57) and additional information $\boldsymbol{v}_{ij}$ ***intended to be collected*** are

$$\widetilde{\mathcal{Z}}_i = \{(Z_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Z_{in}, \boldsymbol{v}_{in}^T)^T\}^T. \tag{8.60}$$

along with $\boldsymbol{x}_i$, which is ***always observed***.

Here, $\mathcal{R}_i = (R_{i1}, \ldots, R_{in})^T$ in (8.58) is such that

$$R_{ij} = \begin{cases} 1 & \text{if } (Z_{ij}, \boldsymbol{v}_{ij}^T)^T \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases}$$

If $R_{ij} = 1$, then $(Z_{ij}, \boldsymbol{v}_{ij}^T)^T = (Y_{ij}, \boldsymbol{v}_{ij}^T)^T$.

DROPOUT INDICATOR: It is conventional in the situation of dropout to define for each i the ***dropout indicator***

$$D_i = 1 + \sum_{j=1}^{n} R_{ij}. \tag{8.61}$$

- It is straightforward from (8.61) that $D_i = j$ implies that i was ***last seen*** at t_{j-1}, so dropped out sometime in the time interval (t_{j-1}, t_j). In this case, ***by convention***, i is said to have dropped out at t_j.
- Moreover, from (8.59), the possible values of $\mathcal{R}_i$ are $\boldsymbol{r}^{(j)}$, $j = 1, \ldots, n$, where $\boldsymbol{r}^{(j)}$ corresponds to dropout at time $j+1$; i.e., being last seen at time t_j. It is thus clear that the events

  $$\{D_i = j + 1\} \qquad \text{and} \qquad \{\mathcal{R}_i = \boldsymbol{r}^{(j)}\} \tag{8.62}$$

  are ***equivalent*** (check).
- Because $R_{i1} = 1$ always, D_i has possible values $2, \ldots, n+1$, where $D_i = n+1$ corresponds the situation where the ***full data are observed***.

DROPOUT MECHANISMS: Recall from (5.96) that the various ***missing data mechanisms*** of MCAR, MAR, and MNAR are defined in terms of the density of $\mathcal{R}_i$ given the full data, which are $\widetilde{\mathcal{Z}}_i$ here, and $\boldsymbol{x}_i$; that is,

$$p(\boldsymbol{r}_i | \boldsymbol{Z}_i, \boldsymbol{x}_i) = \text{pr}(\mathcal{R}_i = \boldsymbol{r}_i | \widetilde{\mathcal{Z}}_i = \boldsymbol{Z}_i, \boldsymbol{x}_i).$$

Because, under dropout, the possible values of $\boldsymbol{r}_i$ are ***restricted*** to be $\boldsymbol{r}^{(j)}$, $j = 1, \ldots, n$, it is straightforward from (8.61) and (8.62) that we can express these mechanisms ***equivalently*** in terms of

$$\text{pr}(D_i = j+1 | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(\mathcal{R}_i = \boldsymbol{r}^{(j)} | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i), \qquad j = 1, \ldots, n. \tag{8.63}$$

Letting $\widetilde{\mathcal{Z}}_{(\boldsymbol{r}^{(j)}),i}$ denote the part of $\widetilde{\mathcal{Z}}_i$ that is ***observed*** when $\mathcal{R}_i = \boldsymbol{r}^{(j)}$, ***equivalently*** , when $D_i = j+1$, from (8.60),

$$\widetilde{\mathcal{Z}}_{(\boldsymbol{r}^{(j)}),i} = \{(Z_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Z_{ij}, \boldsymbol{v}_{ij}^T)^T\}^T = \{(Y_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Y_{ij}, \boldsymbol{v}_{ij}^T)^T\}^T.$$

For convenience shortly, define the ***history*** through time t_j as

$$\mathcal{H}_{ij} = [\{(Z_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Z_{ij}, \boldsymbol{v}_{ij}^T)^T\}^T, \boldsymbol{x}_i], \qquad j = 1, \ldots, n, \tag{8.64}$$

recognizing that $\boldsymbol{x}_i$ is known throughout time. If $D_i = j+1$, then note that

$$\mathcal{H}_{ik} = [\{(Y_{i1}, \boldsymbol{v}_{i1}^T)^T, \ldots, (Y_{ik}, \boldsymbol{v}_{ik}^T)^T\}^T, \boldsymbol{x}_i], \qquad k = 1, \ldots, j.$$

Using (8.63) and (8.64), the ***dropout mechanisms*** are conventionally represented as follows.

- ***Missing Completely at Random (MCAR).*** The probability of dropout at $j+1$ does not depend on $\widetilde{\mathcal{Z}}_i$; that is

  $$\text{pr}(D_i = j+1 | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j+1 | \boldsymbol{x}_i) = \pi(j+1, \boldsymbol{x}_i). \tag{8.65}$$

  Analogous to (5.97), (8.65) implies that

  $$D_i \perp\!\!\!\perp \widetilde{\mathcal{Z}}_i \,|\, \boldsymbol{x}_i \tag{8.66}$$

  which is equivalent to $\mathcal{R}_i \perp\!\!\!\perp \widetilde{\mathcal{Z}}_i \,|\, \boldsymbol{x}_i$.

