# 5 Population-Averaged Linear Models for Continuous Response

## 5.1 Introduction

We begin our discussion of modern models and methods for longitudinal data analysis by considering a general class of models and associated methods for ***continuous response*** that arises from taking a ***population-averaged*** perspective. This class of models addresses all of the ***drawbacks*** of classical models and methods summarized in Section 4.2.

Namely, models of this type do not require the data set to be ***balanced***; i.e., the elements of the response vectors $\boldsymbol{Y}_i$ ***do not*** need to be observations taken at the the same $n$ time points. In addition, the model framework allows a very general specification for the form of the ***overall aggregate covariance matrix*** of a data vector and allows it to differ depending on, for instance, the values of ***covariates***.

The ***population mean response*** is represented by a ***linear model*** that allows ***among-*** and ***within-individual*** covariates to be incorporated straightforwardly and involves ***parameters*** that characterize features of the population mean response, such as ***patterns of change*** exhibited over time, and how these features might be associated with ***among-individual covariates***.

Finally, although the model incorporates an assumption of ***multivariate normality*** of a response vector ***conditional on covariates***, using ***large sample*** (large $m$) arguments, as long as $m$ is large enough and the model for the population mean response is ***correctly specified*** , it is possible to show that ***estimators*** of parameters in the models are ***consistent*** for the true values and to deduce an approximate ***normal sampling distribution*** for them, even if the true distribution of the response is ***not normal***. The approximate sampling distribution then forms the basis for ***inferential goals*** such as assessments of uncertainty and hypothesis testing procedures.

Moreover, as we demonstrate, ***even if*** the representation for the overall pattern of covariance is ***not correctly specified*** , estimators for parameters in a correctly specified population mean response model are ***still consistent*** , and an ***approximate sampling distribution*** can be derived.

## 5.2 Model specification

***BASIC MODEL:*** Recall again that the observed 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 $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i^T)^T$.

The ***population-averaged linear model*** we study in this chapter is most relevant when the responses $Y_{ij}$ are ***continuous***. The model is written as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \epsilon_i, \quad i = 1, \ldots, m. \tag{5.1}$$

- In (5.1), $\boldsymbol{X}_i$ is a ***design matrix*** for individual $i$ depending on individual $i$'s ***covariates*** $\boldsymbol{x}_i$, examples of which we present momentarily.
- The ***deviation*** $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$ is such that

  $$E(\epsilon_i|\boldsymbol{x}_i) = \boldsymbol{0}, \quad \text{var}(\epsilon_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i), \tag{5.2}$$

  where $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ $(n_i \times n_i)$ can depend on the covariates $\boldsymbol{x}_i$ and on a vector of ***covariance parameters*** $\xi$, which includes ***correlation parameters*** $\alpha$ $(s \times 1)$ and ***variance parameters*** $\theta$ $(r \times 1)$. We discuss examples shortly. We sometimes suppress this dependence for brevity and simply write $\boldsymbol{V}_i$.
- The form of $\boldsymbol{V}_i$ is specified ***by the data analyst*** in accordance with the features of the given situation. Because of the dependence of $\boldsymbol{V}_i$ on covariates, there is no requirement, for example, that the form of the covariance matrix be the same for all individuals. We elaborate on this point in the examples below.
- Ordinarily, it is assumed that the ***conditional distribution*** of $\epsilon_i$ given $\boldsymbol{x}_i$ is ***multivariate normal*** ,

  $$\epsilon_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{0}, \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)\}, \tag{5.3}$$

  sometimes written more briefly as $\epsilon_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_i)$.

- $\boldsymbol{\beta}$ is a vector of ***parameters*** characterizing the population mean response; that is, with the assumption on $\epsilon_i$ in (5.2), we have that

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\boldsymbol{\beta} \quad (n_i \times 1), \tag{5.4}$$

  representing the population mean response for individual *i*, or indeed any individual in the population with covariates $\boldsymbol{x}_i$.

- From (5.2), it follows that

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i) \quad (n_i \times n_i), \tag{5.5}$$

  the overall population covariance matrix for an individual with covariates $\boldsymbol{x}_i$, characterizing the ***aggregate pattern of covariance*** combining among- and within-individual sources for such an individual.

- With the normality assumption (5.3), the model can be written succinctly as

$$\boldsymbol{Y}_i|\boldsymbol{x}_i \sim \mathcal{N}\{\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i)\}, \quad i = 1, \ldots, m, \tag{5.6}$$

  which we often abbreviate as

$$\boldsymbol{Y}_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i), \quad i = 1, \ldots, m.$$

***REPRESENTATION OF COVARIANCE MATRIX:*** To facilitate thinking about models $\boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i)$, it is sometimes convenient to represent this covariance matrix as a ***product*** of "***standard deviation matrices***" and a ***correlation matrix***. Let $\boldsymbol{T}_i(\theta, \boldsymbol{x}_i)$ be the $(n_i \times n_i)$ ***diagonal matrix*** whose diagonal elements are models for $\text{var}(Y_{ij}|\boldsymbol{x}_i)$, depending on a parameter $\theta$ as above. Let $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ be a $(n_i \times n_i)$ correlation matrix, depending on a parameter $\alpha$. Then it is straightforward to deduce (try it) that a model for the overall covariance structure can be obtained as

$$\boldsymbol{V}_i(\boldsymbol{\xi}, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i), \quad \boldsymbol{\xi} = (\theta^T, \alpha^T)^T, \tag{5.7}$$

where $\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)$ is the matrix whose diagonal elements are the models for the ***standard deviations*** $\{\text{var}(Y_{ij}|\boldsymbol{x}_i)\}^{1/2}$. Clearly, $\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i) = \boldsymbol{T}_i(\theta, \boldsymbol{x}_i)$. We sometimes write $\boldsymbol{T}_i$ and $\boldsymbol{\Gamma}_i$ for brevity, suppressing dependence on $\theta$, $\alpha$, and $\boldsymbol{x}_i$.

The representation (5.7) allow features of overall variance and the overall pattern of correlation to be thought of ***separately***. That is, one can entertain models for correlation structure and beliefs about variance separately to arrive at an overall specification. We demonstrate in examples below.

***MODEL SUMMARY:*** It is often convenient to summarize the model as follows. Recall that the total number of observations $Y_{ij}$ is $N = \sum_{i=1}^{m} n_i$. Define

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix} \quad (N \times 1), \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix} \quad (N \times p), \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{pmatrix} \quad (N \times 1). \tag{5.8}$$

Then (5.1) can be expressed compactly as (try it)

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}. \tag{5.9}$$

It follows from (5.2) that

$$E(\boldsymbol{\epsilon}|\tilde{\boldsymbol{x}}) = \boldsymbol{0},$$

where $\tilde{\boldsymbol{x}}$ is the collection of all covariates $\boldsymbol{x}_i$, $i = 1 \ldots, m$, for all $m$ individuals, so that, from (5.4),

$$E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta.$$

Define the ***block diagonal*** matrix

$$\boldsymbol{V}(\xi, \tilde{\boldsymbol{x}}) = \begin{pmatrix} \boldsymbol{V}_1(\xi, \boldsymbol{x}_1) & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_2(\xi, \boldsymbol{x}_2) & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{V}_m(\xi, \boldsymbol{x}_m) \end{pmatrix} \quad (N \times N). \tag{5.10}$$

We often write (5.10) for brevity as

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{V}_m \end{pmatrix} \quad (N \times N). \tag{5.11}$$

Then, from (5.2) and (5.5), defining similarly $\boldsymbol{T}(\theta, \tilde{\boldsymbol{x}}) = \boldsymbol{T}$ and $\boldsymbol{\Gamma}(\alpha, \tilde{\boldsymbol{x}}) = \boldsymbol{\Gamma}$ $(N \times N)$,

$$\text{var}(\boldsymbol{\epsilon}|\tilde{\boldsymbol{x}}) = \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V}(\xi, \tilde{\boldsymbol{x}}) = \boldsymbol{V} = \boldsymbol{T}^{1/2}\boldsymbol{\Gamma}\boldsymbol{T}^{1/2},$$

which follows by the ***independence*** of $\boldsymbol{\epsilon}_i$ (and $\boldsymbol{Y}_i$) for $i = 1, \ldots, m$.

Note that $\boldsymbol{V}$ in (5.11) has a different definition from that in Chapter 3. Henceforth, we use the symbol $\boldsymbol{V}$ in this way to represent the covariance matrix of the "***stacked***" random vectors $\boldsymbol{\epsilon}$ and $\boldsymbol{Y}$ (conditional on the $\boldsymbol{x}_i$).

The model (5.6) can then be summarized by

$$\boldsymbol{Y}|\tilde{\boldsymbol{x}} \sim \mathcal{N}\{\boldsymbol{X}\beta, \boldsymbol{V}(\xi, \tilde{\boldsymbol{x}})\}, \tag{5.12}$$

or, briefly,

$$\boldsymbol{Y}|\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{X}\beta, \boldsymbol{V}). \tag{5.13}$$

- In the literature on longitudinal data analysis and in ***software documentation***, it is common to write the model incorporating the normality assumption using this "stacked" notation, suppressing dependence of the overall covariance matrix on parameters and covariate information; that is, as in (5.13).
- We use the more detailed notation (5.12) when we wish to emphasize ***explicitly*** the dependence of the covariance matrix on parameters and covariates.

***REMARK:*** Recall that $\boldsymbol{x}_i$ for individual $i$ includes the ***times*** $t_{ij}$, $= 1, \ldots, n_i$, at which $i$ was observed, which, technically, are not "***covariates***" in the strict sense, although they often play the role of "***covariates***" as far as implementation is concerned. Thus, conditioning on $\boldsymbol{x}_i$ is really meant to imply conditioning on all ***among-*** and ***within-individual covariates***.

We now demonstrate features of the ***population-averaged linear model*** and its interpretation by considering its specification in several examples.

***EXAMPLE 1, DENTAL STUDY:*** We have already considered a population-averaged model for these data in Section 2.4. Recall that there is one among-individual covariate, gender, which we represented for child $i$ as $g_i = 0$ if $i$ is a girl and $g_i = 1$ if $i$ is a boy, so that $\boldsymbol{a}_i = g_i$; there are no within-individual covariates $\boldsymbol{u}_i$. The response was measured for all $m = 27$ children at ages $(t_1, \ldots, t_4) = (8, 10, 12, 14)$. Thus, $\boldsymbol{z}_{ij} = t_j$ for all $i$, and $\boldsymbol{x}_i$ contains $g_i$ (and the four time points). Thus, conditioning on covariates $\boldsymbol{x}_i$ corresponds to conditioning on gender.

From a ***population-averaged*** perspective, the primary question of interest is whether or not the ***rate of change*** of the population mean response profile for boys differs from that for girls. In (2.22), we specified a model for the population mean at time $t_{ij}$ for a child of gender $g_i$ as

$$E(Y_{ij}|\boldsymbol{x}_i) = \{\beta_{0,B} g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B} g_i + \beta_{1,G}(1 - g_i)\} t_{ij}, \tag{5.14}$$

so that $\beta_{1,G}$ and $\beta_{1,B}$ are the ***slopes*** of the ***assumed straight line*** population mean profiles for girls and boys, respectively. Interest is in comparing $\beta_{1,G}$ and $\beta_{1,B}$.

Thus,

$$\boldsymbol{\beta} = (\beta_{0,G}, \beta_{1,G}, \beta_{0,B}, \beta_{1,B})^T,$$

$p = 4$, and, for child $i$,

$$\boldsymbol{X}_i = \begin{pmatrix} (1-g_i) & (1-g_i)t_1 & g_i & g_i t_1 \\ \vdots & \vdots & \vdots & \vdots \\ (1-g_i) & (1-g_i)t_4 & g_i & g_i t_4 \end{pmatrix}, \tag{5.15}$$

so that

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & t_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_4 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} 0 & 0 & 1 & t_1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & t_4 \end{pmatrix}, \tag{5.16}$$

if $i$ is a girl or boy, respectively.

Clearly, $\boldsymbol{X}_i$ in (5.15) is ***not of full rank*** for any $i$. Intuitively, this reflects that fact that a boy does not provide information on parameters describing the population mean for girls, and vice versa. However, it is straightforward to observe that the "***stacked design matrix***" $\boldsymbol{X}$ in (5.8), has ***full column rank*** $p = 4$, as it comprises 11 matrices $\boldsymbol{X}_i$ like that on the left hand side of (5.16) stacked on top of 16 like that on the right hand side; the $p = 4$ columns of $\boldsymbol{X}$ are clearly ***linearly independent*** (check). This demonstrates that the problem of making inference on $\beta$ is feasible from data like those in the study, involving children of both genders.

To complete the model, we specify a model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ for the ***overall pattern of covariance*** $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$. Because these data are ***balanced***, it was straightforward to calculate the sample overall covariance matrices and their associated correlation matrices for each gender in Section 2.6. Recall that the numerical estimates in (2.33) and (2.34) suggest the following:

- Overall variance is likely ***constant over time*** for each gender, but the variance estimates are ***larger*** for boys than for girls. Formally, the data suggest that $\text{var}(Y_{ij}|\boldsymbol{x}_i)$ for boys and girls are ***the same*** for all $j$ but that for boys is larger. Thus, recognizing that conditioning on $\boldsymbol{x}_i$ is really conditioning on $g_i$, a reasonable model is

  $$\text{var}(Y_{ij}|g_i = 0) = \sigma_G^2, \quad \text{var}(Y_{ij}|g_i = 1) = \sigma_B^2. \tag{5.17}$$

  The specification in (5.17) can be represented by taking $\boldsymbol{T}_i(\theta, \boldsymbol{x}_i)$ to be the diagonal matrix with diagonal elements all equal to

  $$\sigma_G^2(1-g_i) + g_i\sigma_B^2,$$

  or, equivalently,

  $$\boldsymbol{T}_i(\theta, \boldsymbol{x}_i) = \{\sigma_G^2(1-g_i) + g_i\sigma_B^2\}\boldsymbol{I}_4, \quad \theta = (\sigma_G^2, \sigma_B^2)^T.$$

- The ages are ***equally-spaced*** in time, so any model that is reasonable under this condition is possible. The empirical evidence suggests that, for each gender, the overall pattern of correlation is approximately ***compound symmetric*** with a ***different*** correlation parameter $\alpha$ in (2.25) for each gender. That is,

$$\boldsymbol{\Gamma}_i(\boldsymbol{\alpha}, \boldsymbol{x}_i) = [1 - \{(1 - g_i)\alpha_G + g_i\alpha_B\}]\boldsymbol{I}_4 + \{(1 - g_i)\alpha_G + g_i\alpha_B\}\boldsymbol{J}_4,$$

  where thus $\boldsymbol{\alpha} = (\alpha_G, \alpha_B)^T$.

Combining the above, the suggested covariance model is

$$\sigma^2_G \begin{pmatrix} 1 & \alpha_G & \cdots & \alpha_G \\ \alpha_G & 1 & \cdots & \alpha_G \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_G & \cdots & \alpha_G & 1 \end{pmatrix} = \sigma^2_G\{(1 - \alpha_G)\boldsymbol{I}_4 + \alpha_G\boldsymbol{J}_4\} \tag{5.18}$$

for girls, and

$$\sigma^2_B\{(1 - \alpha_B)\boldsymbol{I}_4 + \alpha_B\boldsymbol{J}_4\} \tag{5.19}$$

for boys. The ***covariance parameter*** $\boldsymbol{\xi}$ characterizing $\boldsymbol{V}_i$ is then $\boldsymbol{\xi} = (\sigma^2_G, \sigma^2_B, \alpha_G, \alpha_B)^T$.

***ALTERNATIVE PARAMETERIZATION:*** As with any ***linear model*** , it is possible to represent the population mean response model (5.14) using a ***different parameterization***. Because interest focuses on the ***difference in slopes*** characterizing the rates of change of population mean dental distance for boys and girls, it is natural to express the population mean ***directly*** in terms of a parameter representing this difference. Thus, an equivalent alternative to (5.14) is

$$E(Y_{ij}|\boldsymbol{x}_i) = \{\beta_{0,G} + \beta_{0,B-G}g_i\} + \{\beta_{1,G} + \beta_{1,B-G}g_i\}t_{ij}. \tag{5.20}$$

In (5.20), $\beta_{0,B-G}$ and $\beta_{1,B-G}$ represent the ***difference*** in intercept and slope between boys and girls and will be positive if that for boys exceeds that for girls. Moreover, for example, the slope of the population mean response for boys is then $\beta_{1,B-G} + \beta_G$, and similarly for intercept.

***REMARK:*** The population mean response model in (5.14) or (5.20) in no way requires the time points $t_{ij}$ to be the same for each child. Even if these data were ***not balanced*** , there would be ***no problem*** specifying such a model. Specification of the covariance model when data are not balanced does require some special consideration; we discuss this shortly.

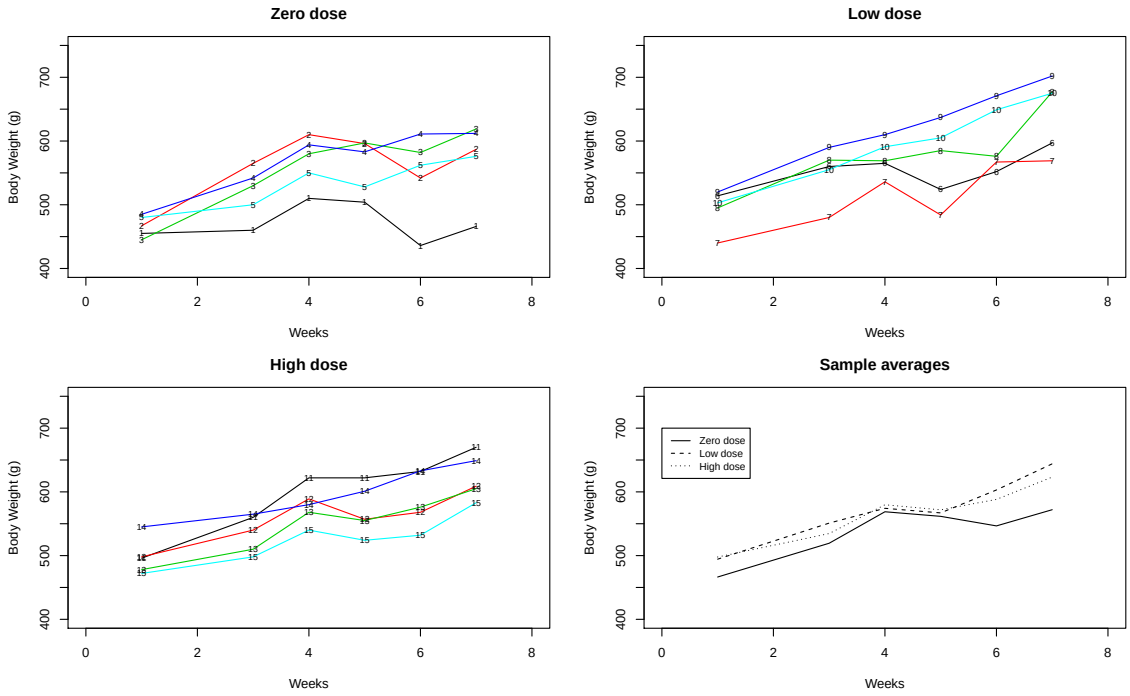


Figure 5.1: *Growth of guinea pigs receiving different doses of vitamin E diet supplement.*

***EXAMPLE 2, GUINEA PIG DIET STUDY:*** The same considerations apply to specification of a population-averaged model for these data, which are also ***balanced***. We discuss specification of a model for population mean response, which illustrates some key issues.

