10 Quadratic vs. linear estimating equations for $\boldsymbol{\beta}$

10.1 Introduction

For the general mean-variance model

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \boldsymbol{\beta}), \quad \text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j), \tag{10.1}$$

we have identified two main approaches to estimation of $\boldsymbol{\beta}$ (possibly jointly with σ and $\boldsymbol{\theta}$):

- *Linear* estimating equations, with the optimal choice being the GLS equation. These equations require only that the analyst be willing to specify the first two moments of $Y_j|\boldsymbol{x}_j$, as in (10.1).
- *Quadratic* estimating equations, which for g depending on $\boldsymbol{\beta}$, involve not only the moments in (10.1), but also require a specification for the third and fourth moments of $Y_j|\boldsymbol{x}_j$. An example of such a quadratic equation is that arising from consideration of the normal distribution, which corresponds to making the particular assumption about the third and fourth moments of this distribution (skewness=0, excess kurtosis=0).

An obvious question, posed at the end of Chapter 9, is whether or not there are trade-offs between the two types of equations.

- If, *in truth*, the distribution of $Y_j|\boldsymbol{x}_j$ is exactly normal, and (10.1) holds, then standard likelihood theory tells us that (in a large sample sense) estimation of $\boldsymbol{\beta}$ via the normal theory ML equation, which is quadratic when g depends on $\boldsymbol{\beta}$, is "optimal." Under these circumstances, as the GLS equation is linear in the data and thus a different estimating equation, it must be inefficient, and it seems that the extra trouble to implement the quadratic normal ML approach is worthwhile (although how worthwhile would depend on the specific problem). We will demonstrate this formally in Section 10.2.
- However, how do these two approaches compare when the data are *not* really normal? If the data are not really normal, then the third and fourth moment assumptions implicit in the normal ML approach are not correct.
- As noted above, writing $\epsilon_j = \{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta})\}/\{\sigma g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j)\}$, the quadratic equations require the analyst to specify $E(\epsilon_j^3|\boldsymbol{x}_j) = \zeta_j$ and $\text{var}(\epsilon_j^2|\boldsymbol{x}_j) = 2 + \kappa_j$, say.

  Here, ζ_j and κ_j could conceivably be functions of $\boldsymbol{x}_j$ and the parameters or fixed constants (e.g. the same for all j), chosen by the data analyst.

Intuition would suggest that, even if the data are not normal, were we able to specify these third and fourth moments correctly, we would stand to gain efficiency over the linear GLS approach, which only uses information on the first two moments.

- However, again, if we are wrong about ζ_j and κ_j, how do the approaches compare? Certainly, it would be quite difficult to specify these moments in practice based on a finite sample. Making the assumption of normality is of course tantamount to assuming that $\zeta_j \equiv 0$ and $\kappa_j \equiv 0$ for all j. If we are not willing to assume normality, these moments are likely something else. If we really felt capable of understanding perfectly the form of ζ_j and κ_j for all j, chances are we would be willing to specify an entire distribution for $Y_j|\boldsymbol{x}_j$, in which case we might consider alternatively maximum likelihood under that distribution (which could very well lead to something other than a linear or quadratic estimating equation).

- Even if we were willing to believe that the ϵ_j are such that $\zeta_j \equiv \zeta$ and $\kappa_j \equiv \kappa$ for all j (e.g. assuming ϵ_j are i.i.d.), but that ζ and κ are not both equal to zero (normality), it would still be challenging to specify them based just on observed data. Are we "safer" not even trying?

- Another concern is that of misspecification of the variance model. We have already seen that failure to get the variance function correct can result in loss of efficiency for GLS, but consistency is preserved. What are the implications for quadratic equations? These equations seem more highly dependent on the form of g, so intuition would suggest that the effects of misspecification could be more deadly.

In this chapter, we will investigate these issues via large sample arguments. First, we will derive the asymptotic behavior of estimators solving general quadratic estimating equations for $\boldsymbol{\beta}$. This will allow us to make direct comparisons in terms of asymptotic relative efficiency not only with GLS but within the class of quadratic estimating equations in some special cases. We will also consider the effect of misspecification of the variance function on estimation via quadratic equations and verify that the effects can indeed be more serious than for GLS.