- ***Missing at Random (MAR).*** The probability of dropout at $j+1$ depends on $\widetilde{\mathcal{Z}}_i$ ***only*** through components of $\widetilde{\mathcal{Z}}_i$ that ***are observed*** under dropout at $j+1$,

  $$\text{pr}(D_i = j+1 | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j+1 | \widetilde{\mathcal{Z}}_{(\boldsymbol{r}^{(j)}),i}, \boldsymbol{x}_i) = pr(D_i = j+1 | \mathcal{H}_{ij}) = \pi(j+1, \mathcal{H}_{ij}). \tag{8.67}$$

- ***Missing Not at Random (MNAR).*** The probability of dropout at $j+1$ depends on components of $\widetilde{\mathcal{Z}}_i$ that ***are not observed*** under dropout at $j+1$.

ADDITIONAL INFORMATION $\boldsymbol{v}_{ij}$: Although the $\boldsymbol{v}_{ij}$ are not of ***direct interest*** for modeling, the definitions above demonstrate that they may be implicated in the ***missingness mechanism***. Thus, if the $\boldsymbol{v}_{ij}$ are available to the data analyst, they are ***critical*** for justifying the ***assumption of MAR*** , as we do shortly.

OBSERVED DATA GEE: We are now in a position to consider the behavior of the estimator $\widehat{\beta}$ for β obtained by solving the usual linear estimating equation (8.51) based on the ***observed data***. It proves convenient to ***modify the notation*** as follows.

Recall from above that we restrict attention to $\boldsymbol{x}_{ij} = (\boldsymbol{a}_i, t_j)$, where $\boldsymbol{a}_i$ are ***exogenous covariates*** known throughout time and observed for all i. Suppose we have posited a model for the mean of the ***intended full response vector*** $\mathcal{Z}_i$ in (8.57), $E(\mathcal{Z}_i|\boldsymbol{x}_i)$, of the form

$$\boldsymbol{f}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f(\boldsymbol{x}_{in}, \beta) \end{pmatrix} = \begin{pmatrix} f(\boldsymbol{a}_i, t_1, \beta) \\ \vdots \\ f(\boldsymbol{a}_i, t_n, \beta) \end{pmatrix} = \begin{pmatrix} f_1(\boldsymbol{a}_i, \beta) \\ \vdots \\ f_n(\boldsymbol{a}_i \beta) \end{pmatrix} \qquad (n \times 1). \tag{8.68}$$

Note that conditioning on $\boldsymbol{x}_i$ here is ***equivalent*** to conditioning on $\boldsymbol{a}_i$. Assume that this model is ***correctly specified***.

With $\boldsymbol{f}_i(\boldsymbol{x}_i, \beta)$ defined as in (8.68), let

$$\mathcal{X}_i(\boldsymbol{x}_i, \beta) = \begin{pmatrix} f_\beta^T(\boldsymbol{x}_{i1}, \beta) \\ \vdots \\ f_\beta^T(\boldsymbol{x}_{in}, \beta) \end{pmatrix} = \begin{pmatrix} f_{1\beta}^T(\boldsymbol{a}_i, \beta) \\ \vdots \\ f_{n\beta}^T(\boldsymbol{a}_i, \beta) \end{pmatrix} \qquad (n \times p) \tag{8.69}$$

be the ***gradient matrix*** of the mean model for the intended full response vector in (8.68), and likewise and let $\mathcal{V}_i(\beta, \xi, \boldsymbol{x}_i)$ be a $(n \times n)$ ***working covariance model*** for $\text{var}(\mathcal{Z}_i|\boldsymbol{x}_i)$.

If $D_i = j + 1$, then the ***observed response vector*** on i is $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ij})^T$ $(n_i = j)$. Let $\boldsymbol{X}_i^{(j)}(\beta)$ $(j \times p)$ and $\boldsymbol{V}_i^{(j)}(\beta, \xi, \boldsymbol{x}_i)$ $(j \times j)$ be the corresponding submatrices of $\mathcal{X}_i(\boldsymbol{x}_i, \beta)$ and $\mathcal{V}_i(\beta, \xi, \boldsymbol{x}_i)$. These are as defined previously in this chapter for each i, where we have added the superscript (j) to ***emphasize*** that these correspond an ***observed data vector*** of length $n_i = j \leq n$.

With these definitions, the ***linear estimating equation*** (8.51) can be written as (verify)

$$\sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} I(D_i = j + 1) \boldsymbol{X}_i^{(j)T}(\beta) \{ \boldsymbol{V}_i^{(j)}(\beta, \xi, \boldsymbol{x}_i) \}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i, \beta) \\ \vdots \\ Y_{ij} - f_j(\boldsymbol{a}_i, \beta) \end{pmatrix} \right\} = \boldsymbol{0}. \tag{8.70}$$

Consider the conditional expectation of the ith summand in (8.70) given $\boldsymbol{x}_i$ $(\boldsymbol{a}_i)$.

Assuming expectation is under the parameter values β and ξ, the estimator $\widehat{\beta}$ solving (8.70) will be ***consistent*** for β_0 if

$$E\left\{ I(D_i = j+1)\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i,\beta) \\ \vdots \\ Y_{ij} - f_1(\boldsymbol{a}_i,\beta) \end{pmatrix} \middle| \boldsymbol{x}_i \right\} = \mathbf{0} \qquad j = 1,\ldots,n. \tag{8.71}$$

Consider the left hand side of (8.71) under ***different dropout/missingness mechanisms***.