Recall from Section 1.2 that 15 guinea pigs were given a growth-inhibiting substance at ***baseline*** (time 0, beginning of the first week). At weeks 1, 3, and 4, body weight we measured. Immediately after the week 4 measurement (so at the start of week 5), the pigs were ***randomized*** to receive zero, low, or high dose of vitamin E, 5 pigs per group, and body weight was subsequently recorded subsequently at weeks 5, 6, and 7. Thus all $m = 15$ pigs were observed at times $(t_1, \ldots, t_6) = (1, 3, 4, 5, 6, 7)$, so $\boldsymbol{z}_{ij} = t_j$ for all $i$, and $\boldsymbol{a}_i$ is the ***among-individual covariate*** dose group, with three possible values, which can be represented as $\boldsymbol{a}_i = (d_{i1}, d_{i2}, d_{i3})^T$, where $d_{i\ell} = 1$ if pig $i$ was randomized to dose group $\ell$ and $= 0$ otherwise, where $\ell = 1, 2, 3$ correspond to zero, low, and high dose.

We reproduce Figure 1.3 from Chapter 1 for convenience as Figure 5.1.

Because the pigs were treated ***identically*** until the end of week 4, a reasonable model for population mean response takes it to be ***identical*** for pigs in all three groups through week 4. Because pigs were then ***randomized*** at this time to receive one of the three doses, a model should allow the population mean response profile to be potentially ***different*** for each dose group henceforth.

That is, a plausible population mean model has two "***phases***," before and after introduction of vitamin E, where the second "***phase***" is different for each group.

From the plot of ***sample averages*** over time in Figure 5.1, a model that takes each of these phases to be a ***straight line*** is reasonable, where the intercept and slope of the first phase is the same for all groups. A model that incorporates these features is the ***linear spline*** model

$$E(Y_{ij}|\boldsymbol{x}_i) = \beta_0 + \beta_1 t_{ij} + \sum_{\ell=1}^{3} \beta_{2\ell} d_{i\ell}(t_{ij} - 4)_+ \tag{5.21}$$

with a ***knot*** at week 4, where

$$\begin{aligned} x_+ &= x \quad \text{if } x \geq 0, \\ &= 0 \quad \text{if } x < 0. \end{aligned}$$

From (5.21), for any pig, population mean response follows the straight line

$$\beta_0 + \beta_1 t$$

through week $t = 4$. For $t \geq 4$, for a pig in group $\ell$, population mean response is represented as

$$\beta_0 + \beta_1 t_{ij} + \beta_{2\ell}(t - 4) = \{\beta_0 + \beta_1(4)\} + (\beta_1 + \beta_{2\ell})(t - 4),$$

so that, with $t = 4$ as the "***origin***," population mean weight follows a straight line for $t \geq 4$ with "***intercept***" (value at $t = 4$ when the dose was administered) $\beta_0 + \beta_1(4)$ and ***slope*** $\beta_1 + \beta_{2\ell}$.

Differences in population mean response trajectory are reflected in (5.21) by differences among the $\beta_{2\ell}$, $\ell = 1, 2, 3$. The model (5.21) could of course be parameterized in ***alternative*** ways. The model allows the possibility that the population mean profile for the zero dose group ***changes*** after week 4, even though the pigs in this group did not receive vitamin E. If there were reason to believe that the population mean trajectory for pigs not receiving vitamin E before week 4 should ***continue*** after week 4, a modification of the model would be to take $\beta_{21} = 0$ in (5.21); however, the visual evidence in Figure 1.3 does not support this. Perhaps the effect of the growth-inhibiting substance begins to manifest at week 4, leading to a downward trend, but the addition of vitamin E mitigates this effect.

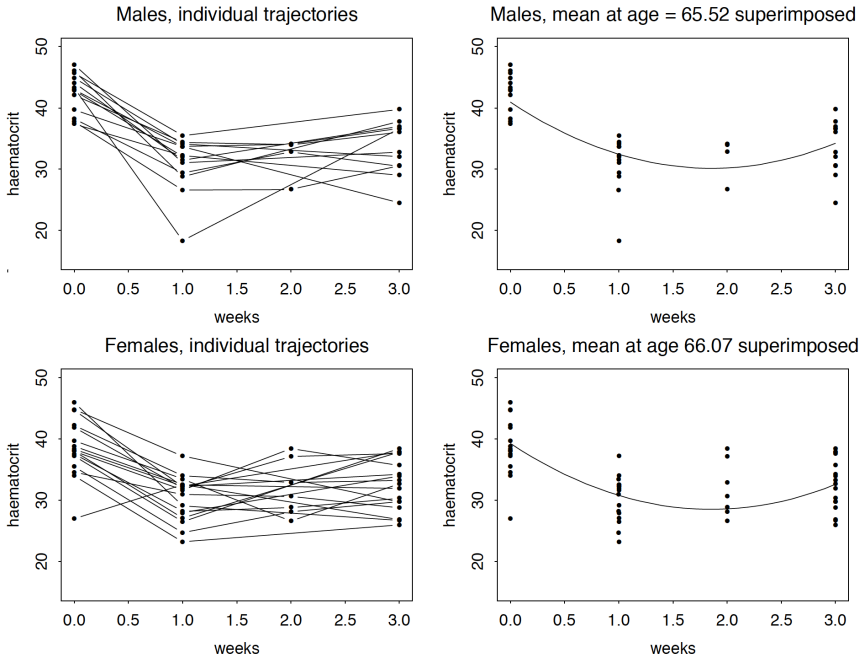


Figure 5.2: *Hæmatocrit trajectories for hip replacement patients. The left hand panels show individual profiles by gender; the right hand panels show a fitted quadratic model for the mean superimposed.*

***HIP REPLACEMENT STUDY:*** These data are adapted from Crowder and Hand (1990, Section 5.2). Thirty patients underwent hip replacement surgery, 13 males and 15 females. Hæmatocrit, the ratio of volume packed red blood cells relative to volume of whole blood recorded as a percentage, was planned to be measured on each patient at ***baseline***, week 0, prior to surgery, and then at weeks 1, 2, and 3 post-surgery. In addition to gender, ***age*** of each patient was also recorded. The data are shown in Figure 5.2.

The primary objectives are to determine if there are differences in the ***population mean pattern of change*** of hæmatocrit following surgery between genders and to characterize the patterns of change.

It is evident in the left hand panels of Figure 5.2 that several patients of both genders are ***missing*** the measurement at week 2; there is also female who is missing both this and the baseline measurement. Crowder and Hand do not offer an explanation; because this is so ***systematic***, occurring for about half of the male and half of the female patients, it is plausible that these observations are missing for reasons having ***nothing*** to do with the health status of the patients but rather might reflect, for example, failure of the equipment used ascertain hæmatocrit values during week 2. We downplay this complication for now and return to the issue of ***missing responses*** later in this chapter.

These data exemplify the common situation where, although it was ***planned*** to record the response at $n = 4$ prespecified times (0,1,2,3 weeks), not all individuals have all responses recorded, so that $n_i$ varies with $i$, although those that are available are at the prespecified times. That is, $n_i = 4$ for some patients, for whom $t_{ij} = 0, 1, 2, 3$ for $j = 1, \ldots, 4$; $n_i = 3$ for those missing the week 2 measurement, so that $t_{ij} = 0, 1, 3$, $j = 1, \ldots, 3$; and $n_i = 2$ for the female patient missing the baseline and week 2 responses, so that $t_{ij} = 1, 3$, $j = 1, 2$. For patient $i$, $\boldsymbol{z}_{ij} = t_{ij}$, $j = 1, \ldots, n_i$, and $\boldsymbol{a}_i = (g_i, a_i)^T$, where gender $g_i = 0$ for females and $g_i = 1$ for males; and $a_i$ is the age of the patient (years), ranging from 47 to 79 for females (sample average 66.07) and 44 to 74 for males (65.52).

For both genders, Figure 5.2 shows that hæmatocrit ***drops*** from baseline after surgery and then begins to ***rebound*** over the 3 weeks post-surgery. This suggests that the following ***quadratic*** model for population mean is reasonable, which allows the pattern to ***differ*** between genders:

$$E(Y_{ij}|\boldsymbol{x}_i) = \{\beta_{0,F}(1 - g_i) + \beta_{0,M}g_i\} + \{\beta_{1,F}(1 - g_i) + \beta_{1,M}g_i\}t_{ij} + \{\beta_{2,F}(1 - g_i) + \beta_{2,M}g_i\}t_{ij}^2. \quad (5.22)$$

The basic model (5.23) can be modified to incorporate the possibility that features of the mean response are ***age-dependent*** ; for example,

$$\begin{aligned} E(Y_{ij}|\boldsymbol{x}_i) &= \{\beta_{0,F}(1 - g_i) + \beta_{0,M}g_i\} + \{\beta_{3,F}(1 - g_i) + \beta_{3,M}g_i\}a_i \\ &\quad + \{\beta_{1,F}(1 - g_i) + \beta_{1,M}g_i\}t_{ij} + \{\beta_{2,F}(1 - g_i) + \beta_{2,M}g_i\}t_{ij}^2. \end{aligned} \quad (5.23)$$

allows ***mean hæmatocrit at baseline*** depend on patient age of patient in a way that is different for each gender. The ***linear*** and ***quadratic*** effects that govern the pattern of change post-baseline could be modified similarly, and any of these models could be***reparameterized*** in terms of parameters representing the ***differences*** in intercept and linear and quadratic effects between genders.

Plausible models $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ for the ***overall pattern of covariance*** include those that are suited to what are ideally ***balanced*** data with ***equally-spaced*** time points; however, fitting of such models requires that the ***missing values*** for some patients be taken into account appropriately. We discuss this shortly.

***HIV CLINICAL TRIAL:*** These data are reported in Fitzmaurice, Laird, and Ware (2011) and are from a randomized, double-blind clinical trial, AIDS Clinical Trials Group (ACTG) Study 193A, in patients infected with ***human immunodeficiency virus*** (HIV) exhibiting advanced immune suppression; i.e., CD4 T-cell counts $\leq$ 50 cells/mm$^3$. CD4 count is a standard measure reflecting the status of the ***immune system*** , which is compromised in patients with HIV infection.

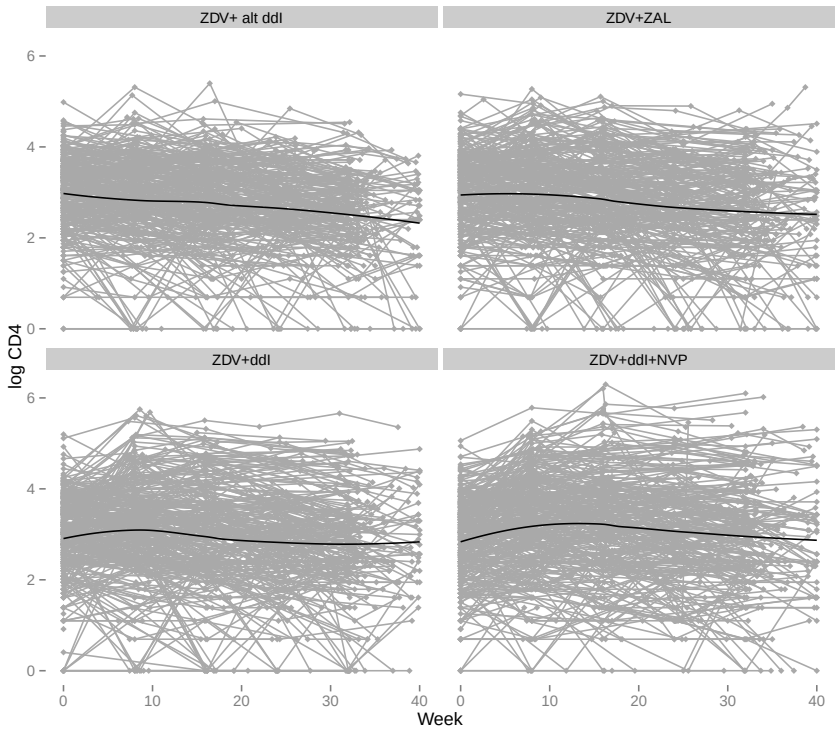


Figure 5.3: *log(CD4+1) profiles for subjects in ACTG Study 193A. A loess smoother fitted to all the data for each treatment is superimposed on the individual profiles in each panel.*

1313 subjects were ***randomized*** to one of four daily treatment regimens consisting of dual or triple combinations of drugs in the class of HIV-1 reverse transcriptase inhibitors: (1) 600 mg of zidovudine (ZDV) alternating monthly with 400 mg of didanosine (ddI); (2) 600 mg ZDV plus 2.25 mg zalcitabine (ZAL); (3) 600 mg ZDV plus 400 mg ddI; or (4) 600 mg ZDV plus 400 mg ddI and 400 mg nevirapine (NVP) (triple therapy).

CD4 measurements were planned at baseline (week 0) and then at 8-week intervals during follow-up, at weeks 8, 16, 24, 32, and 40. Figure 5.3 shows the individual ***log-transformed*** CD4 profiles for subjects randomized to each treatment regimen; because CD4 count of zero is possible, it is customary to take the response variable to be log(CD4+1) (transformed CD4 counts appear ***approximately normally distributed***). As can be seen from the plots, ***actual visits*** did not necessarily take place at ***exactly*** these times; moreover, some subjects ***skipped visits*** altogether or ***dropped out*** of the study before 40 weeks.

For example, visit times for the first subject in the ZDV+ZAL group were $t_{ij}$ = 0, 7.6, 15.6, 23.6, 32.6, and 40 weeks; the first subject in the ZDV+ddI group had actual visits at $t_{ij}$ = 0, 7.1, 16.1, and 32.4 weeks. The number of CD4 measurements per subject ranged from 1 to 9, with a median of 4.

An approximation to addressing this issue would be to "***bin***" actual visit times to correspond to the intended times, so that, for example, 7.6 and 7.1 weeks would be rounded to 8 weeks. However, as discussed in Section 4.2, treating all responses within some interval of an intended visit time as if they were all observed at that time is ***ad hoc***, with unknown effects on inference. If the actual visit times are available, clearly it is ***preferable*** to incorporate them in an analysis.

In addition to treatment regimen, also recorded for each subject is age (years) and gender; thus, the ***among-individual covariates*** are $\boldsymbol{a}_i = (g_i, a_i, \delta_{i1}, \ldots, \delta_{i4})^T$, where $g_i = 0$ (1) for a female (male) subject; $a_i$ is age; and $\delta_{i\ell} = 1$ if subject $i$ was randomized to treatment regimen $\ell$ and 0 otherwise, $\ell = 1, \ldots, 4$.

A ***local polynomial regression*** (***loess***) curve naïvely fitted to all the data for each treatment is superimposed on each panel in Figure 5.3 as suggested in Section 2.6 to give a rough idea of the overall population mean trend. The visual evidence suggests that a ***straight line*** might provide a reasonable representation of the overall population mean response in each group, although the triple therapy group shows a subtle rise followed by a decay, which might be better captured by a quadratic. Downplaying this for now, a simple model that allows a separate, straight line mean trajectory for each treatment is

$$E(Y_{ij}|\boldsymbol{x}_i) = \beta_0 + \{\beta_{14} + \beta_{11}\delta_{i1} + \beta_{12}\delta_{i2} + \beta_{13}\delta_{i3}\}t_{ij}. \tag{5.24}$$

In (5.24), the intercept is taken to be ***the same*** for all regimens; because subjects were ***randomized*** to the four regimens, the mean response at ***baseline*** (week 0), prior to the start of treatment, should be ***identical*** for all regimens, assuming that the randomization was carried out faithfully. Indeed, the ***sample averages*** of log-transformed CD4 at baseline are 2.98, 2.93, 2.91, and 2.84 for subjects randomized to regimens $1 - 4$.

We have parameterized the slope term in braces so that the triple therapy regimen 4 is the ***reference*** regimen. That is, $\beta_{14}$ is the slope for the mean CD4 profile for regimen 4, and $\beta_{14} + \beta_{1\ell}$ is the slope for regimen $\ell = 1, 2, 3$, so that $\beta_{1\ell}$, $\ell = 1, 2, 3$ represents the difference in slope relative to triple therapy. Of course, an ***alternative parameterization*** in terms of separate slopes for each regimen is possible; likewise, allowing for ***separate intercepts*** would allow investigation of the integrity of the randomization. Model (5.24) could also be modified to incorporate dependence of intercept and slope on age and gender or to allow quadratic effects.

Specification of a covariance model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ requires some care. Because all individuals were seen at potentially different times, with different numbers of visits, models for ***balanced*** and ***equally-spaced*** data might not be suitable.

***CONSIDERATIONS FOR COVARIANCE MODELS, BALANCED DATA:*** When the data are ***balanced*** , as for the dental study, as discussed in Section 2.6, inspection of sample covariance and correlation matrices, scatterplot matrices, autocorrelation functions, and lag plots can assist the analyst in identifying plausible models.

In fact, these approaches can be refined to take into account a postulated population mean model so as to take advantage of the belief that the mean follows a ***smooth trajectory***. Instead of basing these diagnostic aids on ***sample means*** at each time point, one can instead estimate those means by a ***preliminary fit*** of the mean model using ***ordinary least squares*** , treating the observations from all individuals as if they are all ***mutually independent***. Although this sounds suspect, as we discuss in Section 5.5, if the mean model is ***correctly specified*** in the sense defined in Section 4.3, then the OLS estimator for $\beta$ in the overall population mean model $\boldsymbol{X}\beta$ is ***consistent*** for the true value $\beta_0$. Thus, at least for $m$ "large," using the ***predicted values*** from the OLS fit to estimate the population means should be reasonable.