We will focus in particular on the special case of the normal theory ML quadratic equation. There is some rationale for devoting some specific effort to this special case.

- The normal theory ML equation is the simplest case of a quadratic estimating equation. Because it is so hard to specify third and fourth moments in practice with confidence, it is common practice when using quadratic equations to use those dictated by normality, even if it is not believed the data are really normal.

This is actually more of an issue in the case of multivariate response, and we will discuss it in Chapter 14. Two references in this context where adopting normal higher moments is advocated as a practical approach are Prentice and Zhao (1991) and Liang, Zeger, and Qaqish (1992).

- As we have noted previously, in the univariate case, normal theory ML is considered the method of choice over GLS; this is true in the pharmacokinetics literature, for example, where the ML method is known by the alternative name *extended least squares* (ELS).
- Some of the issues that we will discuss, such as the effect of misspecification of the variance function in different ways, are relevant to general quadratic equations but are easier to see in the specific case of the normal ML equations.

10.2 Limiting distribution of $\hat{\boldsymbol{\beta}}$ based on quadratic estimating equations

We will begin by appealing to the usual M-estimator argument to derive the large sample normal distribution of the estimator for $\boldsymbol{\beta}$ solving a general quadratic estimating equation.

As we did in Chapter 9, we will allow for the possibility that $\boldsymbol{\theta}$ is unknown and estimated. Recall from Chapter 5 that, with a quadratic estimating equation for $\boldsymbol{\beta}$, it is necessary to consider *joint* estimation of $\boldsymbol{\beta}$, σ, and $\boldsymbol{\theta}$; in fact, even if $\boldsymbol{\theta}$ is known, we still must consider estimation of $\boldsymbol{\beta}$ and σ jointly, in contrast to the linear GLS case. Of course, results for the case of $\boldsymbol{\theta}$ known will follow from simplification of those for when it is estimated.

As we discussed in Chapter 6, there is an entire class of possible estimating equations for $\boldsymbol{\theta}$ based on different transformations of absolute residuals. Here, we will restrict attention to what would likely be done in practice: If the analyst chooses to solve a quadratic equation for $\boldsymbol{\beta}$, it is almost always the case that s/he would also choose to solve a quadratic equation for $(\sigma, \boldsymbol{\theta}^T)^T$ jointly. Estimation of variance parameters is almost always carried out using quadratic equations in practice, usually via the PL method. Recall from Chapter 6 that this method is predicated on the assumption that $\text{var}(\epsilon_j^2|\boldsymbol{x}_j)$ is constant for all j (see page 124). In fact, as we have noted previously in passing, when the PL equation is coupled with an equation for $\boldsymbol{\beta}$, we have always used implicitly the assumption that $\zeta_j \equiv 0$ for all j. Here, we will consider a more general formulation.

Throughout this section, we will assume that the variance model in (10.1) is correctly specified, and we will write generically $\epsilon_j = \{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta})\}/\{\sigma g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j)$ when discussing assumptions on higher moments. As before, we denote the true values of the model parameters as $\boldsymbol{\beta}_0$, σ_0, and $\boldsymbol{\theta}_0$. We will also use ϵ_j later to represent the "true" standardized errors at these true values, as in Chapters 8 and 9.

From our discussion in Chapters 5 and 6, it is clear that the most general form of quadratic estimating equation of the type we have discussed involving a "gradient matrix" and "covariance matrix" may be written as follows, using the shorthand notation used previously:

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma^2 g_j^2 \nu_{\beta j} \\ \mathbf{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\theta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & \zeta_j \sigma^3 g_j^3 \\ \zeta_j \sigma^3 g_j^3 & (2+\kappa_j)\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}. \qquad (10.2)$$

The equation is thus constructed under the assumption that $E(\epsilon_j^3|\boldsymbol{x}_j) = \zeta_j$ and $\text{var}(\epsilon_j^2|\boldsymbol{x}_j) = 2+\kappa_j$; this equation of course follows by noting that

$$\text{cov}\{Y_j, (Y_j - f_j)^2 \,|\, \boldsymbol{x}_j\} = \sigma^3 g_j^3 E(\epsilon_j^3|\boldsymbol{x}_j).$$

In the following argument, we assume that ζ_j and κ_j are either fixed constants or functions of $\boldsymbol{x}_j$ and possibly the parameters. These may not necessarily be correctly specified (i.e., equal to or of the same form as the true skewness and excess kurtosis for each j).