- ***MCAR.*** Writing this as

$$E\left[E\left\{ I(D_i = j+1)\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i,\beta) \\ \vdots \\ Y_{ij} - f_1(\boldsymbol{a}_i,\beta) \end{pmatrix} \middle| \boldsymbol{x}_i, \widetilde{\mathcal{Z}}_i \right\} \middle| \boldsymbol{x}_i \right], \tag{8.72}$$

  and using the fact that, as in (8.66), D_i is ***independent of*** $\widetilde{\mathcal{Z}}_i$ given $\boldsymbol{x}_i$, the left hand side of (8.71) becomes (verify)

$$E\left\{ \pi(j+1,\boldsymbol{x}_i)\,\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1} \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i,\beta) \\ \vdots \\ Y_{ij} - f_1(\boldsymbol{a}_i,\beta) \end{pmatrix} \middle| \boldsymbol{x}_i \right\} = \mathbf{0}.$$

  Thus, under ***MCAR***, the estimating equation (8.70) is ***unbiased***, as intuition would suggest. Accordingly, if ***dropout*** is completely at random, so is ***unrelated*** to the evolving response, as might be the case if study participants dropped out to move to another city, $\widehat{\beta}$ obtained by solving the usual linear GEE is ***consistent***.

- ***MAR.*** Using the equivalent form (8.72), (8.71) can be written using (8.67) as (verify)

$$\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1} E\left\{ \pi(j+1,\mathcal{H}_{ij}) \begin{pmatrix} Y_{i1} - f_1(\boldsymbol{a}_i,\beta) \\ \vdots \\ Y_{ij} - f_1(\boldsymbol{a}_i,\beta) \end{pmatrix} \middle| \boldsymbol{x}_i \right\}. \tag{8.73}$$

  The kth element in the expectation in (8.72) is (verify), $k = 1,\ldots,j$,

$$\text{cov}\{\pi(j+1,\mathcal{H}_{ij}),\, Y_{ik}\}.$$

  From (8.67), $\pi(j+1,\mathcal{H}_{ij})$ ***depends on*** $Y_{i1},\ldots,Y_{ij}$; thus, this covariance is almost certainly ***not equal to zero*** in general. Thus, under MAR, the usual estimator $\widehat{\beta}$ ***is not consistent for*** β_0 ***in general***.

- ***MNAR.*** It should be clear that ***similar considerations*** apply here, and the estimator is ***not consistent*** in general.

RESULT: Even under the simpler condition of ***exogenous covariates***, when there are missing data due to ***dropout***, the usual GEE estimator obtained by solving (8.51) based on the ***observed data*** is ***only guaranteed to be consistent*** if the dropout mechanism is MCAR.

- Given that MCAR is often ***not realistic*** in practice, this is a rather ***scary*** result.
- The data analyst thus must take ***great care*** to understand the possible reasons for dropout.

REMARK: The foregoing developments were derived in the context of ***dropout***, which is a ***monotone missingness mechanism***. Under ***nonmonotone patterns***, the same scary considerations apply: the estimator for β based on the ***observed data*** is only guaranteed to be consistent if the missingness mechanism is MCAR.

Moreover, methods to achieve a consistent estimator for β under the assumption of ***MAR*** with ***nonmonotone missingness*** are ***very difficult*** to develop. However, with ***monotone missingness***, i.e., dropout, such methods are possible, as we now discuss.

MODIFIED ESTIMATING EQUATIONS UNDER MAR DROPOUT: When the dropout mechanism is thought to be ***MAR***, it is possible to ***modify*** the usual ***linear estimating equation*** (8.70) to obtain estimating equations that ***are unbiased***. We briefly sketch the two main approaches.

We first consider a convenient representation of $\text{pr}(D_i = j+1|\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i)$. Define the ***cause-specific hazard function*** of dropout as

$$\lambda_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j|D_i \geq j, \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i), \qquad j = 2, \ldots, n. \tag{8.74}$$

Note that $\lambda_1(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = 1|D_i \geq 1, \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = 0$ because $(Z_{i1}, \boldsymbol{v}_{i1})$ is ***always observed*** for all i; and $\lambda_{n+1}(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = n+1|D_i \geq n+1, \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = 1$ by construction. It can then be deduced (do it) that

$$\overline{\pi}_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(R_{ij} = 1|\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \prod_{\ell=1}^{j}\{1 - \lambda_\ell(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i)\}, \qquad j = 2, \ldots, n, \tag{8.75}$$

where $\overline{\pi}_1(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(R_{i1} = 1|\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(R_{i1} = 1) = \overline{\pi}_1 = 1$, because all individuals are observed at baseline, and thus (verify)

$$\text{pr}(D = j+1|\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \overline{\pi}_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i)\lambda_{j+1}(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i), \qquad j = 1, \ldots, n. \tag{8.76}$$

Under MAR, (8.74) can be written as

$$\lambda_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j | D_i \geq j, \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D_i = j | D_i \geq j, \mathcal{H}_{i,j-1}) = \lambda_j(\mathcal{H}_{i,j-1}), \qquad j = 2, \ldots, n. \quad (8.77)$$

so that the ***hazard of dropping out*** at time t_j, so last being seen at time t_{j-1}, depends only on the ***observed history*** through time t_{j-1}. Similarly, (8.75) and (8.76) become

$$\overline{\pi}_j(\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \overline{\pi}_j(\mathcal{H}_{i,j-1}) = \text{pr}(R_{ij} = 1 | \mathcal{H}_{i,j-1}) = \prod_{\ell=1}^{j} \{1 - \lambda_\ell(\mathcal{H}_{i,j-1}\}, \qquad j = 2, \ldots, n, \quad (8.78)$$

where $\overline{\pi}_1 = 1$, and

$$\text{pr}(D = j + 1 | \widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i) = \text{pr}(D = j + 1 | \mathcal{H}_{ij}) = \overline{\pi}_j(\mathcal{H}_{i,j-1})\lambda_{j+1}(\mathcal{H}_{ij}), \qquad j = 1, \ldots, n. \quad (8.79)$$

