4 Modern Methods: Preliminaries

4.1 Introduction

In this chapter, we set the stage for study of modern statistical models and methods for longitudinal data analysis that are ***favored*** over the classical repeated measures analysis of variance methods we reviewed in Chapter 3.

First, we recount the ***limitations*** of the classical methods that make them less attractive in practice than modern methods. We then review basic principles of large sample theory that we will use in subsequent chapters to deduce approximate approaches to inference. In particular, we will be interested in the properties of the methods under general conditions. For example, although the methods of Chapter 5 and 6 are derived based on the assumption of ***multivariate normality*** of response vectors, we will deduce the properties of the methods even if normality does not hold. In Chapter 8, we will consider methods that do not make a specific distributional assumption. Accordingly, we briefly review the concept of ***estimating equations*** that define ***estimators*** for parameters in a model of interest and the standard general arguments that lead to approximate ***large sample properties*** of such estimators. We will take this point of view and invoke this type of argument in subsequent chapters to establish approximate large sample properties of estimators for parameters in various longitudinal data models.

4.2 Drawbacks of classical methods

The following is a summary of the major drawbacks of classical methods.

1. BALANCE: Ideally, ***univariate*** repeated measures analysis of variance methods are based on the assumption that each individual is observed at the ***same*** n time points. The multivariate repeated measures analysis of variance methods seem to depend critically on this assumption.

In experiments in settings such as agriculture and manufacturing, this may not be much of a restriction, as investigators can plan and execute experiments carefully and have a good deal of ***control***. Even in this situation, unforeseen mishaps can lead to unobserved or unrecorded observations.

In most application areas, this can be a serious limitation, particularly when the individuals are ***human subjects***. For example, in many health sciences studies, subjects are asked to return to the clinic at specific time points so that the response and other information can be ascertained. However, all subjects do not always return at precisely the time instructed, and some miss visits or, more ominously ***drop out*** of the study altogether. Even if subjects do show up as required, mishaps can also occur in processing lab samples or recording information.

It is thus more often than not ***unrealistic*** to expect the final data set to be ***balanced*** in this sense. "Fixes" such as treating all responses ***within some interval*** of an intended time point as if they all observed at that time point are possible, but are ***ad hoc***, with unknown implications for inference. "Adjusted" approximate F tests that account for imbalance have been proposed. However, it seems more productive to adopt a model framework that ***does not require*** balance, under which principled methods can be developed.

2. FORM OF OVERALL COVARIANCE MATRIX: As we have noted, the univariate methods are predicated on the induced assumption that the overall, aggregate covariance structure of a data vector is ***compound symmetric***. This may be ***too restrictive*** if ***within-individual*** sources of correlation are nonnegligible. Likewise, the multivariate methods assume ***no particular structure*** for the covariance matrix, so allow for overall patterns of covariance that may be ***very unlikely*** to arise in longitudinal data. Models and methods that offer some "middle ground" or allow among- and within-individual sources of correlation to be acknowledged and modeled faithfully would offer more ***flexibility*** to the data analyst.

3. COMMON OVERALL COVARIANCE MATRIX: Both the univariate and multivariate approaches are predicated on the assumption that the overall covariance matrix of a data vector is ***the same*** for all individuals, ***regardless*** of group or any other factor. This may or may not be a reasonable assumption, just as the assumption of ***constant variance*** in classical regression analysis is sometimes violated.

For instance, it is often the case for ***biological phenomena*** that variance ***increases*** as the magnitude of the response increases. This is the case of ***pharmacokinetics***, as in the theophylline pharmacokinetics study in ***EXAMPLE 4*** of Section 1.2. In this setting, ***within-subject*** variance is well known to ***increase*** as with the magnitude of drug concentration, often in a way that appears ***proportional*** to the ***square*** of the within-subject ***inherent trend***. In later chapters, we will characterize this phenomenon formally. Under these conditions, we would expect the ***diagonal elements*** of the overall covariance matrix of a response vector to ***change over time***.

We also saw evidence of violation of this assumption in the dental study data in Chapter 2, where the sample covariance matrices calculated separately by gender and their associated correlation matrices in (2.33) and (2.34) suggested that overall variance and the magnitude of correlations might be ***different*** by gender.