***CONSIDERATIONS FOR COVARIANCE MODELS, DATA NOT BALANCED:*** When a longitudinal data set is ***not balanced*** , not only is it more difficult to think about plausible covariance models, ***more ominously*** , if the intention was to record the response at the same ***prespecified times*** for all individuals, but some observations are ***missing*** for some individuals, then things become more complicated. In Section 5.6, we discuss the challenges associated with such ***missing data*** and the assumptions that must be fulfilled to enable ***valid inferences*** to be drawn using the models and methods in this and the next chapter.

For now, we limit our discussion to ***operational issues*** associated with specifying a covariance structure in this situation. Consider the hip replacement study, where the times of observation are the ***same*** for all individuals except that some individuals are ***missing*** the response at some of these times.

Recall that, as discussed in Section 1.3, our notational convention is that $\boldsymbol{Y}_i$ is the $(n_i \times 1)$ vector of responses ***actually observed and recorded*** at times $t_{i1}, \ldots, t_{in_i}$ on individual $i$. Let

$$\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{in})^T \tag{5.25}$$

be the $(n \times 1)$ vector of ***intended*** responses to be collected at times $t_1, \ldots, t_n$, where $n \geq n_i$ for all $i = 1, \ldots, m$. In the literature on ***missing data*** methods, $\mathcal{Z}_i$ is referred to as the ***full data*** for subject $i$; see Section 5.6.

- Clearly, for an individual for whom all intended responses are observed, $n_i = n$ and $\boldsymbol{Y}_i = \mathcal{Z}_i$. Thus, $\boldsymbol{V}_i$ for such a individual is a model for $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \text{var}(\mathcal{Z}_i|\boldsymbol{x}_i)$.

- For an individual with some components of $\mathcal{Z}_i$ ***not observed*** (missing), we can make a correspondence as follows. Consider the hip replacement study, where $n = 4$,

$$\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{i4})^T, \quad (t_1, \ldots, t_4) = (0, 1, 2, 3).$$

  Consider an individual who is missing the intended observation at $t_3 = 2$ weeks. Then

$$\boldsymbol{Y}_i = (Z_{i1}, Z_{i2}, Z_{i4})^T \quad \text{at times} \quad (t_{i1}, t_{i2}, t_{i3}) = (t_1, t_2, t_4) = (0, 1, 3). \tag{5.26}$$

- Here, $\boldsymbol{V}_i$ is a model for the covariance matrix of $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ as in (5.26), namely, for

$$\begin{pmatrix} \text{var}(Z_{i1}|\boldsymbol{x}_i) & \text{cov}(Z_{i1}, Z_{i2}|\boldsymbol{x}_i) & \text{cov}(Z_{i1}, Z_{i4}|\boldsymbol{x}_i) \\ \text{cov}(Z_{i2}, Z_{i1}|\boldsymbol{x}_i) & \text{var}(Z_{i2}|\boldsymbol{x}_i) & \text{cov}(Z_{i2}, Z_{i4}|\boldsymbol{x}_i) \\ \text{cov}(Z_{i4}, Z_{i1}|\boldsymbol{x}_i) & \text{cov}(Z_{i4}, Z_{i2}|\boldsymbol{x}_i) & \text{var}(Z_{i4}|\boldsymbol{x}_i) \end{pmatrix}, \tag{5.27}$$

  which we can write equivalently as in (5.7) as

$$\boldsymbol{V}_i = \boldsymbol{T}_i^{1/2} \boldsymbol{\Gamma}_i \boldsymbol{T}_i^{1/2}, \tag{5.28}$$

  where $\boldsymbol{T}_i = \text{diag}\{\text{var}(Z_{i1}|\boldsymbol{x}_i), \text{var}(Z_{i2}|\boldsymbol{x}_i), \text{var}(Z_{i4}|\boldsymbol{x}_i)\}$, and

$$\boldsymbol{\Gamma}_i = \begin{pmatrix} 1 & \text{corr}(Z_{i1}, Z_{i2}|\boldsymbol{x}_i) & \text{corr}(Z_{i1}, Z_{i4}|\boldsymbol{x}_i) \\ \text{corr}(Z_{i2}, Z_{i1}|\boldsymbol{x}_i) & 1 & \text{corr}(Z_{i2}, Z_{i4}|\boldsymbol{x}_i) \\ \text{corr}(Z_{i4}, Z_{i1}|\boldsymbol{x}_i) & \text{corr}(Z_{i4}, Z_{i2}|\boldsymbol{x}_i) & 1 \end{pmatrix}.$$

It should be clear from (5.27) that there is no conceptual problem in positing an ***unstructured*** covariance matrix under these circumstances; the only caveat is that some ***bookkeeping*** is necessary to establish the correspondence between observed and intended time points.

Similarly, specification of a ***compound symmetric*** correlation structure is not problematic, as correlation between any two elements of $\mathcal{Z}_i$, and thus $\boldsymbol{Y}_i$ (given $\boldsymbol{x}_i$) is ***the same*** under this model.

Here, the ***intended time points*** are equally-spaced, so that the ***one-dependent*** model in (2.27) and the ***AR(1)*** model in (2.28) are also candidates.

It is straightforward to see that the one-dependent model for the situation in (5.27) takes $\mathbf{\Gamma}_i$ in (5.28) to be

$$\begin{pmatrix} 1 & \alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix};$$

and the corresponding AR(1) model is

$$\begin{pmatrix} 1 & \alpha & \alpha^3 \\ \alpha & 1 & \alpha^2 \\ \alpha^3 & \alpha^2 & 1 \end{pmatrix}$$

(check). ***Software packages*** for fitting population-averaged linear models using the methods discussed in the next section incorporate the appropriate bookkeeping for this situation.

In a situation like the HIV clinical trial in ACTG Study 193A, things are more complex. In this study, we can still conceive of the ***full data*** that were intended to be collected on each subject; that is, the vector of ***intended responses*** $\mathcal{Z}_i$ as in (5.25) at prespecified times $(t_1, \ldots, t_n)$.

Strictly speaking, however, each individual $i$ is seen at potentially ***different time points***, so that, ***operationally***, the covariance models that can be feasibly entertained are ***limited***. For example, it is not possible to take the covariance matrix to be completely ***unstructured***, as individuals seen at different time points cannot share the same covariance parameters, so that the vector $\xi$ could be potentially different for each $i$ (and thus infeasible to ***estimate***).

Recognizing that the actual time points for most individuals ***target*** the intended, equally-spaced time points, models such as the compound symmetric, one-dependent, AR(1) might be ***reasonable approximations*** to the true covariance structure. Alternatively, if ***within-individual*** sources of correlation are pronounced, correlation models such as the ***exponential*** (2.31) or ***Gaussian*** (2.32), which depend on the distances between ***actual*** time points, are also feasible.

In Chapter 6, we discuss ***subject-specific*** linear models, for which a model for $\boldsymbol{V}_i$ is induced through specification of ***separate*** models for contributions to the overall covariance structure from ***within-*** and ***among-individual*** sources. This structure "***automatically***" addresses complications arising because of imbalance.

## 5.3 Maximum likelihood estimation under normality

Given a model specification

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

as in (5.4), (5.5), and (5.7), and using the ***independence*** of $(\boldsymbol{Y}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, it is possible to formulate ***estimating equations*** that can be solved to yield estimators for the mean parameters $\beta$ ($p \times 1$) and covariance parameters $\xi = (\theta^T, \alpha^T)^T$ ($r + s \times 1$).

***LOGLIKELIHOOD:*** Specifically, under the ***additional assumption*** that the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is ***multivariate normal*** as in (5.6) and using the independence across $i$, we can appeal to the principle of ***maximum likelihood*** to derive estimators for $\beta$ and $\xi$ as follows.

Writing the model succinctly as in (5.12), the ***joint density*** for $\boldsymbol{Y}$ conditional on $\tilde{\boldsymbol{x}}$ is

$$\begin{aligned} p(\boldsymbol{y}|\tilde{\boldsymbol{x}}; \beta, \xi) &= (2\pi)^{N/2}|\boldsymbol{V}(\xi, \tilde{\boldsymbol{x}})|^{-1/2} \exp\{-(\boldsymbol{y} - \boldsymbol{X}\beta)^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}})(\boldsymbol{y} - \boldsymbol{X}\beta)/2\} \\ &= \prod_{i=1}^{m}(2\pi)^{-n_i/2}|\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)|^{-1/2} \exp\{-(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)/2\}. \quad (5.29) \end{aligned}$$

It follows from (5.29) that the ***loglikelihood*** has the form, ignoring constants,

$$\begin{aligned} l(\beta, \xi) &= (-1/2)\Big\{ \log|\boldsymbol{V}(\xi, \tilde{\boldsymbol{x}})| + (\boldsymbol{Y} - \boldsymbol{X}\beta)^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}})(\boldsymbol{Y} - \boldsymbol{X}\beta) \Big\} \quad (5.30) \\ &= (-1/2)\sum_{i=1}^{m} \Big\{ \log|\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)| + (\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) \Big\} \quad (5.31) \end{aligned}$$

***ESTIMATING EQUATIONS:*** We appeal to standard matrix differentiation results summarized in Appendix A to derive the ***estimating equations*** (***score equations***) whose joint solution in $\beta$ and $\xi$ leads to the ***maximum likelihood estimators*** for these parameters ***under the assumption of multivariate normality***.

***Differentiating*** (5.30) and equivalently (5.31) with respect to $\beta$ ($p \times 1$) yields the estimating equation

$$\boldsymbol{X}^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}})(\boldsymbol{Y} - \boldsymbol{X}\beta) = \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{0}, \quad (5.32)$$

which follows (verify) using the following results in Appendix A:

- For $\boldsymbol{x}$ ($n \times 1$), symmetric ($n \times n$) matrix $\boldsymbol{A}$, and ***quadratic form*** $Q = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$, $\partial Q/\partial \boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}$ ($n \times 1$).
- If $\boldsymbol{x}$ depends on $\beta$ ($p \times 1$), the ***chain rule*** then gives $\partial Q/\partial \beta = (\partial \boldsymbol{x}/\partial \beta)(\partial Q/\partial \boldsymbol{x})$, where $(\partial \boldsymbol{x}/\partial \beta)$ is a ($p \times n$) matrix.

It is straightforward to observe the following.

- The estimating equation (5.32) can be rewritten as

$$\beta = \left\{ \boldsymbol{X}^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}}) \boldsymbol{X} \right\}^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1}(\xi, \tilde{\boldsymbol{x}}) \boldsymbol{Y} = \left( \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i) \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i) \boldsymbol{Y}_i. \tag{5.33}$$

  Thus, if $\boldsymbol{V}$ (equivalently $\boldsymbol{V}_i$, $i = 1, \ldots, m$) were ***known***; i.e., if $\xi$ were ***known***, then (5.33) defines explicitly an ***estimator*** for $\beta$.

  Of course, the covariance parameter $\xi$ is ordinarily ***not known*** and must be ***estimated***, which can be accomplished by solving another ***estimating equation*** discussed below, ***jointly*** with (5.32).

- If the model $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta$ is ***correctly specified***, then it is straightforward to observe that (5.32) is an ***unbiased estimating equation***.

  ***In fact***, even if the model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ is ***not*** a correct specification for $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, the estimating equation is ***still unbiased***. This suggests that, ***even if*** we have specified the covariance structure incorrectly, the maximum likelihood estimator for $\beta$ under normality will be ***consistent*** for the true value $\beta_0$ as long as the mean model is ***correctly specified***.

- Moreover, the foregoing observations hold ***whether or not*** the distribution of $\boldsymbol{Y}_i|\boldsymbol{x}_i$ is actually ***multivariate normal***.

Now consider ***differentiation*** of the loglikelihood (5.30) and equivalently (5.31) with respect to $\xi$ ($r+s\times 1$). This is again straightforward using the following matrix differentiation results from Appendix A. Let $\boldsymbol{V}(\xi)$ be a $(n \times n)$ nonsingular matrix depending on a vector $\xi$.

- If $\xi_k$ is the $k$th element of $\xi$, then $\partial/\partial\xi_k \, \boldsymbol{V}(\xi)$ is the $(n \times n)$ matrix whose $(\ell, p)$ element is the partial derivative of the $(\ell, p)$ element of $\boldsymbol{V}(\xi)$ with respect to $\xi_k$.
- $\partial/\partial\xi_k \{\log|\boldsymbol{V}(\xi)|\} = \text{tr}\left[\boldsymbol{V}^{-1}(\xi)\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\right]$, where $\text{tr}(\boldsymbol{A})$ is the ***trace*** of square matrix $\boldsymbol{A}$.
- $\partial/\partial\xi_k \boldsymbol{V}^{-1}(\xi) = -\boldsymbol{V}^{-1}(\xi)\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\boldsymbol{V}^{-1}(\xi)$.
- For quadratic form $Q = \boldsymbol{x}^T \boldsymbol{V}(\xi)\boldsymbol{x}$, $\partial Q/\partial\xi_k = \boldsymbol{x}^T\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\boldsymbol{x}$. Thus, from the previous result,

$$\partial/\partial\xi_k \{\boldsymbol{x}^T \boldsymbol{V}^{-1}(\xi)\boldsymbol{x}\} = -\boldsymbol{x}^T \boldsymbol{V}^{-1}(\xi)\{\partial/\partial\xi_k \boldsymbol{V}(\xi)\}\boldsymbol{V}^{-1}(\xi)\boldsymbol{x}.$$

Let $\xi_k$, $k = 1, \ldots, r+s$, be the $k$th (scalar) component of $\boldsymbol{\xi}$. Applying the foregoing results to differentiation of the loglikelihood (5.30) and equivalently (5.31) with respect to $\xi_k$, it can be verified (try it) that the result is the following set of $(r+s)$ ***estimating equations***:

$$(1/2)\left((\boldsymbol{Y}-\boldsymbol{X}\beta)^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\,\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})(\boldsymbol{Y}-\boldsymbol{X}\beta) - \text{tr}\left[\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\right]\right)=0,\quad k=1,\ldots,r+s. \tag{5.34}$$

or, equivalently,

$$(1/2)\sum_{i=1}^{m}\left((\boldsymbol{Y}_i-\boldsymbol{X}_i\beta)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta) - \text{tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\right]\right)=0,\quad k=1,\ldots,r+s. \tag{5.35}$$

Stacked, the $(r+s)$ estimating equations in (5.34) or (5.35) define ***implicitly*** the maximum likelihood estimator for the covariance parameter $\boldsymbol{\xi}$ ***under the assumption of normality***. In particular, the estimator is obtained by solving these equations ***jointly*** with the equations in (5.32).

We now demonstrate that these estimating equations are ***unbiased*** if the mean model $\boldsymbol{X}_i\beta$ ***and*** the covariance model $\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)$ are ***correctly specified***. Consider form of the equations in (5.35), where they are written as a sum over $i$ of ***independent*** quantities. It can be shown that the conditional (on $\boldsymbol{x}_i$) expectation of a summand in (5.35) is equal to zero by appealing to the following result:

- If $\boldsymbol{U}$ is a random vector with mean zero and covariance matrix $\boldsymbol{V}$, and $\boldsymbol{A}$ is a square matrix, then $E(\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}) = \text{tr}\{E(\boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{A}\} = \text{tr}(\boldsymbol{V}\boldsymbol{A}) = \text{tr}(\boldsymbol{A}\boldsymbol{V})$ (this is a special case of a more general result in Appendix A).

Using this, we have (assuming expectation is under the parameter values $\boldsymbol{\eta} = (\beta^T, \xi^T)^T$,

$$\begin{aligned} &E_\eta\left[(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta)^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta)\mid\boldsymbol{x}_i\right] \\ &\quad=\text{tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\right] \\ &\quad=\text{tr}\left[\boldsymbol{V}_i^{-1}(\boldsymbol{\xi},\boldsymbol{x}_i)\partial/\partial\xi_k\{\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)\}\right], \end{aligned} \tag{5.36}$$

from whence unbiasedness of (5.35) follows. Of course, if $\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)$ were ***incorrectly specified***, the equation is ***not*** necessarily unbiased.

As above, the argument to show that these estimating equations are unbiased ***does not require*** multivariate normality to hold; all that is necessary is that the ***first two moments*** of the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ are correctly specified.

***SUMMARY:*** The estimators for $\beta$ and $\xi$ in a model of the form

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

under the ***assumption*** that the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is ***multivariate normal*** with these moments are defined as the joint solution to the estimating equations

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{0}, \tag{5.37}$$

$$(1/2)\sum_{i=1}^{m} \Big( (\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\{\partial/\partial\xi_k \, \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i, \beta) - \text{tr}\left[\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\{\partial/\partial\xi_k \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)\}\right]\Big) = 0, \quad k = 1, \ldots, r+s, \tag{5.38}$$

where (5.37) implies

$$\beta = \left\{\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)\boldsymbol{Y}_i. \tag{5.39}$$

***SPECIAL CASE:*** With $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)$ as in (5.7), a common assumption is that

$$\text{var}(Y_{ij}|\boldsymbol{x}_i) = \sigma^2 \quad \text{for all } i,$$

so that $\boldsymbol{T}_i(\theta, \boldsymbol{x}_i) = \sigma^2\boldsymbol{I}_{n_i}$, $r = 1$, and thus

$$\boldsymbol{V}_i = \sigma^2\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i), \quad \xi = (\sigma^2, \alpha^T)^T. \tag{5.40}$$

It can be verified (do it) under these conditions that the estimating equation of form (5.38) corresponding to $\sigma^2$ ($k = 1$) reduces to

$$\sigma^2 = N^{-1}\sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)^T \boldsymbol{\Gamma}_i^{-1}(\alpha, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta). \tag{5.41}$$

We refer to this case further shortly.

***IMPLEMENTATION: Solution*** of the estimating equations to obtain the ***maximum likelihood estimators*** (***MLEs***) for $\beta$ and $\xi$, which we denote as $\widehat{\beta}$ and $\widehat{\xi}$, is of course equivalent to ***maximizing*** the loglikelihood (5.31) in $\beta$ and $\xi$. This is the way it is usually implemented in software packages, using ***standard optimization techniques*** such as a ***Newton-Raphson algorithm***.

The usual implementation takes advantage of the fact that (5.37) leads to the ***expression*** for $\beta$ in (5.39) in terms of $\xi$ and, when $\boldsymbol{V}_i$ is of the form in (5.40), with a multiplicative ***scale parameter*** $\sigma^2$, (5.38) yields the expression for $\sigma^2$ in (5.41) in terms of $\beta$ and $\alpha$. Any $\beta$ and $\sigma^2$ solving the estimating equations, or, equivalently, maximizing the loglikelihood, ***must satisfy*** these expressions.