In principle, this set of equations is just a special case of the general M-estimating equations discussed in Section 8.2. We may write these succinctly as

$$\sum_{j=1}^{n} \boldsymbol{\Psi}_j(Y_j, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}) = \mathbf{0}; \qquad (10.3)$$

for simplicity, we suppress explicit dependence on $\boldsymbol{x}_j$, as our arguments are conditional on the $\boldsymbol{x}_j$.

It is straightforward to observe that the general equation (10.2) is an unbiased estimating equation, as clearly $E\{\boldsymbol{\Psi}_j(Y_j, \boldsymbol{\beta}_0, \sigma_0, \boldsymbol{\theta}_0)|\boldsymbol{x}_j\} = \mathbf{0}$. Thus, we expect that the joint estimators, to be denoted here as $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$, and $\hat{\boldsymbol{\theta}}$, are consistent for the true values $\boldsymbol{\beta}_0$, σ_0, $\boldsymbol{\theta}_0$.

Assuming consistency, we may proceed with the usual argument to derive the asymptotic distribution. Writing for brevity $\boldsymbol{\alpha} = (\boldsymbol{\beta}^T, \sigma, \boldsymbol{\theta}^T)^T$, $\boldsymbol{\alpha}_0 = (\boldsymbol{\beta}_0^T, \sigma_0, \boldsymbol{\theta}_0^T)^T$, and thus $\boldsymbol{\Psi}_j(Y_j, \boldsymbol{\alpha})$, from Section 8.2 we have

$$n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \approx -\boldsymbol{A}_n^{-1} \boldsymbol{C}_n,$$

where

$$\boldsymbol{A}_n = n^{-1} \sum_{j=1}^{n} E\{\partial/\partial\boldsymbol{\alpha}\, \boldsymbol{\Psi}_j(Y_j, \boldsymbol{\alpha}_0)|\boldsymbol{x}_j\} \quad \text{and} \quad \boldsymbol{C}_n = n^{-1/2} \sum_{j=1}^{n} \boldsymbol{\Psi}_j(Y_j, \boldsymbol{\alpha}_0),$$

where expectation is of course with respect to the true distribution. We need only find $\boldsymbol{A} = \lim_{n\to\infty} \boldsymbol{A}_n$ and $\boldsymbol{B}$ from applying the central limit theorem to deduce

$$\boldsymbol{C}_n \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{B}).$$

We may then conclude that

$$n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \xrightarrow{L} \mathcal{N}\{\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^T\}.$$

To determine $\boldsymbol{A}$ and $\boldsymbol{B}$, we first recall that we may write

$$\boldsymbol{\Psi}_j(Y_j, \boldsymbol{\alpha}) = \boldsymbol{D}_j^T(\boldsymbol{\alpha})\boldsymbol{V}_j^{-1}(\boldsymbol{\alpha})\{\boldsymbol{s}_j(\boldsymbol{\alpha}) - \boldsymbol{m}_j(\boldsymbol{\alpha})\}, \tag{10.4}$$

where, from (10.2), suppressing the argument $\boldsymbol{\alpha}$ for brevity,

$$\boldsymbol{s}_j - \boldsymbol{m}_j = \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix}, \quad \boldsymbol{V}_j = \begin{pmatrix} \sigma^2 g_j^2 & \zeta_j \sigma^3 g_j^3 \\ \zeta_j \sigma^3 g_j^3 & (2+\kappa_j)\sigma^4 g_j^4 \end{pmatrix},$$

$$\boldsymbol{D}_j^T = \begin{pmatrix} f_{\beta j} & 2\sigma^2 g_j^2 \nu_{\beta j} \\ \mathbf{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\theta j} \end{pmatrix} \end{pmatrix}$$

We will use a subscript "0" to denote evaluation at the true value $\boldsymbol{\alpha}_0$.

