7 Generalized and Nonlinear Models for Univariate Response

7.1 Introduction

The models for longitudinal data we have discussed so far are suitable for responses that are or can be viewed as approximately ***continuous***. Moreover, the models incorporate the assumption that the overall population mean (PA perspective) and inherent individual trajectories (SS perspective) can be approximated by representations that are ***linear*** in parameters.

Such models are clearly ***unsuitable*** for ***discrete responses***, such as binary or categorical outcomes or responses whose values are small counts, for which standard models are not linear. They are also not appropriate for continuous outcomes when population or individual trajectories ***cannot*** be well-approximated by linear functions of parameters.

For instance, in ***EXAMPLE 4*** of Chapter 1 on the pharmacokinetics of theophylline, the ***mechanistic model*** for (continuous) drug concentration at time t within an ***individual subject*** in (1.3) and (2.1), derived from the one-compartment representation of the body in Figure 1.6, is a natural way to represent the ***inherent individual trajectory*** of drug concentrations over time. As we review shortly, this model is ***nonlinear*** in individual-specific parameters k_a, Cl, and V reflecting absorption rate; drug clearance, which has to do with how the drug is eliminated from the body; and volume of distribution, which is related to the extent to which the drug distributes through the body, respectively. These individual-specific parameters thus have ***meaningful scientific interpretations***, so an appropriate analysis should incorporate the mechanistic model.

Likewise, in ***EXAMPLE 6*** of Chapter 1, the Six Cities Study, the wheezing response is ***binary***. Thus, if $Y_{ij} = 0$ if the ith child is not wheezing at time (age) j and 1 if s/he is, the "typical" or ***population mean*** response at age j given covariates is $\text{pr}(Y_{ij} = 1|\boldsymbol{x}_i)$. Popular regression models for probabilities, such as ***logistic*** or ***probit*** regression models, are ***nonlinear*** in parameters, as we demonstrate in the next section.

Clearly, ***population-averaged*** and ***subject-specific*** models for longitudinal data in these situations are required. In this chapter, as a prelude to discussing these longitudinal models and associated inferential methods, we review classical ***nonlinear regression models*** for ***univariate response***.

7.2 Nonlinear mean-variance models

GENERAL NONLINEAR MODEL: We consider the following situation and notation. Let Y denote a scalar response of interest and $\boldsymbol{x}$ denote a vector of covariates, and suppose we observe $(Y_j, \boldsymbol{x}_j)$, $j = 1, \ldots, n$, ***independent*** across j. Here, we use j as the index in anticipation of our discussion of SS nonlinear models; see below. In this chapter, we focus on models of the general form

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta), \qquad \text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_j), \quad j = 1, \ldots, n. \tag{7.1}$$

where $\theta = (\sigma^2, \delta^T)^T$ is $(r \times 1)$, and β is $(p \times 1)$.

- In (7.1), $f(\boldsymbol{x}, \beta)$ is a ***nonlinear*** function of parameters β depending on the covariates $\boldsymbol{x}_j$.
- $g^2(\beta, \delta, \boldsymbol{x}_j)$ is the ***variance function***, which allows variance to be ***nonconstant*** over j in a systematic fashion depending on $\boldsymbol{x}_j$ and which is also possibly ***nonlinear*** in β and possibly additional ***variance parameters*** δ. Here, σ^2 is a ***scale parameter***.

EXAMPLES: The model (7.1) is used to represent a variety of situations, depending on the context.

- As noted above, when Y_j is ***binary*** taking values 0 or 1, $E(Y_j|\boldsymbol{x}_j) = \text{pr}(Y_j = 1|\boldsymbol{x}_j)$, and a natural model is the classical ***logistic regression model***

  $$f(\boldsymbol{x}_j, \beta) = \frac{\exp(\boldsymbol{x}_j^T\beta)}{1 + \exp(\boldsymbol{x}_j^T\beta)}, \quad \text{or equivalently} \quad \text{logit}\{f(\boldsymbol{x}_j, \beta)\} = \log\left\{\frac{f(\boldsymbol{x}_j, \beta)}{1 - f(\boldsymbol{x}_j, \beta)}\right\} = \boldsymbol{x}_j^T\beta, \tag{7.2}$$

  where $\text{logit}(u) = \log\{u/(1-u)\}$. Here, then, $f(\boldsymbol{x}_j, \beta)$ represents a ***probability***.

  For binary response with mean $f(\boldsymbol{x}_j, \beta)$, it is immediate that we ***must have***

  $$\text{var}(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)\{1 - f(\boldsymbol{x}_j, \beta)\}, \tag{7.3}$$

  so that $\sigma^2 \equiv 1$, and there is no unknown parameter δ. Implicit is the assumption that the binary response can be ascertained ***perfectly***, with no potential ***misclassification error***, which is analogous to measurement error in the case of binary response.

  This situation might arise in a study where the jth of n participants has baseline covariates $\boldsymbol{x}_j$, and the single binary response Y_j is ascertained on each individual j at some follow-up time. Here, $\boldsymbol{x}_j$ is an ***among-individual*** covariate, and interest focuses on the probability of positive response in the population as a function of these covariates.

Thus, the ***scope of inference*** is the ***entire population*** from which the sample of n individuals was drawn, and the parameter β has a ***PA interpretation***. The extension to a ***longitudinal study*** is if the response were ascertained ***repeatedly*** over time on each individual j.

Alternatively, the n binary responses might all be on the ***same individual*** after s/he was given different doses x_j of a drug on occasions $j = 1, \ldots, n$, where these responses are assumed to be ascertained ***sufficiently far apart*** in time to be ***approximately independent***. In this case, interest focuses on the dose-response relationship for this individual, so that the ***scope of inference*** is this ***single individual***, and x_j is a ***within-individual*** covariate. In this case, the parameter β characterizes the probability of positive response for ***this individual only*** as a function of dose.

- As discussed in Section 2.2, a model of the form (7.1) is often used to describe ***individual pharmacokinetics***. For example, from (2.1), if our focus is on a ***given individual*** who received dose D of theophylline at time 0, and Y_j is drug concentration measured on this individual at time t_j, then $\boldsymbol{x}_j = (D, t_j)$, $j = 1, \ldots, n$, and

$$f(\boldsymbol{x}_j, \boldsymbol{\beta}) = \frac{\beta_1}{\beta_3(\beta_1 - \beta_2/\beta_3)}\{\exp(-\beta_2 t_j/\beta_3) - \exp(-\beta_1 t_j)\}, \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T. \tag{7.4}$$

In (7.4), $\boldsymbol{x}_j$ has the interpretation as what we have referred to as a ***within-individual covariate*** (appended by time); we have used the notation $\boldsymbol{z}_{ij}$ for the jth such covariate on individual i in a longitudinal data context.

As noted previously, it is often further assumed that the sampling times t_j are ***sufficiently intermittent*** that ***serial correlation*** among the Y_j is ***negligible***, so that the assumption of ***independence*** of the $(Y_j, \boldsymbol{x}_j)$ over j is taken to hold ***approximately***.

Here, $\text{var}(Y_j|\boldsymbol{x}_j)$ in (7.4) reflects the ***aggregate*** variation due to the ***within-individual realization process*** and ***measurement error*** in ascertaining drug concentrations. As noted in Section 2.2, in pharmacokinetics this aggregate variance typically exhibits ***constant coefficient of variation***, so a popular ***empirical model*** for aggregate within-individual variance in practice is

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 f^2(\boldsymbol{x}_j, \beta), \tag{7.5}$$

which is of the form in (7.1) with $g^2(\beta, \delta, \boldsymbol{x}_j) = f^2(\boldsymbol{x}_j, \beta)$, so that σ is the coefficient of variation (CV). In (7.5), there is no unknown variance parameter δ.