4. INCORPORATION OF COVARIATE INFORMATION: In the classical set-up, an ***among-individual covariate*** is treated as the categorical ***group*** factor, with g levels; as noted in Chapter 3, two or more covariates can be included this way by considering a factorial arrangement. However, it may be of interest to view such a covariate as ***continuous***; for example, in the guinea pig diet study, the dose groups "zero," "low," and "high" might correspond to numerical doses, 0, 100, and 500 μg, say, and interest might be in how mean response changes ***smoothly with dose***. Of course, ***time*** is also treated as categorical, involving the same limitation.

Moreover, it may be relevant to incorporate other (***among-individual***) covariate information. For example, it might be believed that a subject's ***age*** at the start of the study is implicated in his/her later response to treatment (group), suggesting that a relevant model for population mean response should depend on age as well as treatment group. Although univariate and multivariate repeated measures analysis of variance methods can be extended to accommodate this (see Sections 2.4 and 3.4 of Vonesh and Chinchilli, 1997), the way in which such covariates can be incorporated in the statistical model is ***limited***.

Incorporation of covariates that ***change over time***, such as maternal smoking status in the Six Cities study, is in principle also possible in the univariate methods; however, there are ***conceptual*** issues in dealing with such covariates, not limited to these methods, and we discuss these in later chapters.

5. QUESTIONS OF INTEREST: The classical methods emphasize ***hypothesis testing***. However, it is more often than not the case that scientific questions are ***not addressed*** by carrying out a hypothesis test. Investigators often wish to obtain ***estimates*** of meaningful quantities, such as ***rates of change*** and ***differences*** among them, along with appropriate ***measures of uncertainty*** (standard errors and confidence intervals). They might also want to evaluate the extent to which rate of change of mean response over time itself ***changes*** with an individual characteristic like age. Moreover, investigators sometimes wish to make inference about mean responses at times ***other than*** those included in a study.

Clearly, the classical methods are ***too restrictive*** to accommodate these objectives, and a ***more flexible model framework*** is required.

6. NORMALITY: The classical methods are based on the assumption that a data vector has a ***multivariate normal distribution*** with a specific mean structure and the relevant (assumed ***common*** to all individuals) covariance matrix. In the case of the univariate methods, this must hold ***exactly*** to ensure that the test statistics of interest have an F distribution with appropriate degrees of freedom, so that reliable inferences can be drawn. As we have discussed, for outcomes that are ***discrete***, this assumption is clearly inappropriate. Even for ***continuous response***, normality may not be a reasonable representation.

For outcomes that are ***not continuous***, an alternative modeling framework is needed, and we discuss such frameworks in Chapters 7 - 9. For ***continuous*** response, we would like to use methods that yield reliable inferences even if the true distribution of the response is not exactly normal.

The rest of the course is devoted to study of models and associated methods that address these limitations.

4.3 Large sample theory and estimating equations

LARGE SAMPLE THEORY: As is the case with most modern statistical models, the statistical models and methods we discuss in the remainder of the course are sufficiently ***complex*** that it is not possible to derive ***exact results***. In particular, it may not be possible to express ***estimators*** for ***parameters*** involved in the models in a ***closed form***. Rather, the estimators are defined ***implicitly*** as maximizing an ***objective function*** or a solving a set of ***equations***, as discussed momentarily.

Accordingly, it is customary to appeal to ***large sample theory*** to derive ***approximate*** results. Namely, one shows that that estimators are ***consistent*** and ***asymptotically normal***. Consistency provides assurance that, for "large" sample size, an estimator "approaches" the quantity of interest. Asymptotic normality is the basis establishing an approximate ***sampling distribution*** that can be used to derive approximate standard errors, confidence intervals, test procedures, and so on.

Appendix C presents a generic review of concepts and principles of large sample theory.

DATA STRUCTURE, RESTATED: As discussed in Section 2.2, in their most general form, the data are

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

where these are ***independent*** across i.

- The response vectors $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ are of possibly different dimension, with Y_{ij} recorded at time t_{ij}, $j = 1, \ldots, n_i$, on individual i.
- $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{in_i}^T)^T$ comprises for individual i a vector of ***within-individual covariates*** $\boldsymbol{u}_i$ describing conditions under which the Y_{ij} were collected on i along with the times t_{ij}.
- $\boldsymbol{a}_i$ is a vector of ***among-individual covariates*** for individual i.
- $\boldsymbol{x}_i = (\boldsymbol{z}_i^T, \boldsymbol{a}_i)^T$ is the full set of covariates associated with individual i.