Now by the matrix chain rule, recalling that, by $\partial/\partial\boldsymbol{\alpha}\,\boldsymbol{\Psi}_j(Y_j, \boldsymbol{\alpha}_0)$ we mean the indicated partial derivative matrix $\partial/\partial\boldsymbol{\alpha}\,\boldsymbol{\Psi}_j(Y_j, \boldsymbol{\alpha})$ evaluated at $\boldsymbol{\alpha}_0$, we have

$$\partial/\partial\boldsymbol{\alpha}\,\boldsymbol{\Psi}_j(Y_j, \boldsymbol{\alpha}_0) = \{\partial/\partial\boldsymbol{\alpha}\,(\boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1})\}(\boldsymbol{s}_{0j} - \boldsymbol{m}_{0j}) + \boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\{\partial/\partial\boldsymbol{\alpha}\,(\boldsymbol{s}_{0j} - \boldsymbol{m}_{0j})\}.$$

The expectation of the first term is equal to zero because $E(\boldsymbol{s}_{0j} - \boldsymbol{m}_{0j}|\boldsymbol{x}_j) = \mathbf{0}$, so that the derivative of $(\boldsymbol{D}_J^T\boldsymbol{V}_j)$ (which contains ζ_j and κ_j) does not play a role. Thus, if ζ_j and κ_j in $\boldsymbol{V}_j$ depend on $\boldsymbol{\alpha}$, this does not alter the properties of the estimator $\hat{\boldsymbol{\alpha}}$ (through $\boldsymbol{A}$) from those if ζ_j and κ_j were specified as fixed constants. For the second term, it is straightforward to see that $\partial/\partial\boldsymbol{\alpha}\,(\boldsymbol{s}_{0j} - \boldsymbol{m}_{0j}) =$

$$-\begin{pmatrix} f_{\beta 0j}^T & \mathbf{0} \\ 2(Y_j - f_{0j})f_{\beta 0j}^T + 2\sigma_0^2 g_{0j}^2 \nu_{\beta 0j}^T & 2\sigma_0^2(1/\sigma_0,\ g_{0j}^2\nu_{\theta 0j}^T) \end{pmatrix} = -\boldsymbol{D}_{0j} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -2\sigma_0 f_{\beta 0j}^T g_{0j}\epsilon_j & \mathbf{0} \end{pmatrix}.$$

The second term in this expression has expectation zero, as $E(\epsilon_j|\boldsymbol{x}_j) = 0$.

We thus obtain

$$E\{\partial/\partial\boldsymbol{\alpha}\,\boldsymbol{\Psi}_j(Y_j, \boldsymbol{\alpha}_0)|\boldsymbol{x}_j\} = -\boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j}.$$

Write ζ_{0j} and κ_{0j} to denote the expressions for ζ_j and κ_j with $\boldsymbol{\alpha}_0$ substituted in the event these quantities are specified as functions of $\boldsymbol{x}_j$ and $\boldsymbol{\alpha}$; otherwise, ζ_{0j} and κ_{0j} are the assumed fixed constants. Let

$$\boldsymbol{V}_{0j} = \begin{pmatrix} \sigma_0^2 g_{0j}^2 & \zeta_{0j}\sigma^3 g_{0j}^3 \\ \zeta_{0j}\sigma_0^3 g_{0j}^3 & (2+\kappa_{0j})\sigma_0^4 g_{0j}^4 \end{pmatrix}. \tag{10.5}$$

Note that this matrix depends on the *assumed* ζ_j and κ_j.

Thus, $\boldsymbol{A}_n = -n^{-1}\sum_{j=1}^n \boldsymbol{D}_{0j}^T \boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j}$, and $\boldsymbol{A}$ is the limit of this quantity.

Also, we have clearly that $E\{\boldsymbol{\Psi}_j(Y_j,\boldsymbol{\alpha}_0)|\boldsymbol{x}_j\} = \boldsymbol{0}$ and

$$\text{var}\{\boldsymbol{\Psi}_j(Y_j,\boldsymbol{\alpha}_0)|\boldsymbol{x}_j\} = \boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j)\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j}.$$

Here, $\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j) = \boldsymbol{V}_{0j}$ *if* the assumptions $E(\epsilon_j^3|\boldsymbol{x}_j) = \zeta_{0j}$ and $\text{var}(\epsilon_j^2|\boldsymbol{x}_j) = 2+\kappa_{0j}$ are correct.