Equations (8.77)–(8.79) demonstrate that, ***under the assumption of MAR***, it is possible to develop ***models*** for the ***hazard functions***

$$\lambda_j(\mathcal{H}_{i,j-1})$$

based on the ***observed data*** and thereby obtain models for

$$\text{pr}(R_{ij} = 1 | \mathcal{H}_{i,j-1}) = \overline{\pi}_j(\mathcal{H}_{i,j-1}) \quad \text{and} \quad \text{pr}(D = j + 1 | \mathcal{H}_{ij}).$$

These models can then be ***fitted and substituted*** into the ***modified estimating equations*** we now discuss.

WEIGHTED GENERALIZED ESTIMATING EQUATIONS (WGEEs): The modified estimating equations involve ***weighting*** each the summand of the usual linear estimating equation, which we wrote in the form (8.70), in a way that yields unbiasedness. The two main approaches are:

- ***Inverse probability weighting at the individual level.*** This was proposed by Fitzmaurice, Molenberghs, and Lipsitz (1995) and involves ***weighting*** the contribution to (8.70) for each individual by the ***inverse*** of the probability of that individual's observed dropout time, conditional on his/her observed history.
- ***Inverse probability weighting at the occasion level.*** This was proposed by Robins, Rotnitzky, and Zhao (1995) and involves ***weighting*** contributions to (8.70) at ***each time point*** for ***each individual*** by the inverse of the probability of having an observed response at that time point, conditional on observed history.
- These and more advanced techniques can be justified theoretically by appealing to the general theory of ***semiparametrics*** and missing data as in Tsiatis (2006).

INVERSE PROBABILITY WEIGHTING AT THE INDIVIDUAL LEVEL: Define the weight

$$w_{ij} = \frac{I(D_i = j+1)}{\overline{\pi}_j(\mathcal{H}_{i,j-1})\lambda_{j+1}(\mathcal{H}_{ij})},$$

where, from (8.79), the denominator of w_{ij} is $\text{pr}(D = j+1|\mathcal{H}_{ij})$. The ***WGEE*** is then

$$\sum_{i=1}^{m}\left\{\sum_{j=1}^{n} w_{ij}\boldsymbol{X}_i^{(j)T}(\beta)\{\boldsymbol{V}_i^{(j)}(\beta,\xi,\boldsymbol{x}_i)\}^{-1}\begin{pmatrix} Y_{i1}-f_1(\boldsymbol{a}_i,\beta)\\ \vdots \\ Y_{ij}-f_j(\boldsymbol{a}_i,\beta)\end{pmatrix}\right\} = \boldsymbol{0}. \tag{8.80}$$

Comparing (8.80) to (8.70), the only difference is the ***inverse weighting*** by $\text{pr}(D = j+1|\mathcal{H}_{ij})$.

The ***diligent student*** can verify that, if $\lambda_j(\mathcal{H}_{i,j-1})$ are the ***true hazard functions***, the expectation of a summand of (8.80) is indeed equal to zero, so that (8.80) is an ***unbiased estimating equation*** under MAR. This can be accomplished by representing the expectation of a summand as in (8.72), with inner conditioning on $\widetilde{\mathcal{Z}}_i, \boldsymbol{x}_i$.

INVERSE PROBABILITY WEIGHTING AT THE OCCASION LEVEL: Define the $(n \times n)$ diagonal weight matrix

$$\mathcal{W}_i = \text{diag}\left(\frac{R_{i1}}{\overline{\pi}_1}, \frac{R_{i2}}{\overline{\pi}_2(\mathcal{H}_{i1})}, \ldots, \frac{R_{in}}{\overline{\pi}_n(\mathcal{H}_{i,n-1})}\right).$$

The ***WGEE*** is then

$$\sum_{i=1}^{m}\mathcal{X}_i^T(\boldsymbol{x}_i,\beta)\mathcal{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\mathcal{W}_i\begin{pmatrix} Z_{i1}-f_1(\boldsymbol{a}_i,\beta)\\ \vdots \\ Z_{in}-f_n(\boldsymbol{a}_i,\beta)\end{pmatrix} = \boldsymbol{0}, \tag{8.81}$$

where $\mathcal{X}_i^T(\boldsymbol{x}_i\beta)$ and $\mathcal{V}_i(\beta,\xi,\boldsymbol{x}_i)$ are defined in and below (8.69). From the form of $\mathcal{W}_i$, when $R_{ij} = 1$, so that i has not yet dropped out, $Z_{ij} = Y_{ij}$, the observed response at t_j. If then $R_{i,j+1} = 0$, all subsequent $R_{ik} = 0$, $k > j$, and it is straightforward to observe that the summand will depend only on $Y_{i1}, \ldots, Y_{ij}$, $\boldsymbol{X}_i^{(j)}(\beta)$ $(j \times p)$, and the upper left $(j \times j)$ submatrix of $\mathcal{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)$ (verify).

As with (8.80), it can be shown by a similar conditioning argument that the conditional (on $\boldsymbol{x}_i$) expectation of a summand of (8.81) is equal to zero, so that (8.81) is an ***unbiased estimating equation*** under MAR.