In the study of methods for longitudinal analysis based on these data, "***large sample***" ordinarily refers to the ***number of individuals*** m being "large," while the numbers of observations per individual, n_i, remain fixed. Thus, large sample approximations are relevant to the situation where individuals are observed at intermittent, possibly prespecified time points, and the sample of individuals available is ***sufficiently large*** to provide "good" information on the population of individuals.

ESTIMATING EQUATIONS: Most of the estimators for parameters in the longitudinal data models we study in subsequent chapters can be expressed as ***solutions*** to sets of equations commonly referred to as ***estimating equations***. Such estimators are cases of a general class of estimators known as ***M-estimators***. Viewing estimators we discuss in subsequent chapters as M-estimators, we are able to deduce properties of estimators even when some model assumptions are ***not correct***.

As background for these developments, we now present a ***generic and nonrigorous*** overview of ***M-estimation*** and ***estimating equations***, to which we refer in later chapters.

M-ESTIMATOR: Let $\mathcal{U}_i$, $i = 1, \ldots, m$, be ***independent*** random vectors with cdf F_i (we may or may not know F_i). Let η ($k \times 1$) be a parameter in a statistical model for the $\mathcal{U}_i$. E.g., if F_i has density p_i, one might specify a model for p_i depending on η. Alternatively, one might specify a model only for some features of F_i, such as a mean and covariance matrix, in terms of a parameter η.

A ***M-estimator*** for η, $\widehat{\eta}$, can be defined two ways:

(1) $\widehat{\eta}$ minimizes $\sum_{i=1}^m \rho_i(\mathcal{U}_i, \eta)$, where $\rho_i(\cdot\,, \cdot)$ are real-valued functions.

(2) $\widehat{\eta}$ is the root of a ($k \times 1$) set of estimating equations such that

$$\sum_{i=1}^{m} \boldsymbol{\Psi}_i(\mathcal{U}_i, \widehat{\eta}) = \boldsymbol{0}, \tag{4.1}$$

where $\boldsymbol{\Psi}_i(\cdot\,, \cdot)$ are vector-valued functions taking values in k-dimensional space.

Ordinarily, $\boldsymbol{\Psi}_i$ satisfies

$$E_\eta\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \eta)\} = \mathbf{0}, \tag{4.2}$$

where E_η refers to expectation under the assumption that the parameter is equal to η.

The subscript *i* may or may not be relevant, depending on the situation:

- If we view $\mathcal{U}_i$ as ***independent and identically distributed*** (***iid***) draws from some joint distribution, then $\rho_i \equiv \rho$ and $\boldsymbol{\Psi}_i \equiv \boldsymbol{\Psi}$. In our context, letting $\mathcal{U}_i = (\boldsymbol{Y}_i^T, \boldsymbol{x}_i^T)^T$, $i = 1, \ldots, m$, this would correspond to the situation where we sample *m* individuals from a population of interest and record all of $\boldsymbol{Y}_i$, $\boldsymbol{u}_i$, and $\boldsymbol{a}_i$.
- If we view parts of $\boldsymbol{x}_i$ as ***fixed constants*** or view the problem ***conditional*** on the $\boldsymbol{x}_i$, then the subscript *i* on ρ_i and $\boldsymbol{\Psi}_i$ is meant to emphasize dependence on such *i*-dependent quantities.

We present the generic argument under the latter condition, but the same ideas apply in the iid case.

- If ρ_i is differentiable with respect to η, then a problem of type (1) implies one of type (2), where $\partial/\partial\boldsymbol{\eta}\, \rho_i(\mathcal{U}_i, \eta) = \boldsymbol{\Psi}_i(\mathcal{U}_i, \eta)$.
- However, a problem of type (2) can be posed without a corresponding problem of type (1), so that these problems are ***more general***.
- If p_i is the assumed density of $\mathcal{U}_i$, then choosing $\rho_i(\cdot, \eta) = \log p_i(\cdot, \eta)$ yields ***maximum likelihood estimation*** under the assumption that p_i is the true density of $\mathcal{U}_i$.