Thus, if the expressions for $\beta$ in (5.39) and, in the case of (5.40), $\sigma^2$ in (5.41) are substituted into the loglikelihood, the result is a function ***solely*** of the covariance or correlation parameters. This practice is referred to as ***profiling***. The result is that the objective function so obtained can be maximized in the covariance or correlation parameters, which is an optimization problem of ***lower dimension*** so hopefully ***more tractable*** than maximizing the loglikelihood in ***all*** parameters as-is, not taking advantage of these expressions. Once the estimates of the covariance/correlation parameters are obtained, the estimates for $\beta$ and, if relevant, $\sigma^2$, maximizing the objective function can be obtained by substitution of $\widehat{\xi}$ in their expressions.

It is also possible to specify an ***iterative algorithm*** to solve the estimating equations that proceeds by cycling between solving the equation for $\beta$ holding $\xi$ fixed at the current estimate and solving that for $\xi$ holding $\beta$ fixed. This is more interesting and useful in the general ***nonlinear*** models we consider in later chapters, so we defer discussion until then.

## 5.4 Restricted maximum likelihood

***BIASED ESTIMATION IN FINITE SAMPLES:*** We have already observed that the MLEs for $\beta$ and $\xi$ under the assumption of normality should be ***consistent estimators*** for their true values $\beta_0$ and $\xi_0$, provided that the models $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta$ and $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ are ***correctly specified*** , under general conditions, as they solve ***unbiased estimating equations***. However, in ***finite samples*** , the estimator for $\xi$ can be subject to ***bias*** due to a phenomenon similar to that encountered in estimation of ***variance*** of a scalar outcome $Y$ from an iid sample or in ordinary linear regression.

In particular, if we have an iid sample $Y_1, \ldots, Y_m$ from some distribution with mean $\mu$ and variance $\sigma^2$, it is well known that the MLE for $\sigma^2$ under the assumption of ***normality*** ,

$$m^{-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2,$$

is a (downwardly) ***biased*** estimator for $\sigma^2$ for fixed $m$, as its expectation is $\sigma^2(m-1)/m$.

Accordingly, the usual ***sample variance*** estimator

$$s^2 = (m-1)^{-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2$$

is ***unbiased*** and thus preferred. Evidently, this bias is a consequence of the ***need to estimate*** $\mu$ rather than knowing it.

***ELIMINATING THE EFFECT OF ESTIMATION OF MEAN PARAMETERS:*** Thus, although the rationale for $s^2$ is immediate from this calculation, $s^2$ can also be deduced by viewing it as the result of an approach that does not rely on estimation of $\mu$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)^T$, with $\mathbf{1}$ a $(m \times 1)$ vector of 1s, and let $\boldsymbol{A}$ be a $(m \times m-1)$ matrix of column rank $m-1$ such that $\boldsymbol{A}^T\mathbf{1} = \mathbf{0}$. Defining the so-called vector of $m-1$ ***error constrasts***

$$\boldsymbol{U} = \boldsymbol{A}^T\boldsymbol{Y},$$

if we assume that the $Y_i$ are $\mathcal{N}(\mu, \sigma^2)$, so that $\boldsymbol{Y} \sim \mathcal{N}(\mu\mathbf{1}, \sigma^2\boldsymbol{I}_m)$, then it is straightforward to deduce that $\boldsymbol{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{A}^T\boldsymbol{A})$ and that maximizing the corresponding loglikelihood in $\sigma^2$ yields the estimator

$$\widehat{\sigma}^2 = (m-1)^{-1}\boldsymbol{Y}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = s^2$$

(try it). That is, the sample variance can be derived from effectively eliminating $\mu$ from consideration.

A similar result holds for linear regression. Here, with independent pairs $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, and model $Y_i = \boldsymbol{x}_i^T\beta + \epsilon_i$ with $E(\epsilon_i|\boldsymbol{x}_i) = 0$, $\text{var}(\epsilon_i|\boldsymbol{x}_i) = \sigma^2$, if $\boldsymbol{X}$ is the $(m \times p)$ design matrix with rows $\boldsymbol{x}_i$, if $\widehat{\beta}$ is the OLS estimator, it is well known that the MLE for $\sigma^2$ under the assumption that $\epsilon_i$ given $\boldsymbol{x}_i$ is ***normally distributed*** is $m^{-1}\sum_{i=1}^{m}(Y_i - \boldsymbol{x}_i^T\widehat{\beta})^2$, which can be shown to be biased. Dividing instead by $(m-p)$ yields the usual residual mean square, which is unbiased; a similar argument to that above based on suitably defined "***error contrasts***" can be made to justify this estimator, which is a simpler version of one we give shortly in the context of longitudinal data.

***DEMONSTRATION:*** Given these observations, it is natural to be concerned that normal-theory maximum likelihood estimation of the ***covariance parameters*** $\xi$ in our setting might be subject to similar bias. Clearly, it is not possible for general covariance model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ to carry out a similar explicit argument. To get a sense, however, consider the special case where

$$\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \sigma^2\boldsymbol{\Gamma}_i(\boldsymbol{x}_i), \tag{5.42}$$

where the correlation matrix $\boldsymbol{\Gamma}_i(\boldsymbol{x}_i)$ is a ***known*** function of covariates (so there is no parameter $\alpha$).

Writing $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Gamma}$ for brevity, the MLEs for $\beta$ and $\sigma^2$ in this case are (check)

$$\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{Y}, \qquad \widehat{\sigma}^2 = N^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}). \tag{5.43}$$

It is straightforward (try it) to show that the quadratic form in $\widehat{\sigma}^2$ in (5.43) can be written as

$$\boldsymbol{Y}^T\{\boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Gamma}^{-1}\}\boldsymbol{Y},$$

which, letting $\boldsymbol{Y}_* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{Y}$ and $\boldsymbol{X}_* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{X}$ for $\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Gamma}^{-1/2}$, can be reexpressed as

$$\boldsymbol{Y}_*^T\{\boldsymbol{I}_N - \boldsymbol{X}_*(\boldsymbol{X}_*^T\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\}\boldsymbol{Y}_* = \boldsymbol{Y}_*^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{Y}_*.$$

Here, $E(\boldsymbol{Y}_*|\tilde{\boldsymbol{x}}) = \boldsymbol{X}_*\beta$, $\text{var}(\boldsymbol{Y}_*|\tilde{\boldsymbol{x}}) = \sigma^2\boldsymbol{I}_N$ (verify), and $\boldsymbol{P}_*$ is a ***symmetric***, ***idempotent*** matrix. By the result for the ***expectation of a quadratic form*** in Appendix A,

$$E\{\boldsymbol{Y}_*^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{Y}_*|\tilde{\boldsymbol{x}}\} = \text{tr}\{\sigma^2\boldsymbol{I}_N(\boldsymbol{I}_N - \boldsymbol{P}_*)\} + \beta^T\boldsymbol{X}_*^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{X}_*\beta = \sigma^2\{N - \text{tr}(\boldsymbol{P}_*)\} + \boldsymbol{0} = \sigma^2(N - p),$$

using the fact that (see Appendix A) that the ***trace*** of a symmetric, idempotent matrix is equal to its ***rank***, and the rank of $\boldsymbol{X}$ ($N \times p$) and thus $\boldsymbol{X}_*$ and $\boldsymbol{P}_*$ is $p$, and $\boldsymbol{X}_*^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{X}_* = \boldsymbol{0}$ (check).

It follows that

$$E(\widehat{\sigma}^2|\tilde{\boldsymbol{x}}) = \frac{N-p}{N}\sigma^2,$$

demonstrating that the MLE is biased in finite samples ($m$ individuals, $N$ total observations) and that the alternative estimator

$$\widehat{\sigma}^2 = (N-p)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}) \tag{5.44}$$

is preferred. Again, it is evident that the bias is a consequence of the needing to estimate $\beta$ rather than knowing it. We now consider a ***generalization*** of the approach involving ***error contrasts*** above to estimation of $\xi$ in a covariance model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ that ***eliminates*** estimation of $\beta$ from the calculation.

***RESTRICTED MAXIMUM LIKELIHOOD:*** Analogous to the previous argument, let $\boldsymbol{A}$ be a $(N \times N-p)$ matrix of column rank $N-p$ such that $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$, where of course $\boldsymbol{X}$ is the $(N \times p)$ "stacked" design matrix for all $m$ individuals. Define the vector of $N-p$ ***error contrasts*** to be

$$\boldsymbol{U} = \boldsymbol{A}^T\boldsymbol{Y}.$$

Then if $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\beta, \boldsymbol{V})$, where we suppress dependence on $\xi$ and $\tilde{\boldsymbol{x}}$ for brevity, we can write

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}),$$

and it is straightforward that

$$\boldsymbol{U} = \boldsymbol{A}^T\boldsymbol{X}\beta + \boldsymbol{A}^T\boldsymbol{\epsilon} = \boldsymbol{A}^T\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A}). \tag{5.45}$$

The ***loglikelihood*** corresponding to (5.45) is easily found to be, ignoring constants,

$$l_R(\boldsymbol{\xi}) = (-1/2)\left[\log|\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}| + \boldsymbol{Y}^T\boldsymbol{A}\{\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}\}^{-1}\boldsymbol{A}^T\boldsymbol{Y}\right], \tag{5.46}$$

which does not depend on $\beta$. The claim is that maximizing $l_R(\boldsymbol{\xi})$ in $\boldsymbol{\xi}$ leads to an estimator that "***corrects***" for the finite-sample bias in the spirit of (5.44).

We first rewrite (5.46) in a form that makes it ***directly comparable*** to the usual normal loglikelihood (5.30). Note that an $\boldsymbol{A}$ that satisfies $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$ is, for $(N \times N - p)$ matrix $\boldsymbol{C}$,

$$\boldsymbol{A} = \{\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\}\boldsymbol{C}.$$

First, we show that the second term in (5.46) can be rewritten as

$$\boldsymbol{Y}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}), \tag{5.47}$$

where $\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$ as in (5.33).

- We first demonstrate that

$$\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T = \boldsymbol{P} \quad \text{where } \boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}. \tag{5.48}$$

  Defining

$$\boldsymbol{T} = \boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T,$$

  it is straightforward to observe that $\boldsymbol{T}$ is ***symmetric and idempotent*** , where idempotency can be verified by direct multiplication to show $\boldsymbol{T}\boldsymbol{T} = \boldsymbol{T}$, using $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$. Thus,

$$\text{tr}(\boldsymbol{T}\boldsymbol{T}^T) = \text{tr}(\boldsymbol{T}) = \text{tr}(\boldsymbol{I}_N) - \text{tr}\{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\} - \text{tr}\{\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\} = N - p - (N - p) = 0,$$

  and $\text{tr}(\boldsymbol{T}\boldsymbol{T}^T) = 0$ implies $\boldsymbol{T} = \boldsymbol{0}$ (check), from whence it follows that

$$\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T.$$

  Because

$$\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{A}^T\boldsymbol{V}^{1/2}\boldsymbol{V}^{-1/2}\boldsymbol{X} = (\boldsymbol{V}^{1/2}\boldsymbol{A})^T(\boldsymbol{V}^{-1/2}\boldsymbol{X}) = \boldsymbol{0},$$

  the same result above holds with $\boldsymbol{A}$ replaced by $\boldsymbol{V}^{1/2}\boldsymbol{A}$ and $\boldsymbol{X}$ replaced by $\boldsymbol{V}^{-1/2}\boldsymbol{X}$, yielding

$$\boldsymbol{I}_N - \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1/2} = \boldsymbol{V}^{1/2}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{V}^{1/2}. \tag{5.49}$$

  Pre- and post-multiplying (5.49) by $\boldsymbol{V}^{-1/2}$ then gives (5.48)

- It can then be shown by ***brute-force multiplication*** (try it) that

$$\boldsymbol{P} = \boldsymbol{PVP}. \tag{5.50}$$

Using (5.48) and (5.50), with $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$, we have

$$\begin{aligned}\boldsymbol{Y}^T\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y} &= \boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{V}\boldsymbol{P}\boldsymbol{Y} \\ &= \{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}\}^T(\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1})\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}\} \\ &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}),\end{aligned}$$

demonstrating (5.47).

We can thus rewrite the loglikelihood (5.46) as

$$l_R(\boldsymbol{\xi}) = (-1/2)\left\{\log|\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}| + (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})\right\}. \tag{5.51}$$

We now argue that the first term can be expressed as

$$\log|\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}| = \log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})| + \log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}|. \tag{5.52}$$

Differentiate (5.51) with respect to the $k$th component of $\boldsymbol{\xi}$ to obtain

$$\begin{aligned}(1/2)\Big((\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}) \\ - \text{tr}[\{\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{A}\}^{-1}\boldsymbol{A}^T\{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}\boldsymbol{A}]\Big).\end{aligned}$$

The second term can be written, using shorthand and letting $\boldsymbol{V}_\xi = \{\partial/\partial\xi_k\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\}$, as

$$\begin{aligned}tr\{(\boldsymbol{A}^T\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{V}_\xi\boldsymbol{A}\} &= \text{tr}(\boldsymbol{P}\boldsymbol{V}_\xi) \\ &= \text{tr}\{(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1})\boldsymbol{V}_\xi\} \\ &= \text{tr}(\boldsymbol{V}^{-1}\boldsymbol{V}_\xi) - \text{tr}\{\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1})\boldsymbol{V}_\xi\} \\ &= \{\partial/\partial\xi_k\log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})|\} - \text{tr}\{(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{V}_\xi\boldsymbol{V}^{-1}\boldsymbol{X}\} \\ &= \{\partial/\partial\xi_k\log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})|\} + \{\partial/\partial\xi_k\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}|\}.\end{aligned}$$

Because this shows that the derivative of the left hand side of (5.52) is equal to the derivative of the right hand side, we conclude that the first term in (5.51) can be rewritten as $\log|\boldsymbol{V}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})| + \log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{\xi},\tilde{\boldsymbol{x}})\boldsymbol{X}|$, as required.

Substituting in (5.51) yields what is usually referred to as the ***restricted maximum likelihood*** (***REML***) objective function

$$l_R(\xi) = (-1/2)\Big\{ \log|\boldsymbol{V}(\xi,\tilde{\boldsymbol{x}})| + (\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})(\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta}) + \log|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})\boldsymbol{X}|\Big\} \tag{5.53}$$

$$= (-1/2)\left[\sum_{i=1}^{m}\Big\{ \log|\boldsymbol{V}_i(\xi,\boldsymbol{x}_i)| + (\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\beta})^T\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\beta})\Big\} + \log\left|\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)\boldsymbol{X}_i\right|\right] \tag{5.54}$$

where

$$\widehat{\beta} = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})\boldsymbol{Y} = \left\{\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)\boldsymbol{Y}_i.$$

Note that (5.53) and (5.54) are functions of $\xi$ only, as $\widehat{\beta}$ depends on $\xi$. The suggestion is to maximize (5.53), or equivalently (5.54), in $\xi$ and then substitute the resulting estimator in the expression for $\widehat{\beta}$.

Comparing (5.53) and (5.54) to (5.30) and (5.31) with the expression (5.33) for $\widehat{\beta}$ substituted for $\beta$ shows that they have the ***same*** form except for the third term on the right hand side of (5.53) and (5.54). It is this term that effects the "***correction***" for finite sample bias.

Differentiating with respect to the $k$th component of $\xi$, $k = 1,\ldots,r+s$, and setting equal to zero yields

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta})^T\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\,\boldsymbol{V}(\xi,\tilde{\boldsymbol{x}})\}\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})(\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta})&\\ -\mathrm{tr}\left[\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\boldsymbol{V}(\xi,\tilde{\boldsymbol{x}})\}\right]&\\ +\mathrm{tr}\left[\{\boldsymbol{X}^T\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})\{\partial/\partial\xi_k\boldsymbol{V}(\xi,\tilde{\boldsymbol{x}})\}\boldsymbol{V}^{-1}(\xi,\tilde{\boldsymbol{x}})\boldsymbol{X}\right]&=0\end{aligned} \tag{5.55}$$

or, equivalently,

$$\begin{aligned}\sum_{i=1}^{m}\Big((\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\beta})^T\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\,\boldsymbol{V}_i(\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\beta})&\\ -\mathrm{tr}\left[\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\boldsymbol{V}_i(\xi,\boldsymbol{x}_i)\}\right]\Big)&\\ +\mathrm{tr}\left[\left\{\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)\{\partial/\partial\xi_k\boldsymbol{V}_i(\xi,\boldsymbol{x}_i)\}\boldsymbol{V}_i^{-1}(\xi,\boldsymbol{x}_i)\boldsymbol{X}_i\right]&=0.\end{aligned} \tag{5.56}$$

By the manipulations leading to (5.52), (5.55) can be rewritten, in shorthand, as

$$\boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{V}_\xi\boldsymbol{P}\boldsymbol{Y} - \mathrm{tr}(\boldsymbol{P}\boldsymbol{V}_\xi),$$

and it can be shown that $E(\boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{V}_\xi\boldsymbol{P}\boldsymbol{Y}) = \mathrm{tr}(\boldsymbol{P}\boldsymbol{V}_\xi)$, so that these estimating equations are ***unbiased***; the details are left as an exercise for the diligent student.

As for the MLEs, implementation is via ***maximization of the objective function*** (5.53) using standard optimization algorithms such as ***Newton-Raphson***.

***DEMONSTRATION, CONTINUED:*** We demonstrate that estimation of $\xi$ via REML is expected to lead to "***correction***" for bias due to estimation of $\beta$ in the special case in (5.42) where $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \sigma^2 \boldsymbol{\Gamma}_i(\boldsymbol{x}_i)$ and where the correlation matrix $\boldsymbol{\Gamma}_i(\boldsymbol{x}_i)$ is ***known***.

Writing $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Gamma}$ for brevity, we have as in (5.43) that

$$\widehat{\beta} = (\boldsymbol{X}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{Y},$$

and (5.55) becomes

$$(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})^T \boldsymbol{\Gamma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta})/\sigma^4 - \text{tr}(\boldsymbol{I}_N)/\sigma^2 + \text{tr}\{(\boldsymbol{X}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{X})\}/\sigma^2 = 0.$$

Noting that the last term is equal to $\text{tr}(\boldsymbol{I}_p) = p$, solving yields

$$\widehat{\sigma}_R^2 = (N - p)^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}) \boldsymbol{\Gamma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}). \tag{5.57}$$

The REML estimator in (5.57) can be seen to be identical to that in (5.44).