A common generalization of (7.5) is the so-called "***power of the mean***" variance model

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 f^{2\delta}(\boldsymbol{x}_j, \beta), \quad \delta > 0, \tag{7.6}$$

so $g^{2\delta}(\beta, \delta, \boldsymbol{x}_j) = f^{2\delta}(\boldsymbol{x}_j, \beta)$, which represents aggregate variance as proportional to an ***arbitrary power*** δ of the mean response. This is a popular model when the combined effects of effect of realization and measurement error appear to yield ***more profound*** pattern of variance than dictated by the constant CV model.

From the point of view of the ***conceptual representation*** in Chapter 2 , models like (7.5) and (7.6) are indeed ***approximations*** to a potentially more complex mechanism. To see this, write the *j*th drug concentration as

$$Y_j = f(\boldsymbol{x}_j, \beta) + e_{Pj} + e_{Mj}, \tag{7.7}$$

where as before e_{Pj} represents the within-individual deviation due to the ***realization process*** and e_{Mj} represents the ***measurement error*** deviation at time t_j, with $E(e_{Pj}|\boldsymbol{x}_j) = 0$ and $E(e_{Mj}|\boldsymbol{x}_j) = 0$. Then (7.7) of course implies $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)$ as in (7.1) and allows us to contemplate the ***contributions*** of each to the aggregate within-individual variance $\text{var}(Y_j|\boldsymbol{x}_j)$ as follows.

Many biological processes exhibit approximate ***constant CV*** or other dependence of the variance of the process on the ***level of mean response***. Here, this implies that an appropriate model for the ***variance of the realization process deviation*** is

$$\text{var}(e_{Pj}|\boldsymbol{x}_j) = \sigma_P^2 f(\boldsymbol{x}_j, \beta)^{2\delta_P}, \tag{7.8}$$

say, where δ_P might indeed be equal to 1.

As we have discussed, some measuring techniques commit errors such that the magnitude of the error is ***related*** to the size of the thing being measured. This is sometimes the case for ***assays*** used to ascertain levels of drug or other agents in blood, plasma, or other samples. In (7.7), the thing being measured at time t_j is the ***actual realized drug concentration***

$$f(\boldsymbol{x}_j, \beta) + e_{Pj}.$$

Thus, ideally, this suggests that e_{Mj} and e_{Pj} are ***correlated***, so an overall model for $\text{var}(Y_j|\boldsymbol{x}_j)$ should reflect this. However, it is well accepted in pharmacokinetics that the aggregate variance of drug concentrations is ***dominated by measurement error*** in that the deviations from the inherent drug concentration trajectory $f(\boldsymbol{x}_j, \beta)$ are "negligible" compared to those for measurement error.

From this point of view, at the level of the individual, for whom β is ***fixed***, it is common to view e_{Pj} and e_{Mj} as approximately independent and to approximate $\text{var}(e_{Mj}|\boldsymbol{x}_j)$ as depending on $f(\boldsymbol{x}_j, \beta)$, in which case a model for ***measurement error variance*** might be of the form

$$\text{var}(e_{Mj}|bx_j) = \sigma_M^2 f(\boldsymbol{x}_j, \beta)^{2\delta_M}, \tag{7.9}$$

Following these considerations and combining (7.8) and (7.9), we are led to the representation

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma_P^2 f(\boldsymbol{x}_j, \beta)^{2\delta_P} + \sigma_M^2 f(\boldsymbol{x}_j, \beta)^{2\delta_M} \tag{7.10}$$

A further approximation reflecting the belief that measurement error dominates the realization process would be to ***disregard*** e_{Pi} and thus the first term in (7.10) entirely, in which case the common models (7.5) and (7.6) can be viewed as ***representing primarily*** measurement error variance. Alternatively, these models can be viewed as a "***compromise***" approximation to (7.10).

If it is in fact believed that measurement errors are of ***similar magnitude*** regardless of the size of the thing being measured, so that e_{Mj} and e_{Pj} are reasonably taken as ***independent***, an aggregate variance model representing this is a simplification of (7.10), usually parameterized as

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2\{\delta_1 + f^{2\delta_2}(\boldsymbol{x}_j, \beta)\}, \quad \boldsymbol{\delta} = (\delta_1, \delta_2)^T, \tag{7.11}$$

so that $\sigma_P^2 = \sigma^2$ and $\sigma_M^2 = \sigma^2\delta_1$.

In this example, the ***scope of inference*** is confined to the ***single individual*** on whom the drug concentrations over time were ascertained. Here, then, β pertains to this individual only. The ***same*** modeling considerations would of course apply to ***each individual*** in a sample of m individuals on whom concentration-time data are available, as in the SS longitudinal data model framework we discuss in Chapter 9.

- Although in (7.1) we allow the dependence of the variance function on β and $\boldsymbol{x}_j$ to be arbitrary, as in the foregoing examples, it is almost always the case that if it is taken to depend on ***both*** β and $\boldsymbol{x}_j$, this dependence is solely through the ***mean response*** $f(\boldsymbol{x}_j, \beta)$.

- Note that in (7.1) it could be that the variance function depends only on covariates $\boldsymbol{x}_j$ and variance parameters δ and ***not on*** β or the mean response. For example,

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 \exp(\boldsymbol{x}_j^T \boldsymbol{\delta})$$

is a popular ***empirical model*** that allows variance to change directly with the values of covariates. Such models are widely used in econometrics.

In the most general case of model (7.1), we make no further assumptions on the distribution of Y_j given $\boldsymbol{x}_j$ beyond the first two moments. For binary response Y_j, of course, given a model $f(\boldsymbol{x}_j, \beta)$ for $E(Y_j|\boldsymbol{x}_j)$, the entire (Bernoulli) distribution of Y_j given $\boldsymbol{x}_j$ is ***fully specified***. Likewise, if we take the distribution of $Y_j|\boldsymbol{x}_j$ to be ***normal***, then given a model (7.1) the distribution is fully specified.

SCALED EXPONENTIAL FAMILY: A special case of the general model (7.1) is obtained by making the assumption that the distribution of Y_j given $\boldsymbol{x}_j$ is a member of a particular class of distributions that includes the Bernoulli/binomial and the normal with ***constant variance*** for all j. A random variable *Ylt* is said to have distribution belonging to the ***scaled exponential family*** if it has density or probability mass function

$$p(y\,;\zeta,\sigma) = \exp\left\{\frac{y\zeta - b(\zeta)}{\sigma^2} + c(y,\sigma)\right\}, \tag{7.12}$$

where ζ and σ are real-valued parameters characterizing the density, and $b(\zeta)$ and $c(y,\sigma)$ are real-valued functions.

- If σ is ***known*** (often $\sigma = 1$ in this case), then (7.12) is exactly the density of a ***one-parameter exponential family*** with ***canonical parameter*** ζ.
- It is straightforward to derive (try it) that

  $$E(Y) = b_\zeta(\zeta) = d/d\zeta\, b(\zeta), \qquad \text{var}(Y) = \sigma^2 b_{\zeta\zeta}(\zeta) = \sigma^2 d^2/d\zeta^2\, b(\zeta),$$

  so that if $E(Y) = \mu$ and $b_\zeta(\,\cdot\,)$ is a one-to-one function, ζ can be regarded as a function of μ, namely, $\zeta = b_\zeta^{-1}(\mu)$,, and thus $\text{var}(Y) = \sigma^2 b_{\zeta\zeta}\{(b_\zeta^{-1}(\mu)\} = \sigma^2 g^2(\mu)$. This demonstrates that the density (7.12) induces a ***specific relationship between mean and variance***.
- Common distributions that are members of the class (7.12) are as follows:

Distribution	$b(\zeta)$	$\zeta(\mu)$	$g^2(\mu)$
Normal, constant variance	$\zeta^2/2$	μ	1
Poisson	$\exp(\zeta)$	$\log\mu$	μ
Gamma	$-\log(-\zeta)$	$-1/\mu$	μ^2
Inverse Gaussian	$-(-2\zeta)^{1/2}$	$1/\mu^2$	μ^3
Binomial	$\log(1+e^\zeta)$	$\log\{\mu/(1-\mu)\}$	$\mu(1-\mu)$

For the Poisson and binomial distributions, $\sigma = 1$. For the others, σ is a free parameter characterizing the density.