Although we consider longitudinal methods that are motivated by the principles of ***maximum likelihood***, we view all methods from the perspective of a type (2) problem, so that estimators of interest solve ***estimating equations***. This will allow us to evaluate properties even when the assumptions leading to the maximum likelihood formulation do not hold.

UNBIASED ESTIMATING EQUATIONS: Suppose that η_0 is the ***true value*** of η; that is, the value of η such that the assumed model evaluated at η_0 yields the true density (or features of the density) generating the data. In general, if an assumed model depends on a parameter η, and there is a value η_0 such that the model evaluated at η_0 corresponds to (features of) the true distribution generating the data, the model is said to be ***correctly specified***.

Formal arguments regarding ***consistency*** of M-estimators are quite involved. However, it is well known that, if (4.2) holds, under suitable ***regularity conditions***, the estimator $\widehat{\boldsymbol{\eta}}$ solving (4.1) is a ***consistent estimator*** for $\boldsymbol{\eta}_0$ if

$$E\{\boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0)\} = \mathbf{0}, \tag{4.3}$$

where expectation is with respect to the true distribution of the $\boldsymbol{\mathcal{U}}_i$, and $\boldsymbol{\eta}_0$ is the unique value satisfying this requirement.

Estimating equations that satisfy these conditions are referred to as ***unbiased estimating equations***. For our purposes in this course, to ***show consistency*** of an estimator solving a set of estimating equations, it will suffice to note that the estimating equations defining it are ***unbiased***.

APPROXIMATE LARGE SAMPLE DISTRIBUTION OF M-ESTIMATOR: An approximate sampling distribution for $\widehat{\boldsymbol{\eta}}$ can be found via a standard ***Taylor series argument***. We give a ***heuristic sketch***, recognizing that technical conditions are required to validate many of the steps. See Appendix B for a review of Taylor series and notation used here.

Multiplying (4.1) by $m^{-1/2}$, we have

$$\begin{aligned}
\mathbf{0} &= m^{-1/2} \sum_{i=1}^{m} \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \widehat{\boldsymbol{\eta}}) \\
&= m^{-1/2} \sum_{i=1}^{m} \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) + \left\{ m^{-1} \sum_{i=1}^{m} \partial/\partial \boldsymbol{\eta}^T \, \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*) \right\} m^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0),
\end{aligned}$$

where $\boldsymbol{\eta}_*$ is a value between $\widehat{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}_0$, and $\partial/\partial \boldsymbol{\eta}^T \, \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*)$ is the $(k \times k)$ matrix whose ℓth row is

$$\{\partial/\partial \eta_1 \boldsymbol{\Psi}_{i\ell}(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*), \ldots, \partial/\partial \eta_k, \boldsymbol{\Psi}_{i\ell}(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*)\},$$

and $\boldsymbol{\Psi}_{i\ell}(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_*)$ is the ℓth element of $\boldsymbol{\Psi}_i$, $\ell = 1, \ldots, k$.

As $\widehat{\boldsymbol{\eta}}$ is consistent, it is possible to define technical conditions such that $\boldsymbol{\eta}_*$ may be replaced by $\boldsymbol{\eta}_0$ in the partial derivative matrix, and the ***weak law of large numbers*** yields

$$m^{-1} \sum_{i=1}^{m} \partial/\partial \boldsymbol{\eta}^T \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) - m^{-1} \sum_{i=1}^{m} E\left\{ \partial/\partial \boldsymbol{\eta}^T \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) \right\} \xrightarrow{p} \mathbf{0},$$

where the expectation is with respect to the true distribution of $\boldsymbol{\mathcal{U}}_i$. Thus,

$$\mathbf{0} \approx m^{-1/2} \sum_{i=1}^{m} \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) + \left[m^{-1} \sum_{i=1}^{m} E\left\{ \partial/\partial \boldsymbol{\eta}^T \, \boldsymbol{\Psi}_i(\boldsymbol{\mathcal{U}}_i, \boldsymbol{\eta}_0) \right\} \right] m^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0). \tag{4.4}$$