RESULT: If the assumption of ***MAR dropout*** is plausible, then these methods can be used to obtain consistent estimators for β in a ***full data model*** of interest with exogenous covariates.

- However, this requires ***correct specification*** of ***models*** for the ***dropout hazard functions*** $\lambda_j(\mathcal{H}_{i,j-1})$, $j = 2, \ldots, n$. If these models are ***misspecified***, then the estimating equations ***no longer*** need be unbiased.

Considerations for such modeling are discussed in Chapter 5 of the instructor's notes for the course "Statistical Methods for Analysis With Missing Data," where it is demonstrated that one approach is to adopt ***logistic regression models*** for the hazards for each j.

- It is ***not possible*** to show that one weighting approach yields ***more efficient inferences*** than the other in general. The individual-level approach has been preferred in practice on the grounds that it is ***simpler to implement***. Specifically, one can model and fit the dropout hazard functions as above and form ***fixed, estimated weights*** w_{ij}. Many software packages for solving the linear GEE allow fixed weights for each individual, so that these estimated weights can be incorporated straightforwardly.

  The only widely available software that implements both methods, including specification of the ***hazard models***, is SAS `proc gee`. Its use is demonstrated in the instructor's notes for the course "Statistical Methods for Analysis With Missing Data."

8.8 Examples

We briefly describe how modeling might proceed in two examples.

EXAMPLE 5: Epileptic seizures and chemotherapy, continued. Recall that the ***among-individual covariates*** randomized treatment δ_i (= 0 for placebo and = 1 for progabide), baseline seizure count over the 8 weeks prior to the start of the study, c_i, and age a_i at the start of the study ***do not change*** over time. Thus, $\boldsymbol{x}_{ij} = (t_{ij}, \delta_i, c_i, a_i)^T$ are ***exogenous***. All $m = 59$ subjects are seen at all $n = 4$ visits, with no missing responses. The response Y_{ij} is a ***count***, so a natural model is a ***loglinear model*** as in (8.7),

$$f(\boldsymbol{x}_{ij}, \beta) = \exp\{h(\boldsymbol{x}_{ij})^T\beta) \quad \text{or equivalently} \quad \log\{f(\boldsymbol{x}_{ij}, \beta)\} = h(\boldsymbol{x}_{ij})^T\beta,$$

where $h(\cdot)$ is a vector of functions of $\boldsymbol{x}_{ij}$.

In the above, we treat the ***baseline seizure count*** as a ***covariate***. However, strictly speaking, baseline seizure count also a measure of the ***response*** prior to the start of treatment, albeit over an observation period of 8 weeks rather than 2 weeks as is the case for the post-treatment responses. Many authors, including Thall and Vail (1990) themselves, have regarded baseline seizure count as a ***covariate***, in part to avoid the issue of the ***different lengths*** of the observation periods. While this is convenient, it could also be ***inefficient***, as the baseline count also contains information on the ***distribution of responses***.

It is a ***simple matter*** to address the time scale issue, as we demonstrate momentarily, so in what follows, we follow Diggle et al. (2002) and treat the baseline 8-week seizure count as a ***response measure*** Y_{i1} at time $t_{ij} = 0$, and ***reindex*** the responses at periods 1–4 as $Y_{i2}, \ldots, Y_{i5}$, so that $j = 1, \ldots, n = 5$. Accordingly, we ***redefine*** $\boldsymbol{x}_{ij} = (t_{ij}, \delta_i, a_i)^T$. Note that, here, our temporal convention is ***different from*** that in Section 8.6, although this poses no difficulty.

The more ***fundamental issue*** is whether or not it is a good idea to treat a baseline response as a ***covariate*** to take into account the fact that individuals differ in their responses prior to treatment or if it is preferable to treat the baseline value as ***part of the response vector*** for each individual. In the case of ***linear models*** for the mean response, the two strategies can be ***equivalent*** ; however, when the mean response model is ***nonlinear*** as it is here, this is no longer the case.

This is a matter of ***considerable debate*** in the literature, and the choice is in part guided by the nature of the ***questions of interest*** and the ***type of study***. This instructor agrees with Fitzmaurice, Laird, and Ware (2011) that, as a general strategy, it is ***preferable*** to treat a baseline response as part of the response vector rather than as a covariate, partly on ***efficiency grounds***. A very nice, practical discusion of this issue is given in Sections 5.6 and 5.7 of Fitzmaurice et al. (2011).

Accordingly, define $o_{ij} = 8$ if $j = 1$ ($t_{ij} = 0$, baseline) and $o_{ij} = 2$, $j = 2, \ldots, 5$ ($t_{ij} > 0$, observation period/visit 1–4). Thus, o_{ij} records the time scale over which Y_{ij} was ascertained (8 or 2 weeks). Define $v_{ij} = 0$ if $t_{ij} = 0$ (baseline) and $v_{ij} = 1$ if $t_{ij} > 0$ (visit 1–4). We consider the loglinear model

$$E(Y_{ij}|\boldsymbol{x}_i) = \exp(\log o_{ij} + \beta_0 + \beta_1 v_{ij} + \beta_2 \delta_i + \beta_3 \delta_i v_{ij}). \tag{8.82}$$

In (8.82), $\log o_{ij}$ is an ***offset*** in the following sense. On the log scale, (8.82) is equivalent to

$$\log\{E(Y_{ij}|\boldsymbol{x}_i)\} - \log o_{ij} = \log\{E(Y_{ij}/o_{ij}|\boldsymbol{x}_i)\} = \beta_0 + \beta_1 v_{ij} + \beta_2 \delta_i + \beta_3 \delta_i v_{ij}. \tag{8.83}$$

Thus, (8.83) shows that (8.82) is equivalent to modeling the means of $Y_{i1}/8$ and $Y_{ij}/2$, $j = 2, \ldots, 5$; that is, the ***average number of seizures*** per week over each period.