If the third and fourth moment assumptions used to construct the estimating equation are *not* correct, then there are *true* quantities ζ_j^* and κ_j^*, say, that correspond to the actual third and fourth moments (that may depend on $\boldsymbol{x}_j$ and $\boldsymbol{\alpha}$ or may be constant for all j). Evaluating at $\boldsymbol{\alpha}_0$ and writing ζ_{0j}^* and κ_{0j}^*,

$$\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j) = \begin{pmatrix} \sigma_0^2 g_{0j}^2 & \zeta_{0j}^*\sigma_0^3 g_{0j}^3 \\ \zeta_{0j}^*\sigma_0^3 g_{0j}^3 & (2+\kappa_{0j}^*)\sigma_0^4 g_{0j}^4 \end{pmatrix}.$$

- Comparing this matrix to $\boldsymbol{V}_{0j}$ in (10.5), we see that this true covariance matrix of the "response" (at $\boldsymbol{\alpha}_0$) is *not* the same.
- In particular, it is the choice of third and fourth moment specifications used in forming the quadratic estimating equation that are responsible for the difference. If these choices are incorrect, then, the effect of this *misspecification* of third and fourth moments shows up here.

Putting this all together, we have

$$\boldsymbol{B} = \lim_{n\to\infty}\boldsymbol{B}_n, \quad \boldsymbol{B}_n = n^{-1}\sum_{j=1}^n \boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j)\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j},$$

and thus we may conclude the desired result.

GENERAL RESULT FOR QUADRATIC ESTIMATING EQUATIONS:

$$n^{1/2}\begin{pmatrix} \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \\ \hat{\sigma}-\sigma_0 \\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0 \end{pmatrix} \xrightarrow{L} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}), \tag{10.6}$$

$$\boldsymbol{A} = \lim_{n\to\infty} n^{-1}\sum_{j=1}^n \boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j}, \quad \boldsymbol{B} = \lim_{n\to\infty} n^{-1}\sum_{j=1}^n \boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j)\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j}.$$

Practically speaking, this result implies

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \\ \hat{\sigma}-\sigma_0 \\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0 \end{pmatrix} \dot{\sim} \mathcal{N}\left\{\boldsymbol{0}, \left(\sum_{j=1}^n \boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j}\right)^{-1}\left(\sum_{j=1}^n \boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j)\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j}\right)\left(\sum_{j=1}^n \boldsymbol{D}_{0j}^T\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_{0j}\right)^{-1}\right\}. \tag{10.7}$$

REMARKS:

- This result pertains to the entire vector of parameters. To deduce the limiting distribution of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ by itself, it would be necessary to carry out the indicated matrix inversion and multiplications in (10.6) and (10.7) and extract the upper left $(p \times p)$ submatrix of the result.

  Given the definitions of the matrices $\boldsymbol{D}_{0j}$, $\boldsymbol{V}_{0j}$, and $\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j)$, it seems obvious that the ultimate result could be quite unwieldy. The unfortunate consequence is that it is difficult to gain easily constructive insight.

- Note that if the third and fourth moment assumptions used to form the estimating equation are correct, then $\zeta_{0j} = \zeta_{0j}^*$ and $\kappa_{0j} = \kappa_{0j}^*$ for all j, and

$$\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j) = \boldsymbol{V}_{0j}.$$

  Note then that the results in (10.6) and (10.7) become

$$n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\sigma} - \sigma_0 \\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \end{pmatrix} \xrightarrow{L} \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^{-1}),$$

  and

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\sigma} - \sigma_0 \\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \end{pmatrix} \dot{\sim} \mathcal{N}\left\{ \boldsymbol{0}, \left(\sum_{j=1}^{n} \boldsymbol{D}_{0j}^T \boldsymbol{V}_{0j}^{-1} \boldsymbol{D}_{0j} \right)^{-1} \right\},$$

  respectively.

- Thus, the comparison between the quadratic estimating equation using the *correct* covariance matrix ("weights") and that using *misspecified* weights (through incorrect assumptions on skewness and kurtosis) is of exactly the *same form* as that encountered when comparing (linear) GLS estimators with correct and incorrect weights! By the same type of argument, it is possible to show that the asymptotic covariance matrix in the case that the third and fourth moments are correctly specified is "smaller" than that when they are not.