Assuming that the inverse exists, (4.4) can be rearranged as

$$m^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \approx -\left[m^{-1}\sum_{i=1}^{m} E\left\{\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\}\right]^{-1} m^{-1/2}\sum_{i=1}^{m}\boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0). \tag{4.5}$$

Assume that as $m \to \infty$,

$$m^{-1}\sum_{i=1}^{m} E\left\{\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\} \to \boldsymbol{A},$$

say, a nonsingular constant matrix. Now apply the ***central limit theorem*** to the rightmost term on the right hand side of (4.5). As each summand has mean **0**, this term ***converges in distribution*** to a ***multivariate normal random vector*** with mean **0** and covariance matrix

$$\lim_{m\to\infty} m^{-1}\sum_{i=1}^{m} E\left\{\boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\} = \boldsymbol{B},$$

say. ***Slutsky's theorem*** then yields

$$m^{1/2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}\{\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^T\}. \tag{4.6}$$

The notation in (4.6) is ***shorthand*** for the fact that the expression on the left hand side ***converges in distribution*** to a normal random vector with mean zero and covariance matrix as shown on the right hand side.

We can use the result (4.6) to deduce an ***approximate sampling distribution*** for $\widehat{\boldsymbol{\eta}}$. Define

$$\boldsymbol{A}_m = m^{-1}\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0), \quad \boldsymbol{B}_m = m^{-1}\sum_{i=1}^{m}\boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \boldsymbol{\eta}_0).$$

Then from (4.6) and using the weak law of large numbers, we have the approximate result

$$\widehat{\boldsymbol{\eta}} \,\dot{\sim}\, \mathcal{N}\{\boldsymbol{\eta}_0, m^{-1}\boldsymbol{A}_m^{-1}\boldsymbol{B}_m(\boldsymbol{A}_m^{-1})^T\}, \tag{4.7}$$

where $\dot{\sim}$ denotes "approximately distributed as." Note that in (4.7), because we have rescaled (4.6), the m^{-1} terms cancel, and the covariance matrix $m^{-1}\boldsymbol{A}_m^{-1}\boldsymbol{B}_m(\boldsymbol{A}_m^{-1})^T$ can be written as

$$\left\{\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\}^{-1}\left\{\sum_{i=1}^{m}\boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\}\left\{\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \boldsymbol{\eta}_0)\right\}^{-1\,T}. \tag{4.8}$$

Substituting $\widehat{\boldsymbol{\eta}}$ into the expressions in (4.8), we arrive at the so-called ***sandwich estimator*** or ***robust estimator*** for the covariance matrix of $\widehat{\boldsymbol{\eta}}$, namely

$$\left\{\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \widehat{\boldsymbol{\eta}})\right\}^{-1}\left\{\sum_{i=1}^{m}\boldsymbol{\Psi}_i(\mathcal{U}_i, \widehat{\boldsymbol{\eta}})\boldsymbol{\Psi}_i^T(\mathcal{U}_i, \widehat{\boldsymbol{\eta}})\right\}\left\{\sum_{i=1}^{m}\partial/\partial\boldsymbol{\eta}^T\, \boldsymbol{\Psi}_i(\mathcal{U}_i, \widehat{\boldsymbol{\eta}})\right\}^{-1\,T}. \tag{4.9}$$

These results can be used to derive approximate (large-m) standard errors and confidence intervals for the components of $\widehat{\boldsymbol{\eta}}$ in the usual way.

REMARKS:

- Note that this argument ***does not require*** a full assumption about the distribution of the $\mathcal{U}_i$. The argument depends only on ***consistency*** of the estimator $\widehat{\boldsymbol{\eta}}$ solving the estimating equations, which holds (under regularity conditions) if the estimating equations are ***unbiased*** as in (4.3) and suitable conditions hold to allow application of the ***weak law of large numbers***, ***Slutsky's theorem*** and the ***central limit theorem***. Such conditions ordinarily involve the existence of ***higher moments*** and ***differentiability*** of functions of the $\mathcal{U}_i$.
- Thus, the results that the sampling distribution of $\widehat{\boldsymbol{\eta}}$ can be approximated for "large" samples by (4.7) and that the covariance matrix of this approximate sampling distribution can be estimated by the ***sandwich estimator*** (4.9) are not predicated on a full distributional assumption for the data. These results hold only if certain ***regularity conditions*** like those above are satisfied.
- Moreover, even if an estimator is derived under a ***full distributional assumption***; e.g., a ***maximum likelihood estimator*** under a full ***parametric distributional assumption*** (like ***normality***), if it can be expressed as the solution to a set of estimating equations, its properties can be evaluated even if that distributional assumption ***does not hold*** exactly.