GENERALIZED (NON)LINEAR MODEL: If the distribution of $Y_j|\boldsymbol{x}_j$ has density (7.12) with $b_\zeta(\zeta_j) = f(\boldsymbol{x}_j, \beta)$, then it follows that

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta), \quad \text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2\{f(\boldsymbol{x}_j, \beta)\}, \tag{7.13}$$

for function $g^2(\cdot,)$ dictated by $b(\cdot)$, and (7.13) with this density is referred to as a ***generalized (non)linear model***.

- In (7.13), we emphasize that the implied variance function is a ***known function of the mean***.
- Model (7.13) is a slight extension of the ***generalized linear model*** , for which $\boldsymbol{x}_j$ and β enter the mean model ***only*** through the ***linear combination*** $\boldsymbol{x}_j^T\beta$, in which case we write $f(\boldsymbol{x}_j^T\beta)$.
- For a generalized linear model with $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j^T\beta)$ and $f(\cdot)$ ***monotone*** in its single argument, its ***inverse*** $f^{-1}(\cdot)$ is called the ***link function*** , and $\boldsymbol{x}_j^T\beta$ is called the ***linear predictor***. If furthermore the link function satisfies $f^{-1}(\mu) = \zeta$ for ζ as in (7.12), then it is called the ***canonical link***. There is ***no special significance*** to the canonical link as far as data analysis is concerned; e.g., there is ***no reason*** it should provide a better fitting model than some other f.
- The usual ***logistic regression model*** in (7.2) and (7.3) is a special case of a generalized linear model, arising from the simplest ***binomial distribution***, the Bernoulli. This model uses the ***canonical link*** ; the classical ***probit*** model, which instead takes

  $$f(\boldsymbol{x}_j, \beta) = \Phi(\boldsymbol{x}_j^T\beta),$$

  where $\Phi(\cdot)$ is the cdf of the standard normal distribution, is also a generalized linear model that ***does not*** use the canonical link.
- For responses in the form of (nonnegative integer) ***counts*** , as in ***EXAMPLE 5*** of the epileptic seizure study in Chapter 1, the ***Poisson distribution*** is a standard model, and the classical model for $E(Y_j|\boldsymbol{x}_j)$ is the ***loglinear model***

  $$f(\boldsymbol{x}_j, \beta) = \exp(\boldsymbol{x}_j^T\beta),$$

  with $\text{var}(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)$. This is also a generalized linear model with canonical link.
- The classical ***linear regression model*** $f(\boldsymbol{x}_j, \beta) = f(\boldsymbol{x}_j^T\beta) = \boldsymbol{x}_j^T\beta$ where $Y_j|\boldsymbol{x}_j$ is assumed ***normal*** with with ***constant variance*** is also a special case of a generalized linear model, where $f(\cdot)$ is the so-called ***identity link***.

- Despite widespread usage, there is ***no reason*** that dependence on the covariates must be through the ***linear combination*** $\boldsymbol{x}_j^T\beta$ the case except convention. For example, in dose-toxicity modeling, where the response Y_j is binary and x_j is dose given to the jth laboratory rat, modifying the usual logistic model to be

$$E(Y_j|x_j) = \frac{\exp(\beta_0 + \beta_1 x_j^{\beta_2})}{1 + \exp(\beta_0 + \beta_1 x_j^{\beta_2})}$$

  often provides a ***better fit***.

- Model (7.13) can be ***extended*** without altering the foregoing results. For example, if Y_j is the number of "successes" observed in a fixed number r_j trials with success probability $\pi(\boldsymbol{c}_j, \beta)$, say, then letting $\boldsymbol{x}_j = (r_j, \boldsymbol{c}_j^T)^T$,

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta) = r_j\pi(\boldsymbol{c}_j, \beta), \quad \text{var}(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)\{r_j - f(\boldsymbol{x}_j, \beta)\}/r_j = g^2\{f(\boldsymbol{x}_j, \beta), \boldsymbol{x}_j\}.$$

  We suppress this additional dependence of the variance function on $\boldsymbol{x}_j$ in generalized (non)linear models henceforth and continue to write the variance function as in (7.13), but all developments apply to this more general formulation.

- For distributions like the Poisson for counts or binomial for numbers of "successes," the ***scale parameter*** $\sigma^2 = 1$. However, in some circumstances the mean-variance relationship in (7.13) with $\sigma^2 = 1$ may be ***insufficient*** to represent the ***true magnitude*** of the aggregate variation in the data. ***Overdispersion*** refers to the phenomenon in which the variance of the response exceeds the nominal variance dictated by the distributional model. This can be because of ***measurement error*** or due to ***clustering***.

  For example, if r rats are placed in each of n cages, the rats in cage j are given a dose x_j of a toxic agent, and Y_j is the number of rats in cage j having an adverse reaction, then Y_j is the sum of r binary responses, one for each rat. If all rats have the ***same probability*** π_j of having an adverse reaction to the dose x_j, then Y_j is binomial with parameters r and π_j. However, if rats are heterogeneous, so that the kth rat in the cage j has probability p_{jk} of having an advese reaction, where the p_{jk} are such that $E(p_{jk}|x_j) = \pi_j$ and $\text{var}(p_{jk}|x_j) = \tau^2\pi_j(1-\pi_j)$, it can be shown (try it) that $Y_j|x_j$ is such that

$$E(Y_j|x_j) = r\pi_j, \qquad \text{var}(Y_j|x_j) = \sigma^2 r\pi_j(1-\pi_j), \tag{7.14}$$

  where σ^2 is a function of τ^2 and r.

The mean-variance model in (7.14) resembles that of the usual binomial ***except*** for the ***scale factor*** σ^2. Because there is ***additional among-rat variation*** in that all rats do not have the same probability of an adverse reaction, we might expect $\sigma^2 > 1$, which would make the variability ***more profound*** than that dictated by the binomial.

It is thus commonplace to allow a scale factor σ^2 in (7.13) to accommodate potential such ***overdispersion***.

As we discuss next, it turns out that maximum likelihood estimation of $\boldsymbol{\beta}$ in a generalized (non)linear model (7.13) under density (7.12) is equivalent to solving the same ***linear estimating equation*** that one is led to more generally from a variety of viewpoints.

7.3 Estimation of mean and variance parameters

We assume henceforth that the model for the mean $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \boldsymbol{\beta})$ in (7.1) is ***correctly specified***.

MAXIMUM LIKELIHOOD FOR THE SCALED EXPONENTIAL FAMILY: Taking the derivative of the logarithm of (7.12) with respect to $\boldsymbol{\beta}$ with ζ represented as a function of the mean (and thus of $\boldsymbol{\beta}$), using the chain rule, it is straightforward to show (verify) that the ***maximum likelihood estimator*** for $\boldsymbol{\beta}$ is the solution to the ***estimating equation***

$$\sum_{j=1}^{n} f_{\beta}(\boldsymbol{x}_j, \boldsymbol{\beta}) g^{-2}\{f(\boldsymbol{x}_j, \boldsymbol{\beta})\}\{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta})\} = \mathbf{0}, \tag{7.15}$$

where $f_{\beta}(\boldsymbol{x}_j, \boldsymbol{\beta}) = \partial/\partial\boldsymbol{\beta} f(\boldsymbol{x}_j, \boldsymbol{\beta})$ is the $(p \times 1)$ vector of partial derivatives of $f(\boldsymbol{x}_j, \boldsymbol{\beta})$ with respect to the elements of $\boldsymbol{\beta}$. Clearly, this is an ***unbiased estimating equation*** (verify).