  This supports the conjecture at the end of Chapter 9 that, in the quadratic case, the estimating equation of the form (10.4) is "optimal" in the sense of satisfying an "asymptotic Gauss-Markov property." In particular, presumably, a similar argument could be used to show that solving (10.4) in this particular case with the correct third and fourth moment assumptions would lead to the most efficient estimator for $(\boldsymbol{\beta}^T, \sigma, \boldsymbol{\theta}^T)^T$ *within the class of all estimators* solving quadratic estimating equations depending on the data through $\boldsymbol{s}_j - \boldsymbol{m}_j$ defined in (10.4).

- In fact, it is possible to deduce generally that the estimator for $\boldsymbol{\beta}$ found by solving the quadratic estimating equation with the correct "weights" (so correct variance function, skewness, and kurtosis) is more efficient than the GLS estimator for $\boldsymbol{\beta}$ found by solving the linear estimating equation with the correct "weights" (so correct variance function). Momentarily, we will exhibit this result explicitly in a special case.

Now that we have derived the results for the general quadratic estimating equation (10.2), it is worth noting an interesting feature. If one carries out the matrix multiplication in (10.2) (inverting the "covariance" matrix), it should be clear that both the estimating equations corresponding to estimation of $\boldsymbol{\beta}$ and of $(\sigma, \boldsymbol{\theta}^T)^T$ involve linear *and* quadratic terms $(Y_j - f_j)$ and $(Y_j - f_j)^2 - \sigma^2 g_j^2$. Note that this is considerably more complicated than the sets of equations we have discussed previously, where the equation for $(\sigma, \boldsymbol{\theta}^T)^T$ involved *only* a quadratic component.

- If we assume that the third and fourth moment assumptions are correct, with skewness $\neq 0$, then the linear and quadratic terms are *correlated*. Thus, this suggests that when such correlation exists, despite the fact that the mean does not involve σ and $\boldsymbol{\theta}$, there is information to be gained on the variance parameters from the *linear* function of the data. As mean and variance depend on $\boldsymbol{\beta}$, there is possibly information to be gained about $\boldsymbol{\beta}$ from both the linear and quadratic functions, regardless of the correlation.

  This may seem counterintuitive, but it is worth remembering that in simpler problems estimation of variance is often considered separately from that of mean, mostly because of the heavy emphasis on the normal distribution. In the case of more general distributions, this shows that, to gain efficiency, it is necessary to exploit additional information.

SOME SPECIFIC COMPARISONS: To gain some insight into the properties of the estimator for $\boldsymbol{\beta}$, we will consider a simpler form of the quadratic estimating equation (10.2). In particular, we will simplify things by focusing on the situation where the equation is formed by assuming that $\zeta_j \equiv 0$ and $\kappa_j \equiv \kappa$ for all j; for simplicity, we will regard κ as a fixed constant, although it could also depend on the parameters (just not on j). Under these conditions, the estimating equation becomes

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma^2 g_j^2 \nu_{\beta j} \\ \mathbf{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\theta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & 0 \\ 0 & (2+\kappa)\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}. \qquad (10.8)$$

Note that now the "covariance matrix" used to form the estimating equation is a *diagonal* matrix. This feature simplifies the calculations.

Moreover, the assumption of *symmetry* (skewness=0) is one of those corresponding to the normal distribution. Thus, considering this situation subsumes the particular case of the quadratic estimating equations corresponding to normal theory ML.

To deduce the large sample distribution of the estimators solving (10.8), we may appeal directly to the M-estimator argument. This will yield a practical result of the form (10.7), where the covariance matrix $\boldsymbol{V}_{0j}$ is now diagonal, so has a straightforward inverse. This result pertains to all estimators, so to obtain the asymptotic covariance matrix for $\hat{\boldsymbol{\beta}}$ alone, we need to carry out the multiplication to extract the upper left $(p \times p)$ submatrix of the joint covariance matrix of $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}$, and $\hat{\boldsymbol{\theta}}$.