DEMONSTRATION: In subsequent chapters, we deduce the properties of estimators in various ***statistical models for longitudinal data*** under general conditions by recognizing that the estimators can be formulated as ***solutions to estimating equations***. As a prelude to this development, we demonstrate how the foregoing considerations apply in the familiar situation of ***linear regression***.

To place this in the context of our notation for longitudinal data, suppose we have m individuals, on each of whom a ***single*** scalar response Y_i is recorded, so that $n_i = 1$ for $i = 1, \ldots, m$, along with a vector $(p \times 1)$ of covariates $\boldsymbol{x}_i$, where for convenience in the following argument we take the first element of $\boldsymbol{x}_i$ to be identically equal to 1 for all i; and $m > p$. It is reasonable to assume that the pairs $(Y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, m$, are ***independent*** across i. Here, then, identify $\mathcal{U}_i = (Y_i, \boldsymbol{x}_i)$.

Suppose we ***assume*** that, for each $i = 1, \ldots, m$,

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad E(\epsilon_i | \boldsymbol{x}_i) = 0, \ \text{var}(\epsilon_i | \boldsymbol{x}_i) = \sigma^2, \tag{4.10}$$

so that we equivalently assume that

$$E(Y_i | \boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad \text{var}(Y_i | \boldsymbol{x}_i) = \sigma^2, \quad i = 1, \ldots, m. \tag{4.11}$$

In (4.10) and (4.11), we ***do not*** make a full ***distributional assumption***. The ***classical assumption*** for model (4.10) is of course that ϵ_i, $i =, \ldots, m$, are ***iid*** , so that they are independent of $\boldsymbol{x}_i$ for each i, and ***furthermore*** that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, so that $Y_i \sim \mathcal{N}(\boldsymbol{x}_i^T\beta, \sigma^2)$, .

Thus, in postulating the model in (4.10) and (4.11), while we are willing to assume ***constant variance*** for all i, we are not willing to assume ***normality***. Instead, we are willing only to make the assumption about the ***first two moments*** of the conditional distribution of Y_i given $\boldsymbol{x}_i$ given in (4.11).

Now suppose that, ***in truth*** , the distribution of Y_i given $\boldsymbol{x}_i$ has first two moments of the form

$$E(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\beta, \quad \text{var}(Y_i|\boldsymbol{x}_i) = \sigma^2 g^2(\boldsymbol{x}_i), \quad i = 1, \ldots, m, \tag{4.12}$$

for some function $g(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$. Thus, ***in truth*** , instead of being as in (4.10), the actual model generating the data is

$$Y_i = \boldsymbol{x}_i^T\beta + \epsilon_i, \quad E(\epsilon_i|\boldsymbol{x}_i) = 0, \ \text{var}(\epsilon_i|\boldsymbol{x}_i) = \sigma^2 g^2(\boldsymbol{x}_i); \tag{4.13}$$

note that the ϵ_i are ***not*** iid. Suppose that β_0 and σ_0^2 are the ***true values*** of β and σ^2 in (4.12) and (4.13) generating the observed data.

Under the ***assumed model*** in (4.10) and (4.11), it is natural to estimate β and σ by the ***ordinary least squares*** (OLS) estimator and the usual ***residual mean square***

$$\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \left(\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} \sum_{i=1}^m \boldsymbol{x}_i Y_i, \quad \widehat{\sigma}^2 = (m-p)^{-1}\sum_{i=1}^m (Y_i - \boldsymbol{x}_i^T\widehat{\beta})^2, \tag{4.14}$$

where $\boldsymbol{X}$ is the (***full rank***) $(m \times p)$ matrix with rows $\boldsymbol{x}_i^T$, $i = 1, \ldots, m$. Of course, even if the true conditional moments of Y_i given $\boldsymbol{x}_i$ are as in (4.12) and (4.13), we can deduce immediately that the OLS estimator $\widehat{\beta}$ is ***consistent***. If we view this situation as ***conditional*** on the $\boldsymbol{x}_i$, then if