***REMARKS:***

- Although not possible to demonstrate in general, similar "***bias correction***" is achieved for covariance parameters other than scale parameters.
- The original justification for the REML approach is attributed to Patterson and Thompson (1971). See Verbeke and Molenberghs (2000, Section 5.3) for details and other interpretations of the approach.
- It is ***not possible*** to demonstrate theoretically that one of the ML or REML approach is ***uniformly preferable*** for estimation of covariance parameters $\xi$ in general. In the special case of balanced data collected according to a design like that in Chapter 3, with population mean model specified by the classical analysis of variance representation, it turns out that the estimators of the covariance parameters obtained using REML are ***the same*** as the ***classical ANOVA*** estimators obtained by equating mean squares to their expectations; see Verbeke and Molenberghs (2000, Section 5.3) for further references.
- ***In practice***, REML is often used ***by default*** owing to its interpretation given here as providing estimators that should exhibit less bias in finite samples. In fact, software implementing fitting of models like the ones in this and the next chapter ordinarily uses REML as the ***default method*** for estimation of covariance parameters.

## 5.5 Large sample inference

***SAMPLING DISTRIBUTION FOR $\widehat{\beta}$:*** As we have seen, in the context of a particular model

$$E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta, \quad \text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_i = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i) = \boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i)\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)\boldsymbol{T}_i^{1/2}(\theta, \boldsymbol{x}_i), \quad i = 1, \ldots, m, \quad (5.58)$$

most questions of scientific interest can be represented as questions about the components of $\beta$ in (5.58). To make inference on $\beta$ to address the questions formally, we require an estimator for $\beta$ and its ***sampling distribution***.

The obvious estimator for $\beta$ is that solving the estimating equation in (5.37), namely,

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{0}, \quad (5.59)$$

jointly with an estimating equation for $\xi$ such as the ML equation (5.38) or the REML equation (5.56).

- An estimating equation of the general form in (5.59) is often referred to as a ***linear estimating equation*** because it depends on the response through a ***linear function*** of the response, namely $(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)$. This will be important shortly.

Regardless of which method, ML or REML, one uses to estimate the covariance parameter $\xi$, even if the model in (5.58) is ***correctly specified*** and the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is ***exactly multivariate normal*** with these moments, it is ***not possible*** in general to derive the ***exact sampling distribution*** for the resulting estimator

$$\widehat{\beta} = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(\widehat{\xi}, \tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\widehat{\xi}, \tilde{\boldsymbol{x}})\boldsymbol{Y} = \left\{\sum_{i=1}^{m} \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{Y}_i, \quad (5.60)$$

where $\widehat{\xi}$ in (5.60) is either of the MLE or REML estimator for the covariance parameters in the covariance model in (5.58). Clearly, (5.60) is a complicated function of the data.

***LARGE SAMPLE THEORY:*** Accordingly, we appeal to ***large sample theory*** to derive an approximate sampling distribution for $\widehat{\beta}$ using the general approach for ***estimating equations*** discussed in Section 4.3. As we discussed there, the argument ***does not*** require that the assumption of ***normality*** of the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ holds.

We assume that the model for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ in (5.58) is ***correctly specified***. Recall that this means that there is a value $\beta_0$ such that the true expectation of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is $\boldsymbol{X}_i\beta_0$; that is, $\beta_0$ is a parameter of the distribution that ***truly generated the data***.

- Clearly, if this is ***not*** the case, then we are in pretty serious trouble, as we are addressing the questions of interest (which are questions about population mean response) in a framework that may ***not*** be consistent with the truth.

As suggested by our development so far, specification of a model for the overall population-averaged covariance matrix is admittedly ***more difficult*** than specifying a model for the mean. Accordingly, it is reasonable to be concerned that the model we specify for var($\boldsymbol{Y}_i|\boldsymbol{x}_i$) might ***not be correctly specified***. That is, for example we might select a ***correlation model*** $\boldsymbol{\Gamma}_i(\alpha, \boldsymbol{x}_i)$ that does not faithfully represent the true overall correlation structure, and/or we might make incorrect assumptions about the ***overall variance***.

Accordingly, we first consider the ***ideal situation*** in which the models for ***both*** overall mean and covariance posited in (5.58) are ***correctly specified*** , and then consider the case where the latter model might be ***incorrect***.

***COVARIANCE MODEL CORRECTLY SPECIFIED:*** If the model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ in (5.58) is ***correctly specified***, then there is a value $\xi_0$ such that the ***true*** overall covariance matrix

$$\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{V}_{0i} \tag{5.61}$$

is $\boldsymbol{V}_i(\xi_0, \boldsymbol{x}_i)$, $i = 1, \ldots, m$. That is, $\boldsymbol{V}_{0i} = \boldsymbol{V}_i(\xi_0, \boldsymbol{x}_i)$ is the covariance matrix of the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ ***actually*** generating the data.

Rather than just substituting directly into the generic argument in Section 4.3, we carry out the argument from scratch so as to demonstrate a ***fundamental*** and ***well-known result*** that persists across all types of mean-covariance models. The estimator (5.60) satisfies

$$\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}) = \boldsymbol{0}. \tag{5.62}$$

Collecting the parameters as $\eta = (\beta^T, \xi^T)^T$, let $\widehat{\eta} = (\widehat{\beta}^T, \widehat{\xi}^T)^T$. Because both the mean and covariance models are ***correctly specified*** , the estimating equation (5.59 ) and that solved to estimate $\xi$ (ML or REML) are ***unbiased estimating equations***. Thus, we expect that $\widehat{\eta}$ is a ***consistent estimator*** for the ***true value*** $\eta_0 = (\beta_0^T, \xi_0^T)^T$.

Following the argument in Section 4.3, we multiply (5.62) by $m^{-1/2}$ and take a linear Taylor series in $\widehat{\boldsymbol{\eta}}$ about the $\boldsymbol{\eta}_0$. Here, as on the last page of Appendix B (review it), instead of writing the ***linear term*** of the series in terms of this "stacked" parameter vector, we write it as the sum of terms corresponding to each component of $\boldsymbol{\eta}$. That is,

$$\begin{aligned}
\mathbf{0} &= m^{-1/2}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\xi}},\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}) \\
&\approx m^{-1/2}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi}_0,\boldsymbol{x}_i)(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0)+\left\{-m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{\xi}_0,\boldsymbol{x}_i)\boldsymbol{X}_i\right\}m^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\
&\quad+\left[m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\{\partial/\partial\boldsymbol{\xi}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi}_0,\boldsymbol{x}_i)\}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0)\right]m^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0). \qquad (5.63)
\end{aligned}$$

- In the third term on the right hand side in (5.63), we do not attempt to be more precise about the form of the partial derivative of the covariance matrix $\boldsymbol{V}_i(\boldsymbol{\xi},\boldsymbol{x}_i)$ with respect to $\boldsymbol{\xi}$ (which this notation is meant to indicate is evaluated at $\boldsymbol{\xi}_0$). This derivative evidently is rather complicated. As we see momentarily, we needn't worry about this.
- We have used the ***consistency*** of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\xi}}$ to approximate the sums in the second and third terms as evaluated at the true value $\boldsymbol{\eta}_0$ rather than an intermediate value $\boldsymbol{\eta}_*$ as in the argument in Section 4.3.

Write the expansion compactly as

$$\mathbf{0}\approx\boldsymbol{C}_m-\boldsymbol{A}_m\,m^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\boldsymbol{E}_m\,m^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0), \qquad (5.64)$$

where, using $\boldsymbol{V}_i(\boldsymbol{\xi}_0,\boldsymbol{x}_i)=\boldsymbol{V}_{0i}$ as in (5.61),

$$\boldsymbol{C}_m=m^{-1/2}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0),\qquad \boldsymbol{A}_m=m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\boldsymbol{X}_i,$$

$$\boldsymbol{E}_m=m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\{\partial/\partial\boldsymbol{\xi}\boldsymbol{V}_i^{-1}(\boldsymbol{\xi}_0,\boldsymbol{x}_i)\}(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_0).$$

- We in fact assume that

$$m^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0)=O_p(1); \qquad (5.65)$$

  i.e., that this quantity is ***bounded in probability*** (see Appendix C). Under regularity conditions, most estimators that are solutions to unbiased estimating equations satisfy (5.65). This says that $m^{1/2}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0)$ is "***well-behaved***" as $m\to\infty$ and describes the ***rate*** at which $\widehat{\boldsymbol{\xi}}\xrightarrow{p}\boldsymbol{\xi}_0$; i.e., (5.65) is equivalent to $\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0=O_p(m^{-1/2})$. This ensures that the rightmost term in (5.64) does not "***blow up***" as $m\to\infty$.

If we view the argument conditional on $\tilde{\boldsymbol{x}}$, then

$$\boldsymbol{A}_m \to \boldsymbol{A} = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_{0i}^{-1} \boldsymbol{X}_i.$$

By the ***central limit theorem***,

$$\boldsymbol{C}_m \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}),$$

where

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

By the ***weak law of large numbers***, using $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta_0$, it is straightforward that

$$\boldsymbol{E}_m \xrightarrow{p} \boldsymbol{0}. \tag{5.66}$$

Thus, rearranging and applying these results along with ***Slutsky's theorem***, we are left with

$$m^{1/2}(\widehat{\beta} - \beta_0) \approx \boldsymbol{A}^{-1}\boldsymbol{C}_m \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}). \tag{5.67}$$

- Note that (5.66) effectively ***eliminates*** any effect of having to ***estimate*** $\xi$. That is, if $\xi_0$ were ***known*** and substituted in (5.62), we could have immediately concluded (5.67).
- This reflects the ***fundamental result*** that if we obtain a estimator $\widehat{\beta}$ for a parameter $\beta$ in a model for a population mean by solving a ***linear estimating equation***, with an estimated "***weight matrix***," the large sample (normal) distribution of $m^{1/2}(\widehat{\beta} - \beta_0)$ is ***the same*** as that for the (***ideal***) estimator for $\beta$ we could have obtained if the "***weight matrix***" were ***known***.
- This says that there is ***no loss of precision*** suffered by the estimator for $\beta$ due to having had to ***estimate*** covariance parameters versus ***knowing*** them. Intuitively, this seems like a pretty ***optimistic*** result.
- Indeed, in ***small samples*** (small number of individuals $m$), inference based on the result in (5.67) can be optimistic in the sense that, for example, ***standard errors*** for the components of $\widehat{\beta}$ derived from (5.67) as we discuss momentarily will be ***too small*** and thus fail to reflect the true uncertainty associated with estimating $\beta$ (which includes uncertainty due to estimating $\xi$). In "larger" samples, inferences are often fairly reliable. Of course, what comprises "***large enough***" in any particular setting is not known.

To use the result (5.67) in practice, we approximate $\boldsymbol{A}$ by $\widehat{\boldsymbol{A}}_m$, where $\widehat{\boldsymbol{A}}_m$ is $\boldsymbol{A}_m$ with $\widehat{\xi}$ substituted for $\xi_0$ in $\boldsymbol{V}_{i0} = \boldsymbol{V}_i(\xi_0, \boldsymbol{x}_i)$, exploiting the fact that $\widehat{\xi}$ is a consistent estimator for $\xi_0$ under the conditions here. This yields the ***approximate sampling distribution*** for $\widehat{\beta}$ given by

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, m^{-1}\widehat{\boldsymbol{A}}_m^{-1}) = \mathcal{N}(\beta_0, \widehat{\boldsymbol{\Sigma}}_M), \tag{5.68}$$

where

$$\widehat{\boldsymbol{\Sigma}}_M = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\xi}, \boldsymbol{x}_i)\boldsymbol{X}_i\right)^{-1} = \{\boldsymbol{X}^T \boldsymbol{V}^{-1}(\widehat{\xi}, \tilde{\boldsymbol{x}})\boldsymbol{X}\}^{-1}. \tag{5.69}$$

(Note that the $m^{-1}$ on the left hand side of (5.68) "cancels" with that in $\widehat{\boldsymbol{A}}_m$.)

- In practice, ***standard errors*** for the estimators for the components of $\beta$ and associated ***confidence intervals*** for and ***test statistics*** concerning the corresponding components of the true parameter $\beta_0$ can be constructed in the usual way based on (5.68) and (5.69).

***COVARIANCE MODEL POSSIBLY INCORRECTLY SPECIFIED:*** We can generalize the above argument to the case where the posited model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ in (5.58) is ***not necessarily correctly specified***. That is, there is ***no value*** $\xi_0$ such that $\boldsymbol{V}_i(\xi_0, \boldsymbol{x}_i) = \boldsymbol{V}_{0i}$, where, again, $\boldsymbol{V}_{0i}$ is the ***true covariance matrix*** generating the data.

Of course, in practice, we would proceed unknowingly as if the model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ ***is*** correct and solve an estimating equation of the form (5.38) (ML) or (5.56) (REML) to obtain an estimator $\widehat{\xi}$. Because the model is incorrect, it is not even clear that $\xi$ has meaning, as it does not represent a quantity relevant to the ***true mechanism*** generating the data. Accordingly, it is not clear exactly what $\widehat{\xi}$ is "***estimating***."

In the generic argument in Section 4.3, we started from the premise that the model underlying the estimating equations being solved for the parameter $\eta$ is ***correctly specified*** , so that the estimating equations $\sum_{i=1}^{m} \boldsymbol{\Psi}_i(\mathcal{U}_i, \eta) = \boldsymbol{0}$ are ***unbiased*** ; and

$$E\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta_0)\} = \boldsymbol{0},$$

where $\eta_0$ is the true value. Inspection of (5.38) or (5.56) makes clear that, in our problem, if the model $\boldsymbol{V}_i$ is not correct, then a summand of the estimating equations ***does not*** have expectation zero necessarily, so that the estimating equations are ***not unbiased***. In this situation, we can say something about the behavior of $\widehat{\xi}$ in our problem, as follows.

In the generic case of a ***correct*** model, under regularity conditions, it is possible to ***weaken*** the argument in Section 4.3. If instead we have only that

$$\sum_{i=1}^{m} E\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta_0)\} = \mathbf{0} \tag{5.70}$$

(so that each summand does not necessarily have mean zero, but their sum does), then it still holds in general that $\widehat{\eta} \xrightarrow{p} \eta_0$, and the argument leading to the asymptotic normality of the estimator for $\eta$ goes through unchanged, except that the covariance matrix of $\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta_0)$ is no longer equal to $E\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta_0)\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \eta_0)\}$, so that the definitions of the matrices $\boldsymbol{B}_m$ and $\boldsymbol{B}$ in the argument must be changed; e.g., $\boldsymbol{B}_m = m^{-1}\sum_{i=1}^{m} \text{var}\{\boldsymbol{\Psi}_i j(\mathcal{U}_i, \eta_0)\}$ instead.

If the model on which the estimating equations are based is ***incorrect*** under regularity conditions, it is usually the case that there exists $\eta^*$ such that

$$\sum_{i=1}^{m} E\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta^*)\} = \mathbf{0}, \tag{5.71}$$

where this expectation is still with respect to the ***true distribution*** of $\mathcal{U}_i$.

It turns out that, by analogy to (5.70), if (5.71) holds, solving the "***incorrect***" estimating equation will yield an "estimator" such that

$$\widehat{\eta} \xrightarrow{p} \eta^*. \tag{5.72}$$

Although $\eta^*$ does not have any meaning with respect to the ***true distribution*** generating the data, it is a fixed value dictated by (5.71). A value like $\eta^*$ can be thought of as the value that "tries to get closest" to the representing the truth within the confines of an incorrect model, and consequently has been referred to as the ***least false parameter***.

The key point is that, ***even with an incorrectly specified model***, we can still deduce the behavior of an "estimator" for a parameter in that model, even if the parameter has no real meaning.

Returning to our problem, we thus assume that, for ***incorrectly specified*** model $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$, if we solve estimating equations like those in (5.38) or (5.56), the solution $\widehat{\xi}$ satisfies (5.72) for some $\xi^*$; namely,

$$\widehat{\xi} \xrightarrow{p} \xi^*,$$

and, under regularity conditions and analogous to (5.65), $m^{1/2}(\widehat{\xi} - \xi^*) = O_p(1)$.

Suppose then that $\boldsymbol{V}_i$ is incorrectly specified, let

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

denote the "***incorrect covariance matrix***" implied by the choice of this incorrect model, and consider again solving (5.62), namely,

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

First, note that for the estimating equation (5.59),

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

we have

$$E\{\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{\xi}^*, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i\} = E\{\boldsymbol{X}_i^T \boldsymbol{V}_i^{*\,-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)|\boldsymbol{x}_i\} = \boldsymbol{0},$$

$i = 1, \ldots, m$, so that the expectation of ***each summand*** is zero, ***even though*** the covariance model is ***incorrectly specified***, so that the estimating equation is still ***unbiased***, analogous to the demonstration for univariate OLS in Section 4.3. We thus conclude that $\widehat{\boldsymbol{\beta}}$ is a ***consistent estimator*** for $\boldsymbol{\beta}_0$, despite the fact that the "***weight matrix***" used in the linear estimating equation is not the inverse of the true covariance matrix. In fact, this holds even if we take $\boldsymbol{V}_i = \boldsymbol{I}_{n_i}$, $i = 1, \ldots, m$; that is, assume all $N$ observations across all $m$ individuals are ***mutually uncorrelated***. The resulting estimator for $\boldsymbol{\beta}$ is effectively ***OLS***, treating all $N$ observations as if they were independent.

Expanding about $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\xi}}^T)^T = (\boldsymbol{\beta}_0^T, \boldsymbol{\xi}^{*\,T})^T$, analogous to (5.63),

$$\begin{aligned}
\boldsymbol{0} &= m^{-1/2} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}) \\
&\approx m^{-1/2} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{\xi}^*, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0) + \left\{ -m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{\xi}^*, \boldsymbol{x}_i)\boldsymbol{X}_i \right\} m^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad + \left[ m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \{\partial/\partial\boldsymbol{\xi}\, \boldsymbol{V}_i^{-1}(\boldsymbol{\xi}^*, \boldsymbol{x}_i)\}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0) \right] m^{1/2}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*) \\
&= \boldsymbol{C}_m^* - \boldsymbol{A}_m^*\, m^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \boldsymbol{E}_m^*\, m^{1/2}(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*). \qquad (5.73)
\end{aligned}$$

With $\boldsymbol{V}_i^* = \boldsymbol{V}_i(\boldsymbol{\xi}^* \boldsymbol{x}_i)$ and $\boldsymbol{V}_{0i} = \mathrm{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$, the ***true covariance matrix***, it is clear that

$$\boldsymbol{E}_m^* \xrightarrow{p} \boldsymbol{0}, \qquad \boldsymbol{A}_m^* \to \boldsymbol{A}^* = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{V}_i^{*\,-1} \boldsymbol{X}_i,$$

$$\boldsymbol{C}_m^* = m^{-1/2} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{*\,-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}^*)$$

by the ***central limit theorem***, where

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

Thus, rearranging and using ***Slutsky's theorem*** as before, we have

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{*-1}\boldsymbol{B}^{*}\boldsymbol{A}^{*-1}). \tag{5.74}$$

- As in the case where the covariance model is ***correctly specified*** , because $\boldsymbol{E}_m \xrightarrow{p} \mathbf{0}$, there is ***no effect*** of estimating $\xi$ in the incorrect model $\boldsymbol{V}_i$. If the matrix $\boldsymbol{V}_i^*$ had been ***known*** , it is straightforward to observe that (5.74) would still follow.