Model (8.82) specifies that the mean of average numbers of seizures per week at baseline ($j = 1$) is

$$E(Y_{i1}/8|\boldsymbol{x}_i) = \exp(\beta_0 + \beta_2 \delta_i)$$

and for $J = 2, \ldots, 5$ is

$$E(Y_{i1}/2|\boldsymbol{x}_i) = \exp\{\beta_0 + \beta_1 + (\beta_2 + \beta_3)\delta_i\}.$$

This model is consistent with the impression given by the sample mean average numbers of seizures, as summarized below:

Visit	*Placebo*	*Progabide*
0 (baseline)	7.70	7.90
1	9.35	8.58
2	8.29	8.42
3	8.79	8.13
4	7.96	6.71
average over visits 1–4	8.60	7.96

The sample means in each group take a ***jump*** at visit 1 to a higher level that is possibly different in each group, represented in the model by the terms $\beta_1 v_{ij}$ and $\beta_3 v_{ij}\delta_i$. The remain relatively flat in each group thereafter, although there is an apparent drop at visit 4 in the progabide group (see below).

Given that this is a ***randomized study*** , a ***simplification*** of the model would be to ***remove*** the term $\beta_2\delta_i$, which allows the mean to differ at baseline between the two treatment groups. Indeed, the raw sample means of average number of seizures in the two groups are almost ***identical***. Average age at baseline (average of a_i) is 29.6 (SD 6.0) for the placebo group and 27.7 (6.6) for the progabide group, which are very similar, further supporting the contention that the randomization was carried out appropriately. We maintain the $\beta_2\delta_i$ term in analyses on the course website, but it probably could be deleted.

A modification of the model is as follows. As above, the sample means seem to suggest a ***possible drop*** in mean response at the 4th visit, we define an additional indicator variable $v4_{ij} = 0$ unless $j = 5$ for the possibility that the mean response at baseline is associated with age of the subject. The model is modified to

$$E(Y_{ij}|\boldsymbol{x}_i) = \exp(\log o_{ij} + \beta_0 + \beta_1 v_{ij} + \beta_2\delta_i + \beta_3 v_{ij}\delta_i + \beta_4 v4_{ij} + \beta_5 v4_{ij}\delta_i).$$

The parameter β_5 reflects whether or not the difference in post-baseline mean response in fact ***changes*** at the fourth visit, while β_4 allows the possibility that the mean response "shifts" at the 4th visit relative to the earlier ones. A further modification would be to incorporate age at baseline into the model to evaluate if treatment effects differ with age.

Fits of these models in SAS and R are on the course website.

EXAMPLE 6: Maternal smoking and child respiratory health, continued. The ***among-individual covariates*** are city c_i (= 0 for Portage, = 1 for Kingston), which is ***time-independent*** , and mother's smoking status at child's age (time) t_{ij}, s_{ij} (=0, 1, or 2 as the mother's smoking was none, moderate, or heavy). Define $\boldsymbol{x}_{ij} = (t_{ij}, c_i, s_{ij})^T$ for mother-child pair i.

Recall that this is an ***observational study*** ; thus, as discussed in Section 8.6, it is very likely that mother's smoking status s_{ij} is ***not exogenous***. Thus, because of the ***confounding*** that could be present, attempting to draw ***causal interpretations*** regarding the effect of mother's smoking on child respiratory status is ill-advised. Accordingly, we specify a ***marginal model*** only recognizing that we can do nothing more than evaluate the ***association*** between mother's ***current smoking status*** and and the probability her child is currently experiencing respiratory problems. A ***causal analysis*** would require the use of specialized techniques for this purpose, as discussed in Section 8.6.

Because the response is ***binary***, $E(Y_{ij}|\boldsymbol{x}_{ij}) = \text{pr}(Y_{ij} = 1|\boldsymbol{x}_{ij}) = f(\boldsymbol{x}_{ij}, \beta)$. We specify directly such a marginal model as

$$\text{logit}\{f(\boldsymbol{x}_{ij}, \beta)\} = \beta_0 + \beta_1 c_i + \beta_2 I(s_{ij} = 0) + \beta_3 I(s_{ij} = 1). \quad (8.84)$$

Thus, for example, the probability that child i is wheezing at age t_{ij} ($Y_{ij} = 1$) if that child is from Kingston and his mother is a heavy smoker at t_{ij} is

$$\frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)},$$

and thus the ***odds*** that such a child would be wheezing are

$$\exp(\beta_0 + \beta_1).$$

If follows that the ***odds ratio*** comparing the odds that a child from Kingston whose mother is a nonsmoker is wheezing at t_{ij} to the odds that a child from Kingston whose mother is a heavy smoker is then $\exp(\beta_2)$. If $\beta_2 < 0$, the odds of wheezing are smaller under a nonsmoking mother. Of course, these should be interpreted as ***purely associational*** statements.

Following the discussion regarding ***unbiasedness*** of the linear generalized estimating equation at the end of Section 8.6, fitting the ***directly-specified marginal model*** (8.84) should probably be implemented using the ***independence working assumption*** for $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$.