$$m^{-1}\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T = m^{-1}\boldsymbol{X}^T\boldsymbol{X} \ \to \boldsymbol{A}, \tag{4.15}$$

say, the ***weak law of large numbers*** implies

$$m^{-1}\sum_{i=1}^m \boldsymbol{x}_i Y_i - m^{-1}\sum_{i=1}^m \boldsymbol{x}_i(\boldsymbol{x}_i^T\beta_0) \xrightarrow{p} 0,$$

so that

$$m^{-1}\sum_{i=1}^m \boldsymbol{x}_i Y_i \xrightarrow{p} \boldsymbol{A}\beta_0$$

and thus

$$\widehat{\beta} \xrightarrow{p} \boldsymbol{A}^{-1}\boldsymbol{A}\beta_0 = \beta_0.$$

Thus, even though the constant variance assumption ***does not hold*** , the OLS estimator is ***nonetheless*** a ***consistent estimator*** for the true value β_0.

Alternatively, from (4.14), we can write $\widehat{\beta}$ as the solution to the ***estimating equations***

$$\sum_{i=1}^{m}(Y_i - \boldsymbol{x}_i^T\beta)\boldsymbol{x}_i = \boldsymbol{0}, \tag{4.16}$$

the usual ***normal equations***, so that, identifying $\eta = (\beta^T, \sigma^2)^T$,

$$\boldsymbol{\Psi}_i(Y_i, \boldsymbol{x}_i, \eta) = (Y_i - \boldsymbol{x}_i^T\beta)\boldsymbol{x}_i.$$

It is straightforward to observe that, under (4.12) and (4.13), $E_\eta\{\boldsymbol{\Psi}_i(Y_i, \boldsymbol{x}_i, \eta)|\boldsymbol{x}_i\} = \boldsymbol{0}$ so that $E_\eta\{\boldsymbol{\Psi}_i(Y_i, \boldsymbol{x}_i, \eta)\} = \boldsymbol{0}$ for all η, and thus $E\{\boldsymbol{\Psi}_i(Y_i, \boldsymbol{x}_i, \eta_0)\} = \boldsymbol{0}$, so that (4.16) is indeed an ***unbiased estimating equation*** , and consistency of $\widehat{\beta}$ is expected.

We now demonstrate how the large sample distribution of $m^{1/2}(\widehat{\beta} - \beta_0)$ can be derived using the generic estimating equation argument. From (4.16), we have

$$\begin{aligned}\boldsymbol{0} &= m^{-1/2}\sum_{i=1}^{m}(Y_i - \boldsymbol{x}_i^T\widehat{\beta})\boldsymbol{x}_i \\ &= m^{-1/2}\sum_{i=1}^{m}(Y_i - \boldsymbol{x}_i^T\beta_0)\boldsymbol{x}_i - \left(m^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^T\right) m^{1/2}(\widehat{\beta} - \beta_0).\end{aligned} \tag{4.17}$$

(It is in fact the case that (4.17) is an ***exact*** equality.) From (4.13), define

$$\delta_i = \frac{Y_i - \boldsymbol{x}_i^T\beta_0}{\sigma_0 g(\boldsymbol{x}_i)}, \quad \text{var}(\delta_i|\boldsymbol{x}_i) = 1.$$

Thus, rearranging (4.17),

$$m^{1/2}(\widehat{\beta} - \beta_0) = \sigma_0\left(m^{-1}\sum_{i=1}^{m}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1} m^{-1/2}\sum_{i=1}^{m} g(\boldsymbol{x}_i)\boldsymbol{x}_i\delta_i. \tag{4.18}$$

By the ***central limit theorem***, letting $w_i = 1/g(\boldsymbol{x}_i)^2$ and noting that

$$\text{var}\{g(\boldsymbol{x}_i)\boldsymbol{x}_i\delta_i|\boldsymbol{x}_i\} = E\{g^2(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\delta_i^2|\boldsymbol{x}_i\} = w_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T,$$

$$m^{-1/2}\sum_{i=1}^{m} g(\boldsymbol{x}_i)\boldsymbol{x}_i\delta_i \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \sigma_0^2\boldsymbol{B}), \quad \boldsymbol{B} = \lim_{m\to\infty} m^{-1}\sum_{i=1}^{m} w_i^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T.$$