Rather than just proceeding directly in this way, evaluating $\boldsymbol{A}_n = n^{-1}\sum_{j=1}^n \boldsymbol{D}_{0j}^T \boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_j$ and $\boldsymbol{B}_n = n^{-1}\sum_{j=1}^n \boldsymbol{D}_{0j}^T \boldsymbol{V}_{0j}^{-1}\text{var}(\boldsymbol{s}_{0j}|\boldsymbol{x}_j)\boldsymbol{V}_{0j}^{-1}\boldsymbol{D}_j$ and then $n^{-1}\boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{-1}$ and deducing its upper left $(p \times p)$ submatrix, we will modify things slightly in a way that will prove very convenient both now and when we consider properties of variance parameter estimators in Chapter 12.

In particular, if we write the expansion of the estimating equation (10.8) in the usual way, it is straightforward to verify that we obtain

$$n^{1/2}\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\sigma} - \sigma_0 \\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \end{pmatrix} \approx -\boldsymbol{A}_n^{-1}\boldsymbol{C}_n, \quad \text{or} \quad -\boldsymbol{A}_n n^{1/2}\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\sigma} - \sigma_0 \\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \end{pmatrix} \approx \boldsymbol{C}_n, \tag{10.9}$$

where here

$$\boldsymbol{A}_n = -n^{-1}\sum_{j=1}^n \begin{pmatrix} \sigma_0^{-2} g_{0j}^{-2} f_{\beta 0j} f_{\beta 0j}^T + \frac{4}{2+\kappa}\nu_{\beta 0j}\nu_{\beta 0j}^T & \frac{4}{2+\kappa}\nu_{\beta 0j}\begin{pmatrix} 1/\sigma_0 & \nu_{\theta 0j}^T \end{pmatrix} \\ \frac{4}{2+\kappa}\begin{pmatrix} 1/\sigma_0 \\ \nu_{\theta 0j} \end{pmatrix}\nu_{\beta 0j}^T & \frac{4}{2+\kappa}\begin{pmatrix} 1/\sigma_0^2 & 1/\sigma_0\nu_{\theta 0j}^T \\ 1/\sigma_0\nu_{\theta 0j} & \nu_{\theta 0j}\nu_{\theta 0j}^T \end{pmatrix} \end{pmatrix}.$$

We also have, expressing things in terms of ϵ_j,

$$\boldsymbol{C}_n = n^{-1/2}\sum_{j=1}^n \begin{pmatrix} \sigma_0^{-1} g_{0j}^{-1} f_{\beta 0j}\epsilon_j + 2(2+\kappa)^{-1}(\epsilon_j^2 - 1)\nu_{\beta 0j} \\ 2(2+\kappa)^{-1}(\epsilon_j^2 - 1)\begin{pmatrix} 1/\sigma_0 \\ \nu_{\theta 0j} \end{pmatrix} \end{pmatrix}.$$

It is straightforward to verify that (10.9) may be rewritten, by multiplying through each row by a multiplicative factor depending on σ_0, as follows.

Equation (10.9) becomes

$$n^{-1}\sum_{j=1}^{n}\begin{pmatrix} g_{0j}^{-2}f_{\beta 0j}f_{\beta 0j}^T+\frac{4\sigma_0^2}{2+\kappa}\nu_{\beta 0j}\nu_{\beta 0j}^T & \frac{4\sigma_0}{2+\kappa}\nu_{\beta 0j}\begin{pmatrix}1 & \nu_{\theta 0j}^T\end{pmatrix} \\ \frac{4\sigma_0}{2+\kappa}\begin{pmatrix}1\\ \nu_{\theta 0j}\end{pmatrix}\nu_{\beta 0j}^T & \frac{4}{2+\kappa}\begin{pmatrix}1 & \nu_{\theta 0j}^T\\ \nu_{\theta 0j} & \nu_{\theta 0j}\nu_{\theta 0j}^T\end{pmatrix}\end{pmatrix} n^{1/2}\begin{pmatrix}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/\sigma_0\\ (\hat{\sigma}-\sigma_0)/\sigma_0\\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\end{pmatrix}$$

$$\approx n^{-1/2}\sum_{j=1}^{n}\begin{pmatrix} g_{0j}^{-1}f_{\beta 0j}\epsilon_j+2\sigma_0(2+\kappa)^{-1}(\epsilon_j^2-1)\nu_{\beta 0j}\\ 2(2+\kappa)^{-1}(\epsilon_j^2-1)\begin{pmatrix}1\\ \nu_{\theta 0j}\end{pmatrix}\end{pmatrix}.$$

The scaling by σ_0 will turn out to be fortuitous shortly.