To complicate matters ***further***, of the $m = 32$ mother-child pairs, 14 have $(Y_{ij}, \boldsymbol{x}_{ij})$ is ***missing*** at one or more of the ages (times) t_{ij}, and this missingness is ***nonmonotone*** for most of these pairs. As discussed in Section 8.7, unless the missingness mechanism is ***MCAR***, the above associational analysis could be flawed due to possible ***inconsistency*** of the estimator for β used.

Of course, without further information, it is ***difficult*** to make a subject-matter based determination if MCAR is plausible. It ***would be*** possible to investigate, under the assumption that the mechanism is MAR, whether or not it is in fact MCAR by fitting an appropriate models for the missingness (e.g., ***hazard models*** as in Section 8.7) to investigate if missingness is related to wheezing status; this is beyond our scope here. However, as discussed in Section 5.6, it is ***impossible*** to determine from the data if MAR is the true mechanism. Accordingly, analyses based on the ***observed data*** should be viewed with caution.

Even if we are willing to make the assumption of MAR, because the missingness is not ***monotone***, the ***WGEE*** methods at the end of Section 8.7 are not feasible. Other approaches that accommodate nonmonotone patterns must be used. See the instructor's notes for the course "Statistical Methods for Analysis With Missing Data."

Subject to these caveats, fits of (8.84) using SAS and R, which should be viewed as ***purely illustrative***, are on the course website.

8.9 Further results for quadratic equations

As noted in Section 8.3, it is possible to show analytically the equivalence between (8.24), namely,

$$(1/2)\sum_{i=1}^{m}\Big(\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\boldsymbol{\xi},\boldsymbol{x}_i)\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}$$
$$-\operatorname{tr}\left[\boldsymbol{V}_i^{-1}(\beta,\boldsymbol{\xi},\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\beta,\boldsymbol{\xi},\boldsymbol{x}_i)\}\right]\Big)=0,\quad k=1,\ldots,r+s, \tag{8.85}$$

and the alternative form (8.25) with $\boldsymbol{Z}(\beta,\boldsymbol{\xi})$ is chosen according to the Gaussian working assumption,

$$\sum_{i=1}^{m}\boldsymbol{E}_i^T(\beta,\boldsymbol{\xi})\boldsymbol{Z}_i^{-1}(\beta,\boldsymbol{\xi})\{\boldsymbol{u}_i(\beta)-\boldsymbol{v}_i(\beta,\boldsymbol{\xi})\}=\boldsymbol{0}, \tag{8.86}$$

where

$$\boldsymbol{u}_i(\beta)=\operatorname{vech}\left[\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\right].$$

In practice, squared terms are deleted if the model contains no unknown variance parameters. We do not note this explicitly in the following argument.

To show that (8.85) and (8.86) are in fact the ***same estimating equation*** under the conditions above, it ***suffices*** to show that their kth rows coincide. The kth row of (8.85) may be written, using the identity for quadratic forms $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \text{tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T)$, as

$$(1/2)\sum_{i=1}^{m}\text{tr}\Big[\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}\{\boldsymbol{Y}_i-\boldsymbol{f}_i(\boldsymbol{x}_i,\beta)\}^T\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i) - \{\partial/\partial\xi_k\,\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\Big] = 0. \tag{8.87}$$

Noting that $\boldsymbol{E}_i(\beta,\xi)$ has kth row $\{\partial/\partial\xi_k\,\boldsymbol{v}_i(\beta,\xi)\}^T$, we can write the kth row of (8.86) as

$$\sum_{i=1}^{m}\{\partial/\partial\xi_k\,\boldsymbol{v}_i(\beta,\xi)\}^T\boldsymbol{Z}_i(\beta,\xi)\{\boldsymbol{u}_i(\beta)-\boldsymbol{v}_i(\beta,\xi)\} = 0. \tag{8.88}$$

We can thus show the result by showing that the ith summand in (8.87) is equal to that in (8.88).

We use several matrix results given in Appendix A, which we repeat here for convenience; these can be found with discussion in Chapter 16 of Harville (1997) or in Appendix 4.A of Fuller (1987). For matrices $\boldsymbol{A}$ ($a\times a$), $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$,

(i) $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \{\text{vec}(\boldsymbol{A})\}^T\{\text{vec}(\boldsymbol{B}^T)\} = \{\text{vec}(\boldsymbol{A}^T)\}^T\{\text{vec}(\boldsymbol{B})\}$.

(ii) $\text{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{D}^T\boldsymbol{C}^T) = \{\text{vec}(\boldsymbol{A})\}^T(\boldsymbol{B}\otimes\boldsymbol{C})\text{vec}(\boldsymbol{D})$, where "$\otimes$" is Kronecker product.

(iii) For $\boldsymbol{A}$ symmetric, there exists a unique matrix $\boldsymbol{\Phi}$ of dimension $\{a^2\times a(a+1)/2\}$ such that

$$\text{vec}(\boldsymbol{A}) = \boldsymbol{\Phi}\text{vech}(\boldsymbol{A}).$$

Clearly, $\boldsymbol{\Phi}$ is unique and of full column rank, as there is only one way to write the distinct elements of $\boldsymbol{A}$ in a full, redundant vector.

There also exist ***many*** (not unique) linear transformations of vec($\boldsymbol{A}$) into vech($\boldsymbol{A}$). Consider a transformation matrix $\boldsymbol{\Psi}$ of dimension $\{a(a+1)/2\times a^2\}$ such that

$$\text{vech}(\boldsymbol{A}) = \boldsymbol{\Psi}\text{vec}(\boldsymbol{A}).$$

One particular choice of $\boldsymbol{\Psi}$ is the Moore-Penrose generalized inverse of $\boldsymbol{\Phi}$, $\boldsymbol{\Psi} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$. Fuller (1987, page 383) gives the actual form of $\boldsymbol{\Phi}$.