Using (4.15), we conclude by ***Slutsky's theorem*** from (4.18) that

$$m^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \sigma_0^2\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}). \tag{4.19}$$

Thus, if we let

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & w_m \end{pmatrix} = \text{diag}(w_1, \ldots, w_m)$$

then

$$m^{-1} \sum_{i=1}^{m} w_i^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^T = m^{-1} \boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X},$$

and (4.19) yields

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\{\boldsymbol{\beta}_0, \sigma_0^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X}) (\boldsymbol{X}^T \boldsymbol{X})^{-1}\}, \tag{4.20}$$

where the m^{-1} all cancel when $\widehat{\boldsymbol{\beta}}$ is placed on this scale.

REMARKS:

- Note that, if ***in truth*** $g(\boldsymbol{x}) \equiv 1$ for all $\boldsymbol{x}$, so that $w_i \equiv 1$ for all i, then the assumption of ***constant variance*** is ***correct***, and $\boldsymbol{W} = \boldsymbol{I}_m$, an identity matrix. In this case, (4.20) reduces to

  $$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\{\boldsymbol{\beta}_0, \sigma_0^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}\},$$

  which is identical to the familiar exact, classical result when Y_i given $\boldsymbol{x}_i$ is ***normally distributed***. Thus, the exact, ***classical linear model theory*** normal sampling distribution result for the OLS estimator holds ***approximately*** in large samples even if normality of the response does not hold.

- If instead, ***in truth*** $g(\boldsymbol{x})$ depends on $\boldsymbol{x}$, so that w_i are different for each i depending on $\boldsymbol{x}_i$, then (4.20) is an approximate version of the exact linear model theory result when the assumption of ***constant variance does not hold***. Note that this result is ***not immediately useful*** in practice; if we do not know $g(\cdot)$, we cannot calculate the matrix $\boldsymbol{W}$, and, as discussed next, $\widehat{\sigma}^2$ will not be a consistent estimator for σ_0^2. We take up the analogous issue to this in the general longitudinal data setting in the next chapter.

- We can also consider the OLS estimator $\widehat{\sigma}^2$ for σ^2 in (4.14) from the point of view of ***estimating equations***. Recall from classical linear model theory that division here is by $(m-p)$ rather than m to achieve an exactly ***unbiased*** estimator with finite sample size m in the case of ***constant variance***. Because $m/(m-p) \to 1$ as $m \to \infty$, consider instead the estimator

  $$m^{-1} \sum_{i=1}^{m} (Y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})^2.$$

This estimator can be written as the solution to the ***estimating equation***

$$\sum_{i=1}^{m}\{(Y_i - \boldsymbol{x}_i^T\beta)^2 - \sigma^2\} = 0, \tag{4.21}$$

jointly with the equation (4.16). If, ***in truth***, constant variance does hold, so that $g(\boldsymbol{x}) \equiv 1$ for all $\boldsymbol{x}$, then it is clear that (4.21) is an ***unbiased estimating equation***, so we expect that $\widehat{\sigma}^2$ is a consistent estimator for σ_0^2 in this case.

However, if this assumption is ***incorrect***, and $\text{var}(Y_i|\boldsymbol{x}_i) = \sigma^2 g^2(\boldsymbol{x}_i)$ as in (4.12) and (4.13), then it is straightforward that

$$E_\eta\{(Y_i - \boldsymbol{x}_i^T\beta)^2|\boldsymbol{x}_i\} = \sigma^2 g^2(\boldsymbol{x}_i),$$

and thus the summand in the estimating equation (4.21) ***does not*** have conditional (on $\boldsymbol{x}_i$) expectation 0, so that the estimating equation is not ***unbiased***. Intuitively, the meaning of the parameter σ^2 when the true variance is not constant is ***different from*** that when true variance is constant, so this is not surprising.

SUMMARY: In subsequent chapters, we derive large sample results using generalizations of this argument, without making ***full distributional assumptions*** such as normality on the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{x}_i$ but rather making assumptions only on conditional moments. Thus, the results will be broadly applicable as long as the number of individuals m is not too small.