- In the special case of (7.12) corresponding to the ***normal distribution with constant variance*** , (7.15) reduces to

$$\sum_{j=1}^{n} f_{\beta}(\boldsymbol{x}_j, \boldsymbol{\beta})\{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta})\} = \mathbf{0}, \tag{7.16}$$

  which is the estimating equation corresponding to ***ordinary nonlinear least squares***. If in fact $f(\boldsymbol{x}_j, \boldsymbol{\beta}) = \boldsymbol{x}_j^T\boldsymbol{\beta}$, a ***linear*** model, then $f_{\beta}(\boldsymbol{x}_j, \boldsymbol{\beta}) = \boldsymbol{x}_j$, and (7.16) are the usual ordinary least squares ***normal equations*** , as expected.

LINEAR ESTIMATING EQUATION FOR β: For the ***general mean-variance model*** (7.1), with ***no distributional assumptions*** beyond the two specified moments and possibly unknown variance parameters θ, the standard approach to estimation of β is by solving an obvious generalization of the ***linear estimating equation*** (7.15), given by

$$\sum_{j=1}^{n} f_{\beta}(\boldsymbol{x}_j, \beta) g^{-2}(\beta, \delta, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta)\} = \mathbf{0}, \tag{7.17}$$

jointly with estimating equations for the variance parameters θ, discussed momentarily. Obviously, (7.17) is an ***unbiased estimating equation***.

It is common to justify solving (7.17) under these conditions as follows. If the value of the ***variance function*** $g^2(\beta, \delta, \boldsymbol{x}_j)$ were ***known*** for each j, then the reciprocal of the variance function specifies a set of ***fixed weights*** $w_j = g^{-2}(\beta, \delta, \boldsymbol{x}_j)$, $j = 1, \ldots, n$, say. If one were to make the assumption that the distribution of $Y_j|\boldsymbol{x}_j$ is ***normal*** for each j, then the ***maximum likelihood estimator*** for β is the ***weighted least squares estimator*** , which solves

$$\sum_{j=1}^{n} f_{\beta}(\boldsymbol{x}_j, \beta) w_j\{Y_j - f(\boldsymbol{x}_j, \beta)\} = \mathbf{0}, \tag{7.18}$$

(Weighted) least squares estimation is often justified more generally, ***without*** the normality assumption, as minimizing an ***intuitively appealing objective function***, here, the ***weighted least squares*** criterion

$$\sum_{j=1}^{n} w_j\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2. \tag{7.19}$$

Of course, as the variance function ***depends*** on β and δ, which are unknown, the suggestion is effectively to replace the unknown weights w_j in (7.18) and (7.19) by ***estimated weights***, formed by substituting estimators for β and δ, as we demonstrate momentarily.

QUADRATIC ESTIMATING EQUATION FOR θ: Analogous to the approach to estimation of the covariance parameters ξ in the linear longitudinal data models we discussed in Chapters 5 and 6, an appealing estimating equation to be solved to obtain an estimator for $\theta = (\sigma^2, \delta^T)^T$ can be derived by differentiating the ***loglikelihood*** corresponding to taking the distribution of $Y_j|\boldsymbol{x}_j$ to be ***normal*** with mean and variance as in (7.1).

This loglikelihood is given by (ignoring constants)

$$-(n/2)\log\sigma^2 - (1/2)\sum_{j=1}^{n}\log g^2(\beta,\delta,\boldsymbol{x}_j) - (1/2)\sum_{j=1}^{n}\frac{\{Y_j - f(\boldsymbol{x}_j,\beta)\}^2}{\sigma^2 g^2(\beta,\delta,\boldsymbol{x}_j)}. \tag{7.20}$$

This ***does not mean*** that we necessarily ***believe*** normality; we simply use this approach to derive an estimating equation. Differentiating (7.20) yields the $(r \times 1)$ estimating equation (verify)

$$\sum_{j=1}^{n}\left[\frac{\{Y_j - f(\boldsymbol{x}_j,\beta)\}^2}{\sigma^2 g^2(\beta,\delta,\boldsymbol{x}_j)} - 1\right]\begin{pmatrix}1\\ \nu_\delta(\beta,\delta,\boldsymbol{x}_j)\end{pmatrix} = \mathbf{0}, \tag{7.21}$$

where

$$\nu_\delta(\beta,\delta,\boldsymbol{x}_j) = \partial/\partial\delta\ \log g(\beta,\delta,\boldsymbol{x}_j) = \frac{\partial/\partial\delta\, g(\beta,\delta,\boldsymbol{x}_j)}{g(\beta,\delta,\boldsymbol{x}_j)}.$$

The diligent student will be sure to make the ***analogy*** to equation (5.35) for estimation of covariance parameters ξ in the linear longitudinal data models in Chapters 5 and 6.

It is straightforward to observe (verify) that if the variance model $\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\beta,\delta,\boldsymbol{x}_j)$ in (7.1) is ***correctly specified*** , then (7.21) is an ***unbiased estimating equation***.

In the nonlinear modeling literature, this approach to estimation of θ, and thus δ in the "weights," in a mean-variance model (7.1) has been referred to as ***pseudolikelihood***. A ***REML*** version of (7.21) has also been proposed. Other estimating equations for θ based on ***alternatives*** to a quadratic functions of the ***deviations*** , $\{Y_j - f(\boldsymbol{x}_j,\beta)\}^2$, such as the ***absolute deviations*** $|Y_j - f(\boldsymbol{x}_j,\beta)|$, have also been proposed as a way to offer robustness to ***outliers*** ; see Carroll and Ruppert (1988, Chapter 3), Davidian and Carroll (1987), and Pinheiro and Bates (2000, Section 5.2)

GENERALIZED LEAST SQUARES: Of course, the estimating equation (7.21) must be solved ***jointly*** with the equation for β in (7.17); that is, we solve jointly in β and θ the estimating equations

$$\sum_{j=1}^{n} f_\beta(\boldsymbol{x}_j,\beta)g^{-2}(\beta,\delta,\boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j,\beta)\} = \mathbf{0}, \tag{7.22}$$

$$\sum_{j=1}^{n}\left[\frac{\{Y_j - f(\boldsymbol{x}_j,\beta)\}^2}{\sigma^2 g^2(\beta,\delta,\boldsymbol{x}_j)} - 1\right]\begin{pmatrix}1\\ \nu_\delta(\beta,\delta,\boldsymbol{x}_j)\end{pmatrix} = \mathbf{0}. \tag{7.23}$$

This can be implemented by an ***iterative algorithm***, starting from an initial estimate $\widehat{\beta}^{(0)}$, such as the ***nonlinear OLS estimator*** solving (7.16). At iteration ℓ,

1. Holding β fixed at $\widehat{\beta}^{(\ell)}$, solve the quadratic estimating equation (7.23) for θ to obtain $\widehat{\theta}^{(\ell)} = (\widehat{\sigma}^{2(\ell)}, \widehat{\delta}^{(\ell)T})^T$.
2. Holding δ fixed at $\widehat{\delta}^{(\ell)}$, solve the linear estimating equations (7.22) in β to obtain $\widehat{\beta}^{(\ell+1)}$. Set $\ell = \ell + 1$ and return to step 1.