  This reflects a generalization of the result we saw in the case of a ***correctly specified*** covariance model, namely that the large sample distribution of $m^{1/2}(\widehat{\beta} - \beta_0)$ is the ***same*** if the "***weight matrix*** " used in the linear estimating equation for $\beta$ is ***fixed*** or ***estimated***.

- In fact, the argument leading to the result (5.67) in the case of a ***correctly specified*** model is a ***special case*** of this result, where the covariance model $\boldsymbol{V}_i$ is ***correct*** after all, so that the $\xi^* = \xi_0$, the value such that $\boldsymbol{V}_{0i} = \boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$.

Note that (5.74), while informative about the behavior of the estimator for $\beta$ when the posited covariance model is ***incorrect*** , cannot be used as-is in practice, as $\boldsymbol{V}_{0i}$ is of course unknown. We return to this point shortly.

***OPTIMAL LINEAR ESTIMATING EQUATION:*** From (5.67), when the covariance model is ***correctly specified*** , the estimator solving the linear estimating equation satisfies

$$\widehat{\beta}_C \dot{\sim} \mathcal{N}\{\beta_0, (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}\}, \tag{5.75}$$

where the subscript $C$ indicates "correct," and $\boldsymbol{V}_0 = \text{block diag}(\boldsymbol{V}_{01}, \ldots, \boldsymbol{V}_{0m})$. Likewise, when the covariance model is ***incorrectly specified*** , from (5.74), the estimator solving the linear estimating equation satisfies

$$\widehat{\beta}_{IC} \dot{\sim} \mathcal{N}\{\beta_0, (\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{V}_0\boldsymbol{V}^{*-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1}\}, \tag{5.76}$$

where the subscript $IC$ indicates "incorrect," and $\boldsymbol{V}^* = \text{block diag}(\boldsymbol{V}_1^*, \ldots, \boldsymbol{V}_m^*)$.

The covariance matrices of the approximate sampling distributions in (5.75) and (5.76) reflect, at least for $m$ large, the ***precision*** with which $\beta$ can be estimated by solving the linear estimating equation for $\beta$ under correct and incorrect covariance models. Both $\widehat{\beta}_C$ and $\widehat{\beta}_{IC}$ are ***consistent*** estimators for $\beta_0$; thus, we can ***compare*** the covariance matrices of their approximate sampling distributions to examine the ***relative efficiency*** of $\widehat{\beta}_{IC}$ to $\widehat{\beta}_C$.

To this end, consider the difference

$$(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{V}_0\boldsymbol{V}^{*-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}. \tag{5.77}$$

We now argue that the difference (5.77) is a ***nonnegative definite*** matrix; that is,

$$\lambda^T\{(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{V}_0\boldsymbol{V}^{*-1}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{V}^{*-1}\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}\}\lambda \geq \boldsymbol{0} \tag{5.78}$$

for all $\lambda$. It follows that, if (5.78) holds, the ***diagonal elements*** (5.77) are all $\geq 0$, so that the difference in the approximate ***sampling variances*** of the estimators for each component of $\beta$ is $\geq 0$ (check), implying that the components of $\widehat{\beta}_C$ are ***more efficient*** than those of $\widehat{\beta}_{IC}$.

Letting

$$\boldsymbol{X}_* = \boldsymbol{V}^{*-1/2}\boldsymbol{X}, \quad \boldsymbol{V}^{*-1/2}\boldsymbol{V}^{*-1/2} = \boldsymbol{V}^{*-1}, \quad \boldsymbol{W} = \boldsymbol{V}^{*1/2}\boldsymbol{V}_0^{-1}\boldsymbol{V}^{*1/2},$$

$$\boldsymbol{c} = \boldsymbol{W}^{-1/2}\boldsymbol{X}_*(\boldsymbol{X}_*^T\boldsymbol{X}_*)^{-1}\lambda, \quad \boldsymbol{W} = \boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2},$$

rewrite (5.78) as (check)

$$\boldsymbol{c}^T\{\boldsymbol{I}_N - \boldsymbol{W}^{1/2}\boldsymbol{X}_*(\boldsymbol{X}_*^T\boldsymbol{W}\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\boldsymbol{W}^{1/2}\}\boldsymbol{c}. \tag{5.79}$$

It is straightforward to verify (try it) that

$$\boldsymbol{I}_N - \boldsymbol{W}^{1/2}\boldsymbol{X}_*(\boldsymbol{X}_*^T\boldsymbol{W}\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\boldsymbol{W}^{1/2} = \boldsymbol{I}_N - \boldsymbol{P}_*$$

is ***symmetric and idempotent*** , so that (5.79) can be written as

$$\boldsymbol{c}^T(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{c} = \boldsymbol{c}^T(\boldsymbol{I}_N - \boldsymbol{P}_*)(\boldsymbol{I}_N - \boldsymbol{P}_*)\boldsymbol{c} = \boldsymbol{d}^T\boldsymbol{d} \geq \boldsymbol{0},$$

demonstrating (5.78).

The result (5.78) shows that, at least approximately (for "large" $m$), the components of $\widehat{\beta}_C$ are more precise estimators than those of $\widehat{\beta}_{IC}$. Formally, (5.78) demonstrates that, for a given population mean model $\boldsymbol{X}\beta$, among all ***linear estimating equations*** , that formed using a ***correct covariance model*** will yields a (asymptotically) relatively more efficient estimator than any other based on an ***incorrect covariance model***. That is, the linear estimating equation with "weight matrix" based on a correct covariance model is ***optimal*** among all linear equations in this sense. Of course, this comes as no surprise.

The result ***does not*** provide insight into ***how much more precise*** in general. Evidently, the comparison of the large sample covariance matrices will depend on the ***particular situation*** – the population mean response model and covariates $\boldsymbol{x}_i$ on which it is based (assumed correct), the true covariance matrix, and the assumed covariance model.

We demonstrate a more general optimality result in the case of a ***nonlinear model*** for population-averaged mean response in Chapter 8, which subsumes the one here.

***NORMALITY NOT REQUIRED:*** Note that nowhere in these arguments is anything assumed about the ***true distribution*** of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$; e.g., that it is ***multivariate normal***. The only assumption on this distribution required is that it possess ***sufficient moments*** so that application of the weak law of large numbers, the central limit theorem, etc, is justified. Accordingly, even though we derived the ***estimating equations*** for $\beta$ and $\xi$ in the assumed covariance model by starting with the ***normal loglikelihood*** , the resulting estimator for $\beta$ has desirable properties that hold much more generally.

***"ROBUST" COVARIANCE MATRIX:*** In practice, it is natural to be concerned that a posited covariance model is ***not correctly specified***. Identifying an appropriate model is admittedly ***challenging*** ; the structure adopted must faithfully represent the ***aggregate effects*** of both ***among-*** and ***within-individual*** variance and correlation.

Accordingly, rather than carry out inference on $\beta_0$ based on the approximate sampling distribution in (5.68), which is based on the covariance model being ***correct*** , it is conventional to base it on the foregoing argument under the condition that the posited covariance model ***may not be correct*** and the result in (5.74), which we repeat here for convenience, dropping the subscript *IC*:

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{*-1}\boldsymbol{B}^{*}\boldsymbol{A}^{*-1}), \tag{5.80}$$

where

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

Of course, $\boldsymbol{A}^{*}$ can be approximated by

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

the difficulty is that $\boldsymbol{B}^{*}$ depends on the ***true covariance matrix*** $\boldsymbol{V}_{0i}$, which is not known.

However, from the argument in Section 4.3, $\boldsymbol{B}^{*}$ can be approximated by

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

The diligent student can verify that, using the consistency of $\widehat{\beta}$ and weak law of large numbers, this expression converges in probability to $\boldsymbol{B}^{*}$.

Combining, it is thus common to base inference on the approximate sampling distribution

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\Sigma}}_R), \tag{5.81}$$

$$\widehat{\boldsymbol{\Sigma}}_R = \tag{5.82}$$
$$\left\{\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})^T \boldsymbol{V}^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)\boldsymbol{X}_i \left\{\sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}^{-1}(\widehat{\boldsymbol{\xi}}, \boldsymbol{x}_i)\boldsymbol{X}_i\right\}^{-1}.$$

$\widehat{\boldsymbol{\Sigma}}_R$ is often referred to as the ***robust sandwich*** or ***empirical*** (sampling) covariance matrix, in contrast to $\widehat{\boldsymbol{\Sigma}}_M$ in (5.69), which is often called the ***model-based*** covariance matrix, being based on the assumption that the model for ***overall covariance structure*** is ***correctly specified***. "***Robust***" refers to the fact that $m\times$ (5.82) is a ***consistent estimator*** for the true sampling covariance matrix of $m^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ when the covariance model is ***incorrectly specified*** (and even if it ***is correct***). It is thus ***robust*** to possible misspecification of the covariance model $\boldsymbol{V}_i$.

- It is conventional in practice to base inference on the ***robust covariance matrix*** $\widehat{\boldsymbol{\Sigma}}_R$ rather than the ***model-based*** version $\widehat{\boldsymbol{\Sigma}}_M$ to protect against the possibility of an incorrect covariance model.
- ***Software packages*** implementing these methods and those in the next chapter usually use $\widehat{\boldsymbol{\Sigma}}_R$ ***by default*** to compute approximate standard errors, confidence intervals, and so on.
- By the argument leading to (5.77), using $\widehat{\boldsymbol{\Sigma}}_R$ should result in ***less optimistic assessment of the precision*** with which the components of $\boldsymbol{\beta}$ are estimated.

***QUESTIONS OF INTEREST:*** As discussed in the context of the examples in Section 5.2, questions of scientific interest are usually expressed in terms of ***linear functions*** of the components of $\boldsymbol{\beta}$.

For instance, in the population mean response model (5.14) for the dental study given by

$$E(Y_{ij}|\boldsymbol{x}_i) = \{\beta_{0,B}g_i + \beta_{0,G}(1 - g_i)\} + \{\beta_{1,B}g_i + \beta_{1,G}(1 - g_i)\}t_{ij},$$

$$\boldsymbol{\beta} = (\beta_{0,G}, \beta_{1,G}, \beta_{0,B}, \beta_{1,B})^T,$$

interest focuses on the ***difference in slopes*** between the genders, $\beta_{1,B} - \beta_{1,G}$, so that

$$\boldsymbol{L} = (0, -1, 0, 1). \tag{5.83}$$

If interest is in ***estimating*** the population mean response for boys at age $t_0 = 11$, then we focus on

$$\boldsymbol{L}\boldsymbol{\beta} = \beta_{0,B} + \beta_{1,B}t_0, \qquad \boldsymbol{L} = (0, 1, 0, , t_0).$$

Questions of interest can also involve more than one ***contrast*** of the components of $\boldsymbol{\beta}$; for example, continuing with the dental study, whether or not the (assumed straight line) population mean response trajectories for boys and girls in fact ***coincide*** involves the two contrasts $\beta_{0,B} - \beta_{0,G}$ and $\beta_{1,B} - \beta_{1,G}$ (equal intercepts and slopes). The null hypothesis that ***both intercepts and slopes*** for boys and girls are the same, so that the lines coincide, can be expressed as as $\boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$, where

$$\boldsymbol{L} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}. \tag{5.84}$$

In general, questions can be expressed in terms of $\boldsymbol{L}$ $(c \times p)$ (of full rank), $c \geq 1$, corresponding to a set of contrasts of interest, where ordinarily rank($\boldsymbol{L}$) = $c$.

***INFERENCE:*** Using the approximate sampling distribution for $\widehat{\boldsymbol{\beta}}$, an ***estimator*** for $\boldsymbol{L}\boldsymbol{\beta}$ is then $\boldsymbol{L}\widehat{\boldsymbol{\beta}}$, and, with $\widehat{\boldsymbol{\Sigma}}$ either of $\widehat{\boldsymbol{\Sigma}}_M$ or $\widehat{\boldsymbol{\Sigma}}_R$, $\boldsymbol{L}\widehat{\boldsymbol{\beta}}$ has approximate sampling distribution, from (5.68) and (5.81),

$$\boldsymbol{L}\widehat{\boldsymbol{\beta}} \stackrel{\cdot}{\sim} \mathcal{N}(\boldsymbol{L}\boldsymbol{\beta}_0, \boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T). \tag{5.85}$$

Thus, for example, if $\boldsymbol{L}\boldsymbol{\beta}$ represents the difference in slopes in (5.83) ($c$ = 1), a ***standard error*** for $\boldsymbol{L}\widehat{\boldsymbol{\beta}}$ is $(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{1/2}$, and a conventional ***Wald*** type $100(1-\alpha)\%$ ***confidence interval*** for $\boldsymbol{L}\boldsymbol{\beta}_0$ is

$$\boldsymbol{L}\widehat{\boldsymbol{\beta}} \pm c_{\alpha/2}(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{1/2},$$

where $c_{\alpha/2}$ is an appropriate critical value, such as the the $1 - \alpha/2$ quantile of the standard normal or $t$ distribution with some degrees of freedom, discussed further below. A test of $H_0 : \beta_{1,B} - \beta_{1,G} = 0$ versus $H_1 : \beta_{1,B} - \beta_{1,G} \neq 0$ would be based on comparing the test statistic $\boldsymbol{L}\widehat{\boldsymbol{\beta}}/(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{1/2}$ to the appropriate critical value from a normal or $t$ distribution.

More generally, approximate test statistics for the hypotheses

$$H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{h} \;\; vs. \;\; H_1 : \boldsymbol{L}\boldsymbol{\beta} \neq \boldsymbol{h},$$

where $\boldsymbol{L}$ is $(c \times p)$ with (usually) rank($\boldsymbol{L}$) = $c$, and $\boldsymbol{h}$ is a specified $(c \times 1)$ vector (***almost always*** $\boldsymbol{h} = \boldsymbol{0}$), can be constructed based on what is now the $c$-variate approximate sampling distribution (5.85).

- An approximate ***Wald test statistic*** is

$$T_L = (\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})^T(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{-1}(\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h}), \tag{5.86}$$

  which has approximately a ***chi-squared*** distribution with rank($\boldsymbol{L}$) degrees of freedom, so that the test is carried out by comparing $T_L$ to the appropriate $\chi^2$ critical value. If $\boldsymbol{L}$ is a row vector ($c = 1$), then this test is equivalent to the usual "$Z$ test" based on using a standard normal critical value.

- Wald-type tests can be ***optimistic*** in practice and reject $H_0$ more often than they should because either of the large sample approximate sampling distributions (5.68) and (5.81) ***do not take into account*** variability associated with estimating $\xi$, so that the test statistic is ***too large***. In finite samples (finite $m$), this is often addressed by instead using a statistic of the form

$$F_L = \frac{(\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})^T(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}\boldsymbol{L}^T)^{-1}(\boldsymbol{L}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})}{\text{rank}(\boldsymbol{L})}, \tag{5.87}$$

  which is compared to an $F$ distribution with rank($\boldsymbol{L}$) numerator degrees of freedom and denominator degrees of freedom ***estimated*** from the data. When $c = 1$, this reduces to a $t$ test, with degrees of freedom estimated similarly.

  Several methods have been proposed to ***estimate the denominator degrees of freedom*** for the test statistic (5.87), one of which is based on the so-called ***Satterthwaite approximation***. These are implemented in available software. We do not discuss these here; see Verbeke and Molenberghs (1997, Section 3.5.2 and Appendix A) and the documentation for SAS `proc mixed` for details. These methods usually lead to ***different results***; however, with large $m$, all yield degrees of freedom that are sufficiently large that the associated p-values are very similar.

When the null hypothesis corresponds to a comparison of ***nested models***, as for $\boldsymbol{L}$ in (5.83) with equal slopes or in (5.84) where the straight lines for boys and girls ***coincide*** under the null, an alternative approach is to carry out a classical ***likelihood ratio test*** based on the ***normal likelihood***

$$L_{ML}(\beta, \xi) = \prod_{i=1}^{m}(2\pi)^{-n_i/2}|\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)|^{-1/2}\exp\{-(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)^T\boldsymbol{V}_i^{-1}(\xi, \boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{X}_i\beta)/2\}. \tag{5.88}$$

Here, one fits the "***full model***" of interest (5.58) first by solving the estimating equations for $\beta$ and $\xi$ dictated by (5.88) [equivalently, ***maximizing*** (5.88)] to obtain $\widehat{\beta}$ and $\widehat{\xi}$. One then imposes the condition dictated by the null hypothesis $\boldsymbol{L}\beta = \boldsymbol{0}$ and fits the resulting "***reduced model***" by maximizing the corresponding (5.88) (solving the estimating equations) to obtain $\widehat{\beta}^{(0)}$ and $\widehat{\xi}^{(0)}$, say.

The ***likelihood ratio test statistic*** is then

$$T_{LRT} = -2\{\log L_{ML}(\widehat{\beta}^{(0)}, \widehat{\xi}^{(0)}) - \log L_{ML}(\widehat{\beta}, \widehat{\xi})\}. \quad (5.89)$$

Under regularity conditions, the test statistic $T_{LRT}$ in (5.89) has an approximate ***chi-squared distribution*** with degrees of freedom equal to the ***difference in the dimensions*** $p$ of $\beta$ in the "full" model and that for the "reduced" model; this difference is typically equal to $c$.

- Although the test statistic (5.89) comes about from assuming that the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ is ***normal*** , large sample (large $m$) arguments show that the result that $T_{LRT}$ has an approximate $\chi^2$ distribution holds ***even if*** this distribution is ***not normal***.
- If one uses the ***REML objective function*** in place of the normal likelihood (5.88), a valid test is ***not obtained***. This is because the population mean parameter $\beta$ is ***eliminated from consideration*** through the "error contrasts," and this parameter is ***different*** under the "full" and "reduced" models, so that the REML objective function is effectively based on ***different*** (mean zero) responses under each model and thus the two REML "loglikelihoods" are not comparable.