Because we are addressing equivalency under the ***assumption of normality*** , take $\boldsymbol{Y}_i|\boldsymbol{x}_i$ to be normally distributed for the purposes of the argument.

Under these conditions, it is possible to show [see, for example, Fuller (1987, Lemma 4.A.1)] that

$$\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\} = 2\boldsymbol{\Psi}\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\otimes\{\boldsymbol{V}_i(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{\Psi}^T. \tag{8.89}$$

In fact, (8.89) is a compact way of expressing (8.23). Result 4.A.3.1 of Fuller (1987, page 385) then yields that

$$\{\text{var}\{\boldsymbol{u}_i(\beta)|\boldsymbol{x}_i\}\}^{-1} = (1/2)\boldsymbol{\Phi}^T\{\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\otimes\boldsymbol{V}_i^{-1}(\beta,\xi,\boldsymbol{x}_i)\}\boldsymbol{\Phi}. \tag{8.90}$$

We are now in a position to show the desired correspondence. For brevity, we suppress the arguments of all matrices and vectors. The estimating equation in (8.87) has two parts:

$$(1/2)\text{tr}\{(\partial/\partial\xi_k\boldsymbol{V}_i)\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{f}_i)(\boldsymbol{Y}_i-\boldsymbol{f}_i)^T\boldsymbol{V}_i^{-1}\} \tag{8.91}$$

and

$$-(1/2)\text{tr}\{(\partial/\partial\xi_k\boldsymbol{V}_i)\boldsymbol{V}_i^{-1}\}. \tag{8.92}$$

Consider (8.91). By (ii) above, identifying $\boldsymbol{A} = \partial/\partial\xi_k\boldsymbol{V}_i$, $\boldsymbol{B} = \boldsymbol{V}_i^{-1}$, $\boldsymbol{D}^T = (\boldsymbol{Y}_i-\boldsymbol{f}_i)(\boldsymbol{Y}_i-\boldsymbol{f}_i)^T$, and $\boldsymbol{C}^T = \boldsymbol{V}_i^{-1}$, and using the definition of $\boldsymbol{u}_i$,

$$\begin{aligned}
&(1/2)\text{tr}\{(\partial/\partial\xi_k\boldsymbol{V}_i)\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{f}_i)(\boldsymbol{Y}_i-\boldsymbol{f}_i)^T\boldsymbol{V}_i^{-1}\}\\
&\qquad = (1/2)\{\text{vec}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\text{vec}\{(\boldsymbol{Y}_i-\boldsymbol{f}_i)(\boldsymbol{Y}_i-\boldsymbol{f}_i)^T\}\\
&\qquad = (1/2)\{\boldsymbol{\Phi}\text{vech}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\boldsymbol{\Phi}\boldsymbol{u}_i\\
&\qquad = \{\text{vech}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T\{(1/2)\boldsymbol{\Phi}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\boldsymbol{\Phi}\}\boldsymbol{u}_i,
\end{aligned} \tag{8.93}$$

From the definition of $\boldsymbol{v}_i$,

$$\text{vech}(\partial/\partial\xi_k\,\boldsymbol{V}_i) = \partial/\partial\xi_k\,\text{vech}(\boldsymbol{V}_i) = \partial/\partial\xi_k\boldsymbol{v}_i.$$

Moreover, the "middle" term in (8.93) in braces, by (8.90), equals $\boldsymbol{Z}_i^{-1}$, as we are doing these calculations under normality. Substituting these developments into (8.93) yields

$$\{\partial/\partial\xi_k\boldsymbol{v}_i\}^T\boldsymbol{Z}_i^{-1}\boldsymbol{u}_i. \tag{8.94}$$

Now consider (8.92). Applying (ii) gives

$$\begin{aligned}
&-(1/2)\{\text{vec}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\text{vec}(\boldsymbol{V}_i)\\
&\qquad = -(1/2)\{\boldsymbol{\Phi}\text{vech}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\boldsymbol{\Phi}\text{vech}(\boldsymbol{V}_i)\\
&\qquad = -\{\text{vech}(\partial/\partial\,\xi_k\boldsymbol{V}_i)\}^T\{(1/2)\boldsymbol{\Phi}^T(\boldsymbol{V}_i^{-1}\otimes\boldsymbol{V}_i^{-1})\boldsymbol{\Phi}\}\boldsymbol{v}_i\\
&\qquad = -\{\partial/\partial\xi_k\boldsymbol{v}_i\}^T\boldsymbol{Z}_i^{-1}\boldsymbol{v}_i.
\end{aligned} \tag{8.95}$$

Combining (8.94) and (8.95), we obtain that the kth row of the PL summand in (8.87) is equal to the kth row of the GEE summand in (8.88), namely

$$\{\partial/\partial\xi_k \boldsymbol{v}_i\}^T \boldsymbol{Z}_i^{-1}(\boldsymbol{u}_i - \boldsymbol{v}_i),$$

as desired. Of course, it is fact possible to carry out the argument in the reverse direction, starting from (8.88).

The same type of argument can be applied to the second term in the quadratic estimating equation for β, so that the joint normal ML equations may be written in the "GEE-2" form with the Gaussian working assumption.