A variation on step 2 is to substitute $\widehat{\beta}^{(\ell)}$ in $g^{-2}(\beta, \delta, \boldsymbol{x}_j)$ in (7.22) along with $\widehat{\delta}^{(\ell)}$, so that the "weights" are held fixed.

This procedure and variations on it is often referred to as (***estimated***) ***generalized least squares*** (GLS). One would ordinarily iterate between steps 1 and 2 to "***convergence***."

- It is important to recognize that, for arbitrary variance function $g^2(\beta, \delta, \boldsymbol{x})$, it is not necessarily the case that solving the system (7.22)-(7.23) corresponds to ***maximizing*** some ***objective function***. That is, in general, we view the resulting final estimators $(\widehat{\beta}^T, \widehat{\theta}^T)^T$ as ***M-estimators*** of the ***second type*** as in (4.2).
- Thus, there is no reason to expect that there is a ***unique solution*** to (7.22)-(7.23) or that the above algorithm should ***converge*** to a solution. ***Luckily***, in practice, it almost always does.
- Operationally, in this case it is not possible to obtain the solution $(\widehat{\beta}^T, \widehat{\theta}^T)^T$ directly by ***standard optimization techniques*** applied to an ***overall objective function*** as was the case for the longitudinal data methods in Chapters 5 and 6. Instead, an iterative algorithm like that above must be used.

  For fixed $\widehat{\beta}^{(\ell)}$, step 1 of the algorithm can in fact be carried out by maximizing the ***normal likelihood*** corresponding to general model (7.1) in θ. Then, for fixed $\widehat{\theta}^{(\ell)}$, step 2 can be carried out by so-called ***iteratively reweighted least squares*** (IRWLS), which is ***itself an iterative process*** that can be derived by taking a ***linear Taylor series*** of (7.22) in β about some β^*.

Defining $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$,

$$\boldsymbol{X}(\beta) = \begin{pmatrix} f_\beta^T(\boldsymbol{x}_1, \beta) \\ \vdots \\ f_\beta^T(\boldsymbol{x}_n, \beta) \end{pmatrix} \quad (n \times p), \qquad \boldsymbol{W}(\beta) = \text{diag}\{g^{-2}(\beta, \delta, \boldsymbol{x}_1), \ldots, g^{-2}(\beta, \delta, \boldsymbol{x}_n\}$$

for ***fixed*** δ, the ath iteration of IRWLS is

$$\beta_{(a+1)} = \beta_{(a)} + \{\boldsymbol{X}_{(a)}^T \boldsymbol{W}_{(a)} \boldsymbol{X}_{(a)}\}^{-1} \boldsymbol{X}_{(a)}^T \boldsymbol{W}_{(a)} (\boldsymbol{Y} - \boldsymbol{f}_{(a)}), \quad \boldsymbol{W}_{(a)} = \boldsymbol{W}(\beta_{(a)}, \quad \boldsymbol{X}_{(a)} = \boldsymbol{X}(\beta_{(a)}). \quad (7.24)$$

Iteration continues until some convergence criterion is met.

The ***diligent student*** will look up or verify him/herself the derivation of (7.24).

- When the mean-variance model is of the form for a generalized (non)linear model, so that there is ***no unknown*** δ in the variance function, the estimating equation (7.22) for β is in fact the ***score equation*** (7.15), and its solution corresponds to maximizing the loglikelihood, which is carried out by an IRWLS approach. Thus, IRWLS is the standard way to implement ***maximum likelihood*** in the class of generalized (non)linear models.

For future reference, we can write the system of estimating equations (7.22)-(7.23) ***compactly*** in obvious streamlined notation as (check)

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & \boldsymbol{0} \\ \boldsymbol{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\delta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & 2\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \boldsymbol{0}. \quad (7.25)$$

QUADRATIC ESTIMATING EQUATION FOR β***:*** There is a common ***misconception*** that solving (7.22)-(7.23) corresponds to ***maximizing*** the ***normal loglikelihood*** in (7.20). Of course, (7.23) does arise from differentiating (7.20) with respect to $\boldsymbol{\theta}$.

However, it is straightforward to derive (do it) that differentiating (7.20) with respect to β yields the alternative estimating equation

$$\sum_{j=1}^{n} f_\beta(\boldsymbol{x}_j, \beta) g^{-2}(\beta, \delta, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta)\} + \sigma^2 \sum_{j=1}^{n} \left[\frac{\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2}{\sigma^2 g^2(\beta, \delta, \boldsymbol{x}_j)} - 1 \right] \nu_\beta(\beta, \delta, \boldsymbol{x}_j) = \boldsymbol{0}, \quad (7.26)$$

where

$$\nu_\beta(\beta, \delta, \boldsymbol{x}_j) = \partial/\partial\beta \log g(\beta, \delta, \boldsymbol{x}_j) = \frac{\partial/\partial\beta\, g(\beta, \delta, \boldsymbol{x}_j)}{g(\beta, \delta, \boldsymbol{x}_j)}.$$

The second term in (7.26) is a result of the fact that the variance function $g^2(\beta, \delta, \boldsymbol{x}_j)$ ***depends on*** β.

Note that the first term in the estimating equation (7.26) is identical to the ***linear estimating equations*** (7.22). The second term thus demonstrates that, when the variance is believed to depend on β (usually through the ***mean response***), there is ***additional information*** about β in the ***squared deviations*** $\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2$ above and beyond that in the mean itself.

- This is a consequence of the fact that the normal distribution places ***no restrictions*** on the form of the mean and variance. Intuitively, then, when the variance depends on the parameter β that describes the mean, it stands to reason that ***more*** can be learned about it from the quadratic function $\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2$, which obviously reflects the nature of ***variance***.
- This suggests that, under the assumption of normality, it is possible to obtain an estimator for β that is ***more efficient*** than that obtained from the linear GLS equation.
- Of course, if the variance function ***does not depend*** on β, then (7.26) reduces to the linear equation (7.22), in which case the ***maximum likelihood estimators under normality*** for β and θ do jointly solve (7.22)-(7.23).
- In contrast, the ***scaled exponential family*** distributions with density (7.12) are such that the variance is a ***specific function of the mean*** dictated by the particular distribution. Intuitively, this suggests that, under these distributions, there is ***no additional information*** to be gained about β from the variance, reflected in the fact that the resulting estimating equation (7.15) does ***not*** involve a quadratic function of the deviations.

REMARK: A critical feature of the estimating equation (7.26) is that it is ***not enough*** for $f(\boldsymbol{x}_j, \beta)$ to be ***correctly specified*** for this to be an ***unbiased estimating equation***.

- With $f(\boldsymbol{x}_j, \beta)$ correctly specified, (7.26) is an unbiased estimating equation if the ***variance model*** $\sigma^2 g^2(\beta, \delta, \boldsymbol{x})$ is ***also correctly specified***. Thus, in general, for (7.26) to yield a ***consistent estimator*** for the true value β_0, it is necessary to specify ***both*** the mean and variance models correctly.
- Thus, there is a ***trade-off*** between ***gaining information*** about β to obtain a ***more efficient*** estimator and ending up with an ***inconsistent estimator*** for β due to ***misspecification*** of the variance model.
- Intuitively, as it is ***more difficult*** to model ***variances*** than it is to model means, this is a ***non-trivial*** concern.

In summary, under the assumption that the distribution of $Y_j|\boldsymbol{x}_j$ is ***normal*** with first two moments as in (7.1), the maximum likehood estimators for β and θ jointly solve (7.26) and (7.23). For future reference, we write this system of estimating equations ***compactly*** in streamlined form as (verify)

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma g_j^2 \nu_{\beta j} \\ \mathbf{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\delta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & 2\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}. \tag{7.27}$$

Of course, this system of estimating equations differs from the GLS equations in (7.25) only by the presence of the non-zero off-diagonal entry in the leftmost matrix, which serves to introduce the ***quadratic dependence*** of the equation for β and which equals zero when $g^2(\beta, \delta, \boldsymbol{x}_j)$ does not depend on β.