Inference on ***components of*** $\xi$ is also sometimes of interest. We defer discussion of this to Chapter 6. For now, we describe the use of so-called ***information criteria*** as a way of ***informally*** comparing competing models, and in particular competing ***covariance models***.

***INFORMATION CRITERIA:*** Although scientific questions are typically framed in terms of $\beta$ in a model of the form $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta$ and can sometimes be cast as a comparison between ***nested models*** for the population mean response of this form, other questions arise where the models to be compared ***cannot*** be viewed as nested.

For example, in ***building a model*** of the form (5.58), while we have in mind a specific model $E(\boldsymbol{Y}_i|\boldsymbol{x}_i) = \boldsymbol{X}_i\beta$ in which to frame the scientific questions, we may wish to compare the support in the data for several different models $\boldsymbol{V}_i(\xi, \boldsymbol{x}_i)$ for the ***overall covariance structure*** var$(\boldsymbol{Y}_i|\boldsymbol{x}_i)$. Ordinarily, competing models, e.g., compound symmetric versus AR(1), for example, are not nested.

Alternatively, we may wish to compare two competing models for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ that involve different combinations of covariates and consequently are ***not nested***.

***Information criteria*** provide an ***informal approach*** to these challenges. As is well known, the ***more parameters*** that are incorporated in a model, the ***larger the loglikelihood*** becomes; thus, if we wish to compare competing models that are not nested based on the maximized loglikelihoods for these models, we must take this into account. Simply comparing the maximized loglikelihoods directly favors "larger" models. Accordingly, the idea behind information criteria is to incorporate a ***penalty*** for using more parameters and compare instead ***penalized versions*** of the maximized loglikelihoods.

Let $\log \widehat{L}_{ML}$ denote generically the maximized loglikelihood for a specific mean-covariance model, and let $P$ be the total number of parameters (mean and covariance) in the model ($= p + r + s$ for us). Some popular information criteria are as follows; the definitions are such that ***smaller*** values are preferred.

- ***Akaike Information Criterion (AIC):***

$$AIC = -2 \log \widehat{L}_{ML} + 2P. \tag{5.90}$$

- ***Schwarz's Bayesian Information Criterion (BIC):*** With $N$ the total number of observations,

$$BIC = -2 \log \widehat{L}_{ML} + (P \log N). \tag{5.91}$$

- ***Hannan-Quinn Information Criterion (HQ):***

$$HQ = -2 \log \widehat{L}_{ML} + \{P \log(\log N)\}. \tag{5.92}$$

All but $AIC$ involve ***penalties*** depending on ***both*** the ***number of model parameters*** $P$ and the ***total number of observations*** $N$, so that differences in loglikelihood are calibrated relative to both of these factors.

Analogous criteria can be defined based on the logarithm of the ***REML*** objective function. ***However***, as noted above, REML "loglikelihoods" are comparable only if they involve the ***same mean model***; thus, information criteria based on REML should be used only to compare ***covariance models*** paired with the ***same*** population mean response model. Some advocate here setting $P$ equal to the number of covariance parameters ($P = r + s$). In addition, because the REML objective function is formulated based on $N - p$ ***error contrasts***, $N$ in (5.91) and (5.92) should be replaced by $N - p$.

Inspection of information criteria ***should not*** be used to draw formal inferences; rather, they should be viewed only as ad hoc ***rules of thumb***. It is entirely possible in practice that ***different criteria*** will prefer ***different models***. $AIC$ often prefers "larger" models relative to $BIC$, with $HQ$ intermediate. It is beyond our scope to offer a rigorous justification for the use of information criteria for this purpose.

## 5.6 Missing data

Longitudinal data analysis often involves dealing with ***missing data***, most prominently because of ***attrition*** of individuals over time; that is, ***dropout***. This is, of course, a recurrent challenge when the individuals are ***human subjects***.

Here, although it is ***intended*** to ascertain the outcome of interest at specific time points, as in many of the examples we have examined, some individuals ***fail to present*** for the outcome to be recorded ***after a certain time point***, leading to what is often called a ***monotone pattern of missingness***. More generally, it is the case in many longitudinal studies that individuals do not show up at the intended times in a ***haphazard fashion***, so that the pattern of missingness for any individual can be ***nonmonotone***.

We have already discussed in Section 5.2 the hip replacement study, in which several patients exhibit a ***nonmonotone*** missingness pattern in which they are missing the intended response measurement at week 2 (with one patient also missing the baseline measurement). Recall that, because this phenomenon seems ***systematic*** and occurs for about half of the patients of each gender, it is reasonable to speculate that the fact that these observations are missing has ***nothing to do*** with the health status of patients or their genders. We return to this point shortly.

***EXAMPLE: AGE-RELATED MACULAR DEGENERATION CLINICAL TRIAL:*** Figure 5.4 shows data reported by Molenberghs and Kenward (2007) from a multicenter clinical trial comparing an experimental (active) treatment, interferon-$\alpha$, to placebo in $m = 240$ patients with age-related macular degeneration (AMD), a leading cause of vision loss among people aged 50 and older. AMD causes damage to the macula, a spot near the center of the retina and the part of the eye needed for sharp, central vision. Patients with AMD progressively lose vision at varying rates. The response, ***visual acuity***, was assessed at baseline (week 0) and then at weeks 4, 12, 24, and 52, and measured the total number of letters a patient read correctly on a standardized vision chart with lines of letters of decreasing size.

Patients were randomized to the two treatments, and all have baseline responses observed; however, ***only 188 of the 240 patients*** have observed responses at ***all five time points***. Of those remaining, 24 dropped out before the final clinic visit at 52 weeks, 8 before the 24 week visit, 6 before the 12 week visit, and 6 before the 4 week visit, with the remaining 8 missing visits intermittently. These data exemplify the very common situation in longitudinal studies in humans in which missingness is almost entirely due to ***dropout***.

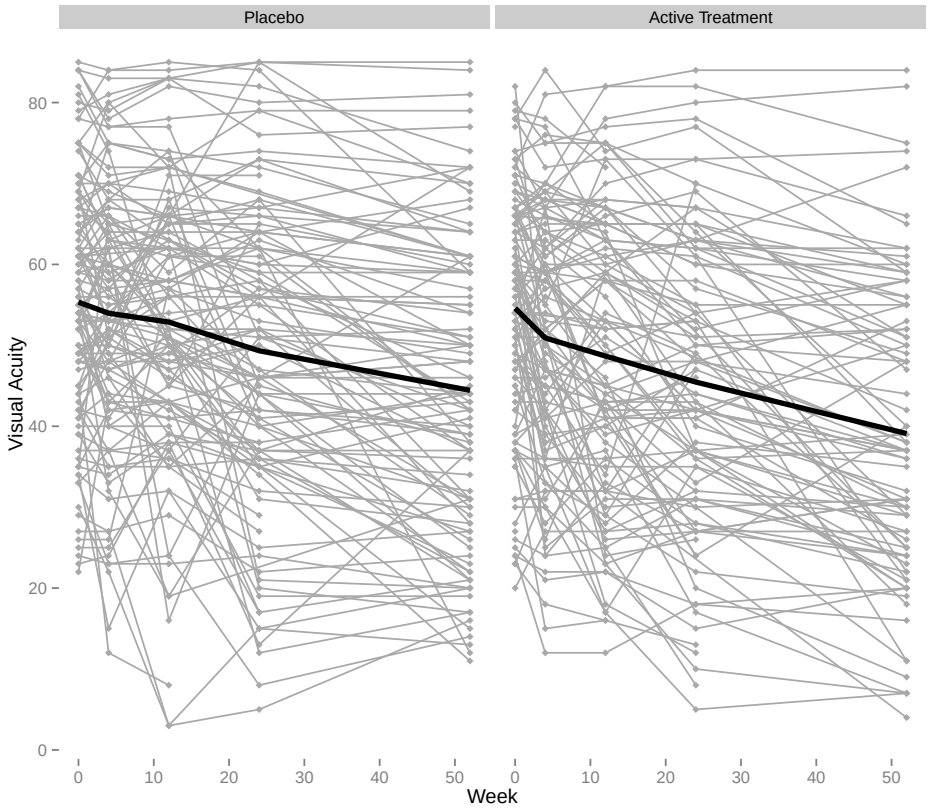


Figure 5.4: *Visual acuity profiles for subjects in the age-related macular degeneration trial. Averages of observed responses at each time point are superimposed on the individual profiles in each panel.*

A full account of the implications of such ***missing data*** on inference and of methods for valid analysis in their presence is the subject of an ***entire course***. Accordingly, we restrict our attention here to implications of missingness for the analysis methods we have discussed; these will also be relevant to those in the next chapter. Whether or not proceeding with an analysis using the ***observed data*** as if they were the ***intended data*** leads to ***valid inferences*** on questions of interest depends on the underlying ***mechanism*** responsible for the missingness, as we now discuss.

***NOTATION:*** We first introduce notation in the context of our longitudinal data framework useful for ***formalizing*** study of missingness. We defined in (5.25) the ***full data***

$$\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{in})^T; \tag{5.93}$$

that is, the responses ***intended*** to be collected on individual $i$ at prespecified times $t_1, \ldots, t_n$. We focus on the situation where the responses ***actually observed*** , which we denote as $\boldsymbol{Y}_i$, have components that are a ***subset*** of those of $\mathcal{Z}_i$, as in (5.26) for the hip replacement data and evidently for the AMD data.

Assume that the ***covariates*** planned to be recorded, $\boldsymbol{x}_i$, are observed for ***all individuals*** $i = 1, \ldots, m$. Of course, in practice, this is ***also not always*** the case, but this is beyond the scope of our discussion here. As is customary in this context, we consider the problem ***conditional*** on $\boldsymbol{x}_i$.

From this point of view, if we ***intend*** to collect the responses (5.93), then it is clear that the ***questions of interest*** pertain to the population mean response for $\mathcal{Z}_i$ given $\boldsymbol{x}_i$. Thus, when we adopt a ***model*** for the population mean response for $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ as we have discussed up to now, implicitly, we are ordinarily ***actually*** specifying a model for the population mean response for $\mathcal{Z}_i$ given $\boldsymbol{x}_i$.

Accordingly, the questions of interest pertain to the ***distribution of the full data***. When data are missing, our objective is thus to address those questions based on the ***observed data***.

Define the ***missing data indicators*** corresponding to the $n$ components of $\mathcal{Z}_i$ as

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

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

$$\mathcal{R}_i = (R_{i1}, \ldots, R_{in})^T, \tag{5.94}$$

so that $\mathcal{R}_i$ records whether or not $Z_{ij}$, $j = 1, \ldots, n$, is observed. Then the ***ideal full data*** are

$$(\mathcal{Z}_i, \mathcal{R}_i),$$

which, unless $\mathcal{R}_i = \mathbf{1}$, can never be fully observed (convince yourself).

***REMARK:*** Some authors refer to $\mathcal{Z}_i$ as the ***complete data*** and $(\mathcal{Z}_i, \mathcal{R}_i)$ as the ***full data***.

Let ***r*** denote a possible ***missingness pattern*** ; that is, a vector of zeroes and ones that is a possible value of $\mathcal{R}_i$ in (5.94). In general, there are $2^n$ ***possible missingness patterns***. If the only missingness patterns observed are those corresponding to ***dropout*** , and all individuals are observed at ***baseline*** (time $t_1$), then there are $n$ possible patterns

$$(1, 0, \ldots, 0), \quad (1, 1, 0, \ldots, 0), \quad \ldots, \quad (1, 1, \ldots, 1).$$

For a specific pattern of missingness $\boldsymbol{r}$, write $\mathcal{Z}_{(\boldsymbol{r})i}$ to denote the part of $\mathcal{Z}_i$ that is observed, and $\mathcal{Z}_{(\bar{\boldsymbol{r}})i}$ to denote the part that is missing. Then (convince yourself), we can represent the data that we ***actually get to see*** as

$$(\mathcal{Z}_{(\mathcal{R}_i)i}, \mathcal{R}_i), \quad i = 1, \ldots, m. \tag{5.95}$$

We have been referring to the ***observed data*** as $\boldsymbol{Y}_i$, which we now identify with $\mathcal{Z}_{(\boldsymbol{r}_i)i}$ when $\mathcal{R}_i = \boldsymbol{r}_i$; however, strictly speaking, the missing indicators are ***also part of the observable information***. In the missing data literature, (5.95) is referred to as the ***observed data***.

Write the density of $\mathcal{R}_i$ given $\mathcal{Z}_i$ and $\boldsymbol{x}_i$ as

$$p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i) = \text{pr}(\mathcal{R}_i = \boldsymbol{r}_i|\mathcal{Z}_i = \boldsymbol{Z}_i, \boldsymbol{x}_i) \tag{5.96}$$

***MISSING DATA MECHANISMS:*** Rubin (1976) pioneered a ***hierarchical taxonomy of missing data mechanisms***, which has become standard:

- ***Missing Completely at Random (MCAR):*** The data are said to be ***MCAR*** if

  $$\text{pr}(\mathcal{R}_i = \boldsymbol{r}_i|\mathcal{Z}_i, \boldsymbol{x}_i) \quad \text{does not depend on } \mathcal{Z}_i; \tag{5.97}$$

  that is, $\mathcal{R}_i \perp\!\!\!\perp \mathcal{Z}_i$ conditional on covariates $\boldsymbol{x}_i$. Then

  $$p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i) = p(\boldsymbol{r}_i|\boldsymbol{x}_i). \tag{5.98}$$

  The MCAR mechanism is ***plausible*** in situations where it is clear that missingness has ***nothing*** to do with the issues under study; for example, human subjects ***drop out*** of a study because they move away for work or family reasons. In the hip replacement study, if the missing values at week 2 are due to faulty equipment, for example, it may be reasonable to assume that the mechanism is MCAR.

  Intuitively, under a MCAR mechanism, it ***should be possible*** to make ***valid inferences*** on the questions of interest. The observed data are ***still representative*** of the information ***intended*** to be collected; there are just ***fewer observations*** than originally planned. Thus, the main consequence of proceeding with an analysis of the observed data will be ***loss of efficiency***.

- ***Missing at Random (MAR):*** The data are said to be ***MAR*** if

  $$\text{pr}(\mathcal{R}_i = \boldsymbol{r}_i|\mathcal{Z}_i, \boldsymbol{x}_i) = \text{pr}(\mathcal{R}_i = \boldsymbol{r}_i|\mathcal{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i); \tag{5.99}$$

  that is, the probability of missingness pattern $\boldsymbol{r}_i$ as a function of $\mathcal{Z}_i$ depends ***only*** on the components of $\mathcal{Z}_i$ that are ***observed*** under $\boldsymbol{r}_i$. Then

  $$p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i) = p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i). \tag{5.100}$$

  If subjects base their decisions to ***drop out*** on their observed response values to that point, and these values ***are available to the data analyst***, then the MAR mechanism is plausible.

Intuitively, if the missingness of intended data is ***associated with*** evolving responses, and those responses reflect, for example, evolving health status, if sicker patients are ***more likely*** to drop out, then the observed data are probably ***not representative*** of the information intended to be collected. Patients who remain in the study may be the ones who are doing better on their assigned treatments; accordingly, proceeding with an analysis to address the questions of interest without taking this into account is likely to lead to ***misleading inferences***.

If ***all data*** implicated in dropout decisions are ***available*** to the data analyst, as they are if the mechanism is MAR, it should be possible to do something to "***adjust***" for the missingness on their basis in an analysis of the observed data.

- ***Missing Not at Random (MNAR):*** The data are said to be ***MNAR*** if pr($\mathcal{R}_i = \boldsymbol{r}_i | \mathcal{Z}_i, \boldsymbol{x}_i$) depends on components of $\mathcal{Z}_i$ that are ***not observed*** when $\mathcal{R}_i = \boldsymbol{r}_i$.

  Intuitively, if a MNAR mechanism governs the missingness, again, the observed data are ***not representative*** of what was intended. However, because the data that are implicated in dropout decisions are ***not available***, "***adjusting***" the analysis for missingness seems ***hopeless***.

***AMD EXAMPLE, CONTINUED:*** In the AMD study, as in most clinical or observational studies of humans with ***dropout***, it is ***unlikely*** that the missingness has ***nothing to do*** with the health status of the subjects. For example, it may well be that patients whose vision ***continues to deteriorate*** might decide to leave the study on the advice of their physicians over concerns that they are ***achieving no benefit***. Here, MCAR is clearly ***implausible***.

If these decisions are based ***solely*** on inspection of the visual acuity measures up to that point, assuming a ***MAR*** mechanism would be reasonable. On the other hand, if the decisions are made based on ***other, unrecorded factors*** that might be associated with patients' ***future prognosis*** and that would be reflected in ***future visual acuity measures***, which are ***not observed***, the mechanism is ***MNAR***.

***FUNDAMENTAL CHALLENGE:*** Of course, we ***cannot determine from the available data*** which of these two explanations reflects the true state of affairs. This conundrum exemplifies the ***fundamental challenge*** of inference with missing data – the ***true missingness mechanism*** is ***not identifiable from the observed data***. Accordingly, whether or not it is plausible to assume that the mechanism is MCAR or MAR, under which methods for achieving ***valid inferences*** on questions of interest based on the observed data are fairly straightforward, ***cannot be determined*** from the data.

Ordinarily, subject-matter expertise and knowledge is incorporated to justify the assumption of MCAR or MAR; however, it remains an ***unverifiable assumption***.

The upshot is that applying the longitudinal analysis methods we have discussed so far and will discuss in the remainder of the course to the observed data when there is missingness ***without acknowledgment*** of this complication can lead to ***misleading inferences***.

A ***full course*** on analysis in the presence of missing data examines this issue in ***excruciating detail***. Here, we focus on one key result that speaks directly to the validity of carrying out an analysis of the observed data using the methods in this and the next chapter under the ***assumption of MAR***.

***OBSERVED DATA LIKELIHOOD:*** Consider the ***joint density*** of the ***ideal full data*** $(\mathcal{Z}_i, \mathcal{R}_i)$, which we write as

$$p(\boldsymbol{Z}_i, \boldsymbol{r}_i|\boldsymbol{x}_i) = p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i)\, p(\boldsymbol{Z}_i|\boldsymbol{x}_i). \tag{5.101}$$

In (5.101), we have ***factorized*** the density in to the product of two terms.