7.4 Large sample results

It is possible via large sample theory arguments to derive approximate sampling distributions for the estimators for β obtained by solving the ***linear estimating equation*** (7.22) jointly with (7.23), i.e., (7.25) ; or the ***quadratic estimating equation*** (7.26) jointly with (7.23), (i.e., (7.27). Here, "***large sample***" implies $n \to \infty$.

The calculations are ***simpler versions*** of those required to deduce the large sample (large m) properties of the estimators for ***general PA longitudinal data models for mean and covariance matrix*** we discuss in Chapter 8). We thus provide a brief sketch of these results for (7.25) and (7.27), whose implications carry over to the longitudinal setting.

LINEAR ESTIMATING EQUATION: Analogous to the situation of a ***possibly incorrectly specified covariance model*** in the case of the linear PA models in Section 5.5, we can carry out a similar M-estimation argument under a misspecified variance model.

Assume that we posit a ***correct*** mean model $E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \beta)$, and suppose that the ***true*** variance ***actually generating the data*** is given by

$$\text{var}(Y_j|\boldsymbol{x}_j) = v_{0j}. \tag{7.28}$$

Suppose, however, that we posit a variance model

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_j) = v(\beta, \theta, \boldsymbol{x}_j)$$

such that ***there is not necessarily*** a $\theta_0 = (\sigma_0^2, \delta_0^T)^T$ such that $v(\beta_0, \theta_0, \boldsymbol{x}_j) = v_{0j}$.

Suppose further that we estimate $\boldsymbol{\theta}$ by solving the estimating equation (7.23) jointly with (7.22). The equation (7.23) is ***not an unbiased estimating equation*** if the variance model is ***incorrect*** ; however, assume that, under this incorrect variance model, the resulting "estimator" $\widehat{\boldsymbol{\theta}}_* = (\widehat{\sigma}^2, \widehat{\boldsymbol{\delta}}^T)^T \xrightarrow{p} (\sigma^{2*}, \boldsymbol{\delta}^{*T})^T = \boldsymbol{\theta}^*$ for some $\boldsymbol{\theta}^*$. Note that for the linear estimating equation (7.22), we then have

$$E[f_\beta(\boldsymbol{x}_j, \beta_0) g^{-2}(\beta_0, \boldsymbol{\delta}^*, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta_0)\} \mid \boldsymbol{x}_j] = \mathbf{0}, \quad j = 1, \ldots, m,$$

so that (7.22) is still an ***unbiased estimating equation***, and thus $\widehat{\beta}$ is a ***consistent estimator*** for β_0 nonetheless.

Define $v_j^* = \sigma^{2*} g^{-2}(\beta_0, \boldsymbol{\delta}^*, \boldsymbol{x}_j)$, $j = 1, \ldots, n$, and let $f_{\beta\beta}(\boldsymbol{x}, \beta) = \partial^2 / \partial\beta \partial\beta^T f(\boldsymbol{x}, \beta)$, the $(p \times p)$ matrix of second partial derivatives of $f(\boldsymbol{x}, \beta)$. Let

$$\boldsymbol{V}_0 = \text{diag}(v_{01}, \ldots, v_{0n}), \quad \boldsymbol{V}^* = \text{diag}(v_1^*, \ldots, v_n^*).$$

Note that we use $\boldsymbol{V}^*$ here ***differently*** from its definition in Chapters 5 and 6. Assume also that $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = O_p(1)$ (bounded in probability).

Expanding the right hand side of

$$\mathbf{0} = \sigma^{-2*} n^{-1/2} \sum_{j=1}^{n} f_\beta(\boldsymbol{x}_j, \widehat{\beta}) g^{-2}(\widehat{\beta}, \widehat{\boldsymbol{\delta}}, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \widehat{\beta})\}$$

in a Taylor series about $(\widehat{\beta}^T, \widehat{\boldsymbol{\delta}}^T)^T = (\beta_0^T, \boldsymbol{\delta}^{*T})^T$, analogous to (5.73), we obtain

$$\mathbf{0} \approx \boldsymbol{C}_n^* + (\boldsymbol{A}_{n1}^* + \boldsymbol{A}_{n2}^* + \boldsymbol{A}_{n3}^*) n^{1/2}(\widehat{\beta} - \beta_0) + \boldsymbol{E}_n^* n^{1/2}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*), \tag{7.29}$$

where (check)

$$\boldsymbol{A}_{n1}^* = n^{-1} \sum_{j=1}^{n} v_j^{*-1} f_{\beta\beta}(\boldsymbol{x}, \beta_0)\{Y_j - f(\boldsymbol{x}_j \beta_0)\} \xrightarrow{p} \mathbf{0},$$

$$\boldsymbol{A}_{n2}^* = -n^{-1} \sum_{j=1}^{n} v_j^{*-1} f_\beta(\boldsymbol{x}_j, \beta_0) f_\beta^T(\boldsymbol{x}_j, \beta_0) \xrightarrow{p} \boldsymbol{A}^* = lim_{n\to\infty} n^{-1} \boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{X}, \qquad \boldsymbol{X} = \boldsymbol{X}(\beta_0)$$

$$\boldsymbol{A}_{n3}^* = -2n^{-1} \sum_{j=1}^{n} v_j^{*-1} f_\beta(\boldsymbol{x}_j, \beta_0) \nu_\beta^T(\beta_0, \boldsymbol{\delta}^*, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta_0)\} \xrightarrow{p} \mathbf{0},$$

$$\boldsymbol{E}_n^* = -2n^{-1} \sum_{j=1}^{n} v_j^{*-1} f_\beta(\boldsymbol{x}_j, \beta_0) \nu_\delta^T(\beta_0, \boldsymbol{\delta}^*, \boldsymbol{x}_j)\{Y_j - f(\boldsymbol{x}_j, \beta_0)\} \xrightarrow{p} \mathbf{0},$$

$$\boldsymbol{C}_n^* = n^{-1/2} \sum_{j=1}^{n} v_j^{*-1} f_\beta(\boldsymbol{x}_j, \beta_0)\{Y_j - f(\boldsymbol{x}_j, \beta_0)\} \xrightarrow{\mathcal{L}} \mathcal{N}\{\mathbf{0}, \boldsymbol{B}^*\}, \qquad \boldsymbol{B}^* = \lim_{n\to\infty} n^{-1} \boldsymbol{X}^T \boldsymbol{V}^{*-1} \boldsymbol{V}_0 \boldsymbol{V}^{*-1} \boldsymbol{X}.$$

It follows that

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{*-1}\boldsymbol{B}^{*}\boldsymbol{A}^{*-1}). \tag{7.30}$$

Moreover, if in fact the variance model is ***correctly specified*** after all, so that $\delta^* = \delta_0$ for which $v_{0j} = g^2(\beta_0, \delta_0, \boldsymbol{x}_j)$, then $v_j^* = v_{0j}$, $j = 1, \ldots, n$, and (7.30) reduces to

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{-1}), \quad \boldsymbol{A} = \lim_{n\to\infty} n^{-1}\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X}. \tag{7.31}$$

- The results in (7.30) and (7.31) are of course entirely ***analogous*** to those we obtained for the ***PA linear model*** in Section 5.5, with the ***exception*** that the matrix $\boldsymbol{X} = \boldsymbol{X}(\beta_0)$ here is a ***nonlinear function*** of the true value β_0 and the covariates rather than a ***fixed design matrix***.
- In the case of a ***generalized (non)linear model*** , so that there is ***no unknown parameter*** δ, $\widehat{\beta}$ is in fact the ***MLE*** and thus (7.31) is the large sample result for maximum likelihood under a scaled exponential family distribution
- These results are used to specify ***approximate sampling distributions*** in the usual way; e.g., under the assumption the ***variance model is correctly specified*** , one would derive ***model-based standard errors*** by substituting the estimates into $\boldsymbol{X}$ and $\boldsymbol{V}_0$ to obtain, in obvious notation,

$$\widehat{\beta} \dot{\sim} \mathcal{N}\{\beta_0, \boldsymbol{X}^T(\widehat{\beta})\boldsymbol{V}^{-1}(\widehat{\beta}, \widehat{\theta})\boldsymbol{X}(\widehat{\beta})\}, \quad \boldsymbol{V}(\beta, \theta) = \sigma^2\text{diag}\{g^{-2}(\beta, \delta, \boldsymbol{x}_j), \ldots, g^{-2}(\beta, \delta, \boldsymbol{x}_j)\}. \tag{7.32}$$

- Likewise, ***robust*** or ***empirical standard errors*** can be derived from (7.30).

In the next chapter, we will see that analogous results hold for a ***general nonlinear population-averaged mean-covariance model***.

QUADRATIC ESTIMATING EQUATION: It is likewise possible to derive the large sample distribution of the estimator for β solving the system in (7.27) jointly in θ; that is, solving the ***quadratic estimating equation*** (7.26). Because this equation is ***not unbiased*** unless the variance function is ***correctly specified*** , the argument proceeds under the assumption that the variance function model is ***correct***. We thus assume that there are ***true values*** β_0 and θ_0 such that the posited mean and variance models yield the true mean and variance relationships.

The resulting approximate sampling distribution can be compared to that we just derived for the estimator for β solving (7.25) to gain insight into the ***potential gains in efficiency*** for estimating β achieved when the variance model is indeed correctly specified by using the ***quadratic*** rather than the ***linear equation*** under different conditions.

The argument entails expanding $n^{-1/2}\times$ (7.26) in $(\widehat{\beta}^T, \widehat{\theta}^T)^T$ about $(\beta_o^T, \theta_0^T)^T$ to find an approximate expression for

$$n^{1/2}\begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\theta} - \theta_0 \end{pmatrix}$$

and then ***isolating the implied distribution*** of $n^{1/2}(\widehat{\beta} - \beta_0)$ by appealing to formulæ for the ***inverse of a partitoned matrix*** (see Appendix A).

It is ***not possible*** to expand the estimating equation (7.26) alone to arrive at this ***directly*** as we did for the ***linear estimating equation*** because it turns out that the dependence of the distribution of $\widehat{\beta}$ on that of $\widehat{\theta}$ ***does not vanish*** as it does for (7.22) above.

The argument is thus ***tedious***; accordingly we do not give it here but only present the result. The argument assumes that, although the equations (7.25) are derived under the assumption of ***normality***, the ***true distribution*** of $\boldsymbol{Y}_j|\boldsymbol{x}_j$ is ***not necessarily normal***.

HIGHER MOMENT PROPERTIES: Letting

$$\epsilon_j = \frac{Y_j - f(\boldsymbol{x}_j, \beta_0)}{\sigma_0 g(\beta_0, \delta_0, \boldsymbol{x}_j)},$$

$E(\epsilon_j^3|\boldsymbol{x}_j) = \zeta$ is the ***coefficient of skewness*** of the distribution of $Y_j|\boldsymbol{x}_j$ (***third*** moment property) and, with $\text{var}(\epsilon_j^2|\boldsymbol{x}_j) = 2 + \kappa$, κ is the ***coefficient of excess kurtosis*** (***fourth*** moment property). For the ***normal distribution***, $\zeta = \kappa = 0$.

Define $\tau_\theta(\beta, \delta, \boldsymbol{x}_j) = \{1, \nu_\delta^T(\beta, \delta, \boldsymbol{x}_j)\}^T$. Using streamlined notation where a "0" subscript indicates evaluation at the ***true values*** of the parameters, let

$$\boldsymbol{R} = \begin{pmatrix} \nu_{\beta 01}^T \\ \vdots \\ \nu_{\beta 0n}^T \end{pmatrix} \quad (n \times p), \quad \boldsymbol{Q} = \begin{pmatrix} \tau_{\delta 01}^T \\ \vdots \\ \tau_{\delta 0n}^T \end{pmatrix} \quad (n \times r), \quad \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T.$$

Then it can be shown that, if the ***skewness and excess kurtosis*** of the ***true distribution*** of $Y_j|\boldsymbol{x}_j$ are ζ and κ,

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda}^{-1}\boldsymbol{\Delta}\boldsymbol{\Lambda}^{-1}), \tag{7.33}$$

$$\boldsymbol{\Lambda} = \lim_{n\to\infty} n^{-1}(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R}),$$

$$\boldsymbol{\Delta} = \lim_{n\to\infty} n^{-1}\left\{\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + (2+\kappa)\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R} + \zeta(\boldsymbol{X}^T\boldsymbol{V}_0^{-1/2}\boldsymbol{P}\boldsymbol{R} + \boldsymbol{R}^T\boldsymbol{P}\boldsymbol{V}_0^{-1/2}\boldsymbol{X})\right\}.$$

- The dependence of $\boldsymbol{\Delta}$ on third and fourth moment properties of the true distribution of $Y_j|\boldsymbol{x}_j$ is a consequence of the fact that the summand of the estimating equation (7.26) involves both ***linear and quadratic terms*** in $\{Y_j - f(\boldsymbol{x}_j, \beta)\}$, so that ζ and κ show up in the ***variance*** of the summand when the ***central limit theorem*** is applied.

- Both components of the covariance matrix in (7.33) depend on the covariance matrix of the linear estimator, $(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}$ ***plus*** additional terms arise because of the ***quadratic component*** of the estimating equation (7.26) for β ($\boldsymbol{R}$) and the need to estimate θ ($\boldsymbol{Q}$). Thus, inclusion of the quadratic term in the estimating equation for β has the effect of making the properties of $\widehat{\beta}$ depend on those of $\widehat{\theta}$.

- When $\zeta = 0$ and $\kappa = 0$, corresponding to the ***third and fourth moments of the normal distribution***, so that the true distribution of $Y_j|\boldsymbol{x}_j$ is ***really normal*** ,

$$\boldsymbol{\Delta} = \boldsymbol{\Lambda}.$$

  Then (7.33) implies approximately that

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, n^{-1}\boldsymbol{\Lambda}^{-1}), \quad n^{-1}\boldsymbol{\Lambda}^{-1} \approx (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1}, \tag{7.34}$$

  whereas, for the ***linear estimating equation*** when the variance function is ***correctly specified*** as we assume here, (7.31) implies approximately that

$$\widehat{\beta} \dot{\sim} \mathcal{N}(\beta_0, n^{-1}\boldsymbol{A}^{-1}), \quad n^{-1}\boldsymbol{A}^{-1} \approx (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}, \tag{7.35}$$

  It is straightforward to observe that the difference

$$(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1} - (\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1}$$

  is ***nonnegative definite*** (check); thus, (7.34) and (7.35) imply that, when the true distribution really is normal, the quadratic estimator for β is ***more efficient*** than the linear estimator.