- The first term on the right hand side of (5.101), $p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i)$, is the density of the ***missingness indicator*** $\mathcal{R}_i$ given the full data $\mathcal{Z}_i$ and covariates $\boldsymbol{x}_i$. As above, depending on the missingness mechanism, this density might ***simplify*** ; we discuss this momentarily.
- The second term on the right hand side, $p(\boldsymbol{Z}_i|\boldsymbol{x}_i)$, is the density of the ***intended, full data*** given covariates. As discussed above, we now see that, from the perspective of the missing data framework, the models we have written for $E(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ and $\text{var}(\boldsymbol{Y}_i|\boldsymbol{x}_i)$ in (5.58), and indeed for the density $p(\boldsymbol{y}_i|\boldsymbol{x}_i)$ under the assumption of ***normality*** in (5.29), are really models ***implicitly*** reflecting our beliefs about the density of the ***full data*** $\mathcal{Z}_i$ given $\boldsymbol{x}_i$.

Thus, the ***ML methods*** derived in Section 5.3 correspond to assuming that $p(\boldsymbol{Z}_i|\boldsymbol{x}_i)$ in (5.101) is the *n*-***variate normal density***, depending on population mean and covariance parameters $\eta = (\beta^T, \xi^T)^T$.

In principle, we could also adopt a model for the ***density of the missingness mechanism*** , involving a parameter $\psi$, say. Thus write the assumed model for (5.101) as

$$p(\boldsymbol{Z}_i, \boldsymbol{r}_i|\boldsymbol{x}_i; \psi, \eta) = p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i; \psi)\, p(\boldsymbol{Z}_i|\boldsymbol{x}_i; \eta). \tag{5.102}$$

For $\mathcal{R}_i = \boldsymbol{r}_i$, we can ***partition*** $\mathcal{Z}_i$ as above into ***observed and missing components*** as $(\mathcal{Z}_{(\boldsymbol{r}_i)i}, \mathcal{Z}_{(\bar{\boldsymbol{r}}_i)i})$. Accordingly, we can write (5.102) as

$$p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{r}_i|\boldsymbol{x}_i; \psi, \eta) = p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{x}_i; \psi)\, p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}|\boldsymbol{x}_i; \eta).$$

It follows that we can obtain the joint density of the ***observed*** component and $\mathcal{R}_i$ as

$$\begin{aligned} p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{r}_i|\boldsymbol{x}_i;\psi,\eta) &= \int p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{r}_i|\boldsymbol{x}_i;\psi,\eta)\, d\boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i} \\ &= \int p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{x}_i;\psi)\, p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}|\boldsymbol{x}_i;\eta)\, d\boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}. \end{aligned} \tag{5.103}$$

This is the density of the ***observed data*** $(\mathcal{Z}_{(\mathcal{R}_i)i}, \mathcal{R}_i)$ in (5.95) as discussed above.

Now ***under MAR***, from (5.100), the first term in the integrand of (5.103) satisfies

$$p(\boldsymbol{r}_i|\boldsymbol{Z}_i, \boldsymbol{x}_i;\psi) = p(\boldsymbol{r}_i|\boldsymbol{Z}_{\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}, \boldsymbol{x}_i;\psi) = p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i;\psi).$$

Substituting in (5.103), we obtain

$$\begin{aligned} p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{r}_i|\boldsymbol{x}_i;\psi,\eta) &= \int p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i;\psi)\, p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}|\boldsymbol{x}_i;\eta)\, d\boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i} \\ &= p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i;\psi) \int p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i}|\boldsymbol{x}_i;\eta)\, d\boldsymbol{Z}_{(\bar{\boldsymbol{r}}_i)i} \\ &= p(\boldsymbol{r}_i|\boldsymbol{Z}_{(\boldsymbol{r}_i)i}, \boldsymbol{x}_i;\psi)\, p(\boldsymbol{Z}_{(\boldsymbol{r}_i)i}|\boldsymbol{x}_i;\eta) \end{aligned} \tag{5.104}$$

Suppose now that we have a sample of ***observed data*** from $m$ individuals, $(\mathcal{Z}_{(\mathcal{R}_i)i}, \mathcal{R}_i)$, $i = 1, \ldots, m$, as in (5.95). Consider the form of the ***likelihood*** for the parameters $(\psi^T, \eta^T)^T$ based on the observed data, often called the ***observed data likelihood***. From (5.104), the ***contribution to the likelihood*** for an individual $i$ with $\mathcal{R}_i = \boldsymbol{r}$ is

$$p(\mathcal{Z}_{(\boldsymbol{r})i}, \boldsymbol{r}|\boldsymbol{x}_i;\psi,\eta) = p(\boldsymbol{r}|\mathcal{Z}_{(\boldsymbol{r})i}, \boldsymbol{x}_i;\psi)\, p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta); \tag{5.105}$$

when $\boldsymbol{r} = \boldsymbol{1}$, in fact $\mathcal{Z}_{(\boldsymbol{r})i} = \mathcal{Z}_i$. It follows that the contribution to the likelihood for the $i$th individual can be written

$$\prod_{\boldsymbol{r}} p(\mathcal{Z}_{(\boldsymbol{r})i}, \boldsymbol{r}|\boldsymbol{x}_i;\psi,\eta)^{I(\mathcal{R}_i=\boldsymbol{r})} = \prod_{\boldsymbol{r}} p(\boldsymbol{r}|\mathcal{Z}_{(\boldsymbol{r})i}, \boldsymbol{x}_i;\psi)^{I(\mathcal{R}_i=\boldsymbol{r})} p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta)^{I(\mathcal{R}_i=\boldsymbol{r})}, \tag{5.106}$$

where the product is over ***all possible missingness patterns r***. The ***observed data likelihood*** is then the product over $i = 1, \ldots, m$ of terms (5.106).

***IGNORABILITY:*** Assume that the parameters $\psi$ and $\eta$ are ***variation independent*** in the sense that their possible values lie in a ***rectangle***, so that the range of $\eta$ is ***the same*** for all possible values of $\psi$, and vice versa. This is often called the ***separability condition***. ***Similar assumptions*** are often made in statistical modeling more generally without comment.

Under the separability condition, there is ***no information*** about the ***parameter of interest***, $\eta$, in the first term on the right hand side of (5.106). Thus, for the purpose of maximizing the likelihood to make inference on $\eta$, we can ***ignore*** this term. Accordingly, we need only maximize in $\eta$

$$\prod_{i=1}^{m} \prod_{\boldsymbol{r}} p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta)^{I(\mathcal{R}_i=\boldsymbol{r})} \quad \text{or equivalently} \quad \sum_{i=1}^{m} \sum_{\boldsymbol{r}} I(\mathcal{R}_i = \boldsymbol{r}) \log p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta). \tag{5.107}$$

In fact, under ***ignorability*** and ***separability*** , it is common to refer to (5.107) as the ***observed data likelihood (loglikelihood)***.

Now consider (5.107) from the perspective of the ML approach in Section 5.3. As we have noted, in the context of intending to collect ***full data*** at prespecified time points, the spirit of the model for the observed response vector $\boldsymbol{Y}_i$ conditional on $\boldsymbol{x}_i$ we have discussed is that it ***really reflects*** a model $\mathcal{Z}_i$ given $\boldsymbol{x}_i$. That is, the ***questions of interest*** are formulated within a model for the data we ***intend*** to collect.

From this perspective, as noted above, we are thus assuming that the distribution of $\mathcal{Z}_i$ given $\boldsymbol{x}_i$ is ***n-variate normal***. If ***no data were missing*** , the likelihood for $\eta$ would be the product of the individual $n$-variate normal densities dictated by our assumptions on the conditional (on $\boldsymbol{x}_i$) population mean and covariance structure.

When some of the intended observations ***are missing*** , and the missingness mechanism is assumed to be ***MAR*** (which, of course, we ***cannot verify*** from the data), the contribution

$$p(\mathcal{Z}_{(\boldsymbol{r})i}|\boldsymbol{x}_i;\eta)$$

for individual $i$ with $\mathcal{R}_i = \boldsymbol{r}$ is the density of the ***corresponding subvector*** of $\mathcal{Z}_i$. As is ***well-known*** , any ***subvector*** of a ***multivariate normal*** random vector is ***itself multivariate normal*** with mean vector and covariance matrix corresponding to the components contained in the subvector; an example of the latter was demonstrated in (5.27) for the hip replacement study.

Accordingly, when some responses are ***missing*** , and we are willing to believe the assumption of ***MAR*** and the ***separability condition*** , we can ***ignore*** the first term on the right hand side of (5.106), and thus we can regard the likelihood (5.30) and (5.31) as the ***observed data likelihood***. That is, the ***usual analysis*** we carry out to estimate $\beta$ and $\xi$ under the assumption the response vectors $\boldsymbol{Y}_i$ are $n_i$-variate normal conditional on $\boldsymbol{x}_i$ corresponds to the likelihood analysis we would perform both if the ***full data*** were observed (i.e., $\boldsymbol{r} = \boldsymbol{1}$ for all $m$ individuals) and with ***missing data*** (i.e., some individuals missing some components of $\mathcal{Z}_i$) ***under MAR***.

The first term on the right hand side of (5.106) represents the ***missing data mechanism*** under MAR. Under these conditions, then, if interest is ***solely*** in the parameters $\beta$ and $\xi$, there is ***no need to model and fit*** the missing data mechanism.

***KEY RESULT:*** The usual conclusion from these developments is thus that, ***under MAR***, we expect the usual analysis to yield ***valid inferences*** on $\beta$ and $\xi$. However, we must be careful to qualify what we mean by "***valid inferences***."

- We emphasize that, for the usual analysis to yield valid inference, ***both*** (i) the assumption that the distribution of $\mathcal{Z}_i$ given $\boldsymbol{x}_i$ is ***multivariate normal*** with mean and covariance structure ***correctly specified and*** (ii) the assumption of ***MAR*** must hold. If either of these assumptions is not true, then it is ***no longer the case*** that the inferences are necessarily valid.
- The estimators for $\beta$ and $\xi$ obtained by maximizing (5.107) in $\eta$ are ***identical*** to those obtained by maximizing (5.106) (under ***separability***). The estimators so obtained will be ***consistent*** for the true values of these parameters assuming, of course, that the ***full data model*** is ***correctly specified***.
- ***Likelihood ratio tests*** comparing ***nested models*** for the ***full data*** based on the statistic in (5.89) will also be ***valid***, as, under ***separability***, the missingness mechanism in (5.106) would have been estimated ***identically*** under the "full" and "reduced" full data models and thus ***cancels*** in the final test statistic.

***WRINKLE:*** Although these results are pleasing, there is a ***catch***: obtaining an appropriate ***approximate sampling distribution*** to use as the basis for ***standard errors*** and ***Wald confidence intervals and tests*** is ***not straightforward***,, as discussed in detail by Verbeke and Molenberghs (2000, Section 17.3 and Chapter 21) and Molenberghs and Kenward (2007, Chapter 12). We focus here as we have previously on inference on $\beta$.

- Recall the Taylor series argument in (5.63) and (5.64) to derive the approximate sampling distribution for the ML estimator $\widehat{\beta}$ when the models for mean and covariance matrix are ***correctly specified***. From the vantage point of missing data, this argument was made and accordingly ***expectations*** of the quantities involved were taken acting as if the lengths $n_i$ of the $\boldsymbol{Y}_i$ were ***fixed by design***. If, as here, ***normality holds*** and the mean and covariance models are ***correct*** this argument yields the ***same*** large-$m$ approximate sampling distribution for $\widehat{\beta}$ as does finding the ***expected information matrix*** for $(\widehat{\beta}^T, \widehat{\xi}^T)^T$ and inverting it, where the expectation is taken from this perspective.

- Specifically, recall the definitions

$$\boldsymbol{A}_m = m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\boldsymbol{X}_i, \quad \boldsymbol{E}_m = m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\{\partial/\partial\boldsymbol{\xi}\boldsymbol{V}_i^{-1}(\xi_0,\boldsymbol{x}_i)\}(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta_0).$$

With $\xi$ $(r+s\times 1)$, using the results for ***matrix differentiation*** in Appendix A, $\boldsymbol{E}_m$ is in fact the $(p\times r+s)$ matrix with $k$th column

$$-m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\{\partial/\partial\xi_k\boldsymbol{V}_i^{-1}(\xi_0,\boldsymbol{x}_i)\}\boldsymbol{V}_{0i}^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta_0), \tag{5.108}$$

Thus, the ***observed information matrix***; that is, the negative of the matrix of second partial derivatives of the loglikelihood (5.31), is

$$\begin{pmatrix} \boldsymbol{A}_m & -\boldsymbol{E}_m \\ -\boldsymbol{E}_m^T & -\boldsymbol{G}_m \end{pmatrix}, \tag{5.109}$$

where $\boldsymbol{G}_m$ is the $(r+s\times r+s)$ matrix of second partial derivatives of the loglikelihood with respect to elements of $\xi$. If we find the ***expected information matrix*** by taking the expectation of (5.109) (conditional on $\tilde{\boldsymbol{x}}$), acting as if the lengths $n_i$ of the $\boldsymbol{Y}_i$ were ***fixed by design***, then

$$E(\boldsymbol{E}_m|\tilde{\boldsymbol{x}}) = \boldsymbol{0}. \tag{5.110}$$

Then the expected information matrix is ***block diagonal*** so that, taking the inverse of the conditional expectation of (5.109), by standard likelihood theory, we are led to the result in (5.67),

$$m^{1/2}(\widehat{\beta}-\beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0},\boldsymbol{A}^{-1}), \quad \boldsymbol{A} = \lim_{m\to\infty} m^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\boldsymbol{X}_i. \tag{5.111}$$

- The foregoing argument ***hinges critically*** on the fact that ***expectation*** was taken as if the lengths $n_i$ of the $\boldsymbol{Y}_i$ were ***fixed by design***, leading to (5.111). ***However***, from the perspective of missing data, these lengths are ***not fixed in advance***; rather, they are a consequence of the ***realized pattern of missingness***. Accordingly, calculation of the expected information must ***acknowledge*** this by placing this problem in the missing framework we have just described.

- Calculation of $E(\boldsymbol{E}_m|\tilde{\boldsymbol{x}})$ or more precisely, from (5.108), the expectation of a summand in the $k$th column of $\boldsymbol{E}_m$,

$$\boldsymbol{X}_i^T\boldsymbol{V}_{0i}^{-1}\{\partial/\partial\xi_k\boldsymbol{V}_i^{-1}(\xi_0,\boldsymbol{x}_i)\}\boldsymbol{V}_{0i}^{-1}(\boldsymbol{Y}_i-\boldsymbol{X}_i\beta_0), \tag{5.112}$$

from this point of view can be accomplished via a conditioning argument where the conditioning set involves ***missingness pattern*** $\mathcal{R}_i$. The details of this formulation and argument are beyond our scope here.

- It can then be shown that the expectation of (5.112) (conditional on $\tilde{\boldsymbol{x}}$) is ***not equal to*** $\mathbf{0}$ in general, so that $E(\boldsymbol{E}_m|\tilde{\boldsymbol{x}}) \neq \mathbf{0}$ and

$$\boldsymbol{E}_m \xrightarrow{p} \boldsymbol{E},$$

for some $\boldsymbol{E} \neq \mathbf{0}$ in general. Thus, the expected information is ***not block diagonal***.

- Consequently, if we appeal to standard likelihood theory, using the formula for the ***inverse of a partitioned matrix*** in Appendix A, we obtain that, acknowledging that the $n_i$ are the result of missingness,

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}[\mathbf{0}, \{\boldsymbol{A} - \boldsymbol{E}(-\boldsymbol{G})^{-1}\boldsymbol{E}^T\}^{-1}], \quad \boldsymbol{G}_m \xrightarrow{p} \boldsymbol{G} \tag{5.113}$$

for $\boldsymbol{G}$ ($r + s \times r + s$) positive definite.

Comparing (5.113) to (5.111) shows that using the ***usual*** large sample approximation the sampling distribution of $\widehat{\beta}$ when the $\boldsymbol{Y}_i$ are ($n_i \times 1$) as a result of a MAR mechanism leads to standard errors that are ***too small*** and Wald test statistics and confidence intervals that are thus ***too optimistic*** when the differing $n_i$ are the consequence of MAR.

- As a way around this, it has thus been advocated that, instead of basing the approximate sampling distribution on the ***usual expected information matrix*** , one should base it on the ***observed information matrix*** (5.109) and obtain standard errors and other inferences based on the inversion of this matrix. This preserves the nonzero off-diagonal, providing an empirical approximation to (5.113). A practical difficulty is that most ***software packages do not*** offer this option and ***do not output*** this matrix as a by-product of the optimization of the loglikelihood.

- It has become common practice (which of course does not make it correct) to ***disregard*** this issue and to use the ***usual*** approximate sampling distribution for inference as if it were valid. Although this has the potential to yield ***misleadingly optimistic*** inferences, there are empirical examples where it does not seem to be too terrible. ***However*** it is important to be aware of this problem. Ideally, calculation of the inverse of the full observed information matrix is ***strongly preferred***.

- It goes without saying that one should ***not use*** the ***robust*** or ***empirical*** covariance matrix (5.82) in this situation. Not only does it suffer from the same drawback, it allows the possibility that the covariance matrix is ***incorrectly specified***.

***REMARK:*** Although in the particular case of a ***correctly specified model*** , ***MAR*** mechanism, and ***likelihood-based*** analysis it is possible to obtain valid inferences on the questions of interest (regarding aspects of the ***full data distribution***), it is important to recognize that this is ***not*** the case in general. Proceeding with a standard analysis in the presence of missing data can lead to substantially ***biased*** results.

Accordingly, it is essential that the data analyst think critically and realistically about possible reasons for missingness. An enormous body of literature exists on methods for achieving valid inferences in the presence of missing data. Verbeke and Molenberghs (2000, Chapters 14-21) and Molenberghs and Kenward (2007) offer extensive discussion of methods for handling missing data in longitudinal data analysis, including alternative approaches under MAR and methods when it is not possible to assume MAR (so that the mechanism is assumed to be MNAR).

***REMARK:*** Contrary to widespread belief, analyses based on so-called ***Big Data*** are ***not*** somehow ***exempt*** from the issues that arise because of missing data. For example, if we have data from ***electronic health records*** on millions of subjects, the fact that some subjects have ***more observations*** on the outcome of interest might reflect that they are having encounters with the health system more frequently because of ***poorer health status***. Thus, subjects with ***fewer observations*** and thus "missing data" by comparison might be healthier, so that inferences on the effects of treatments in the population of all subjects will be compromised if this is not taken into account. With such large $m$, ***bias*** (***inconsistency***) of standard estimators for population quantities of interest will swamp variance. The result will be estimators that are ***very precise*** but that are ***very far off*** from representing the true quantities of interest.