- However, if the true distribution is ***not normal*** and instead has ***arbitrary*** coefficients of skewness and kurtosis ζ and κ, ***relative efficiency*** of the two estimators is less clear. Approximately for large n, analogous to (7.34) and (7.35), this involves comparing $n^{-1}\boldsymbol{A}^{-1}$ in (7.35) to

$$\begin{aligned}(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1} &\left\{\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + (2+\kappa)\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R} + \zeta(\boldsymbol{X}^T\boldsymbol{V}_0^{-1/2}\boldsymbol{P}\boldsymbol{R} + \boldsymbol{R}^T\boldsymbol{P}\boldsymbol{V}_0^{-1/2}\boldsymbol{X})\right\} \\ &\times(\boldsymbol{X}^T\boldsymbol{V}_0^{-1}\boldsymbol{X} + 2\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1}.\end{aligned}$$

  Evidently, whether or not the difference of these two covariance matrices is nonnegative definite depends in a complicated way on ζ, κ, and the matrices $\boldsymbol{R}$ and $\boldsymbol{Q}$.

The takeaway message is that, although estimation of β via the ***quadratic estimating equation*** (7.26), jointly with that of θ via (7.23), will be ***more efficient*** than using the ***linear equation*** (7.22), if $Y_j|\boldsymbol{x}_j$ is ***exactly normal*** , if it is ***not*** , it is not clear that the extra trouble is worthwhile.

Indeed, use of the quadratic equation ***requires*** that the ***variance model*** is ***correctly specified*** to achieve ***consistent estimation*** of β, so that the potential efficiency gain must be weighed against the possibility of ***misspecification*** of this model.

LARGE SAMPLE THEORY FOR VARIANCE PARAMETER ESTIMATORS: It is also possible to derive an approximate sampling distribution for the estimator for the ***variance parameter*** θ in either case. We do not pursue this here.

- From the results for the quadratic estimator for β above, because the estimating equation (7.23) depends on $\{Y_j - f(\boldsymbol{x}_j, \beta)\}^2$, we expect that properties of $\widehat{\theta}$ are sensitive to whether or not the true distribution of $Y_j|\boldsymbol{x}_j$ is ***really normal*** and thus depend on the ***coefficients of skewness and excess kurtosis*** of the true distribution.
- This reflects a ***more general*** phenomenon. The properties of estimators of ***second moment*** properties like ***variance and covariance*** depend on the ***third and fourth moment*** properties of the true distribution of the data. Thus, obtaining ***realistic assessments of uncertainty*** of such estimators is ***inherently challenging***. In particular, unless the true distribution is really ***exactly normal*** , assessments based on the assumption of normality will be ***unreliable***.

GENERALIZATION: All of these results ***generalize*** to the longitudinal data setting. We discuss some of these in Chapter 8.

CURIOSITY: We end this chapter by noting an interesting feature of the linear estimating equations (7.17) for β, namely, in shorthand,

$$\sum_{j=1}^{n} f_{\beta j}\sigma^{-2}g_j^{-2}(Y_j - f_j) = \mathbf{0}, \tag{7.36}$$

and the system of joint estimating equations (7.27) for β and θ,

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma g_j^2 \nu_{\beta j} \\ \mathbf{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\delta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & 2\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}. \tag{7.37}$$

It is straightforward to see or show (verify) that (7.36) and (7.37) are of the general form

$$\sum_{j=1}^{n} \mathcal{D}_j^T(\eta)\mathcal{V}_j^{-1}(\eta)\{\boldsymbol{s}_j(\eta) - \boldsymbol{m}_j(\eta)\} = \boldsymbol{0}, \tag{7.38}$$

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

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

- The ***linear estimating equation*** for β in (7.36) with θ treated as ***fixed*** is trivially of this form, with $\eta = \beta$, $v = 1$, and

$$\boldsymbol{s}_j(\eta) = Y_j, \qquad \boldsymbol{m}_j(\eta) = f(\boldsymbol{x}_j, \beta), \qquad \mathcal{V}_j(\eta) = \sigma^2 g^2(\beta, \delta, \boldsymbol{x}_j), \qquad \mathcal{D}_j^T(\eta) = f_\beta(\boldsymbol{x}_j, \beta).$$

- The joint ***quadratic estimating equations*** in (7.37) are also of this form, with $\eta = (\beta^T, \theta^T)^T$, $v = 2$, and, in shorthand,

$$\boldsymbol{s}_j(\eta) = \begin{pmatrix} Y_j \\ (Y_j - f_j)^2 \end{pmatrix}, \qquad \boldsymbol{m}_j(\eta) = \begin{pmatrix} f_j \\ \sigma^2 g_j^2 \end{pmatrix}, \qquad \mathcal{V}_j(\eta) = \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & 2\sigma^4 g_j^4 \end{pmatrix}, \tag{7.39}$$

$$\mathcal{D}_j^T(\eta) = \begin{pmatrix} f_{\beta j} & 2\sigma g_j^2 \nu_{\beta j} \\ \boldsymbol{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\delta j} \end{pmatrix} \end{pmatrix}.$$

  Note that $\mathcal{V}_j(\eta)$ in (7.39) is $\text{var}(\boldsymbol{s}_j|\boldsymbol{x}_j)$ ***under the assumption of normality*** , so that $\text{cov}\{Y_j, (Y_j - f_j)^2|\boldsymbol{x}_j\} = 0$ and $\text{var}\{(Y_j - f_j)^2|\boldsymbol{x}_j\} = 2\sigma^4 g_j^4$, which of course correspond to the normal, which has coefficients of skewness and excess kurtosis $\zeta = \kappa = 0$.

- This suggests that, if we instead believe that the true distribution of $Y_j|\boldsymbol{x}_j$ has skewness and kurtosis $\zeta \neq 0$, $\kappa > 0$ for some ζ and κ, the "***covariance matrix***" $\mathcal{V}_j(\eta)$ in (7.39) is ***incorrectly specified***.

- To gain insight into the consequences of this, we can make an ***analogy*** to the argument we made in Chapter 5 comparing the covariance matrices (5.75) and (5.76) that resulted from using ***correct and incorrect specifications*** for the overall covariance matrix of a response vector in the ***linear estimating equation*** for β in the linear PA models of that chapter. This argument showed that using an incorrect model for the covariance matrix $\boldsymbol{V}_i$ leads to an estimator for β that is ***inefficient*** relative to that obtained using a ***correct model*** , which corresponds to using the ***optimal linear estimating equation***.

It is straightforward to see (verify) that, if we identify $\mathcal{D}_j^T$ with $\boldsymbol{X}_i^T$, $\mathcal{V}_j$ with $\boldsymbol{V}_i$, $\boldsymbol{s}_j$ with $\boldsymbol{Y}_i$, and $\boldsymbol{m}_j$ with $\boldsymbol{X}_i\beta$ in the estimating equation (5.59), the equation (7.38), namely,

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

is of the ***same form*** and can be viewed as a ***linear estimating equation*** in the "***response***" $\boldsymbol{s}_j$. Thus, the same (large sample) argument regarding inefficiency applies here with these correspondences, and thus suggests that using $\mathcal{V}_j(\eta)$ in (7.39) should result in ***inefficiency*** of the resulting estimators for β and θ ***relative to*** instead taking

$$\mathcal{V}_j(\eta) = \begin{pmatrix} \sigma^2 g_j^2 & \zeta\sigma^3 g_j^3 \\ \zeta\sigma^3 g_j^3 & (2+\kappa)\sigma^4 g_j^4 \end{pmatrix},$$

which is the "***correct covariance matrix***" and should thus result in the "***optimal linear estimating equation***" of the form (7.38).

- Of course, it is ***extremely unlikely*** we would ever know the true ζ and κ in practice. However, this shows that, by assuming normality, we are ***effectively*** making the assumption that the ***first four moments*** of the distribution of $Y_j|\boldsymbol{x}_j$ are the ***same*** as those of the normal distribution with mean and variance given by the posited mean-variance model (7.1).

These considerations will arise in a ***multivariate*** context in the overview of ***generalized estimating equations*** in the next chapter.