Recalling that $\tau_\theta(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{x}_j)=\{1,\nu_\theta^T(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{x}_j)\}^T$, if we define as before $\boldsymbol{X}=\boldsymbol{X}(\boldsymbol{\beta}_0)$, $\boldsymbol{W}=\text{diag}(g_{01}^{-2},\ldots,g_{0n}^{-2})$,

$$\boldsymbol{R}=\begin{pmatrix}\nu_{\beta 01}^T\\ \vdots\\ \nu_{\beta 0n}^T\end{pmatrix}\ (n\times p),\quad \boldsymbol{Q}=\begin{pmatrix}\tau_{\theta 01}^T\\ \vdots\\ \tau_{\theta 0n}^T\end{pmatrix}\ (n\times q+1),$$

then we may write this more succinctly as

$$n^{-1}\begin{pmatrix}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+\frac{4\sigma_0^2}{2+\kappa}\boldsymbol{R}^T\boldsymbol{R} & \frac{4\sigma_0}{2+\kappa}\boldsymbol{R}^T\boldsymbol{Q}\\ \frac{4\sigma_0}{2+\kappa}\boldsymbol{Q}^T\boldsymbol{R} & \frac{4}{2+\kappa}\boldsymbol{Q}^T\boldsymbol{Q}\end{pmatrix} n^{1/2}\begin{pmatrix}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/\sigma_0\\ (\hat{\sigma}-\sigma_0)/\sigma_0\\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\end{pmatrix}$$

$$\approx n^{-1/2}\sum_{j=1}^{n}\begin{pmatrix} g_{0j}^{-1}f_{\beta 0j}\epsilon_j+2\sigma_0(2+\kappa)^{-1}(\epsilon_j^2-1)\nu_{\beta 0j}\\ 2(2+\kappa)^{-1}(\epsilon_j^2-1)\tau_{\theta 0j}\end{pmatrix}. \qquad (10.10)$$

We write this as

$$n^{1/2}\begin{pmatrix}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/\sigma_0\\ (\hat{\sigma}-\sigma_0)/\sigma_0\\ \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\end{pmatrix}\approx \boldsymbol{A}_n^{*-1}\boldsymbol{C}_n^*.$$

We already have the form of $\boldsymbol{A}_n^*$; it remains to determine the form of $\boldsymbol{A}_n^{*-1}$ and the behavior of $\boldsymbol{C}_n^*$. First consider $\boldsymbol{C}_n^*$.

Suppose that, *in truth*, the skewness and kurtosis of the response are not necessarily the same as those assumed in forming the estimating equation (10.8). In particular, suppose that in truth

$$E(\epsilon_j^3|\boldsymbol{x}_j)=\zeta^*,\quad \text{and } \text{var}(\epsilon_j^2|\boldsymbol{x}_j)=2+\kappa^*.$$

Under these conditions, we apply the multivariate central limit theorem to $\boldsymbol{C}_n^*$. We observe that a summand of $\boldsymbol{C}_n^*$ of course has mean zero and covariance matrix

$$\begin{pmatrix} g_{0j}^{-2} f_{\beta 0j} f_{\beta 0j}^T + \frac{4\sigma_0^2(2+\kappa^*)}{(2+\kappa)^2}\nu_{\beta 0j}\nu_{\beta 0j}^T & \frac{2\zeta^*}{(2+\kappa)g_{0j}} f_{\beta 0j}\tau_{\theta 0j}^T + \frac{4\sigma_0(2+\kappa^*)}{(2+\kappa)^2}\nu_{\beta 0j}\tau_{\theta 0j}^T \\ \quad + \frac{2\sigma_0\zeta^*}{(2+\kappa)g_{0j}}\{f_{\beta 0j}\nu_{\beta 0j}^T + f_{\beta 0j}^T\nu_{\beta 0j}\} & \\ \\ \frac{2\zeta^*}{(2+\kappa)g_{0j}}\tau_{\theta 0j} f_{\beta 0j}^T + \frac{4\sigma_0(2+\kappa^*)}{(2+\kappa)^2}\tau_{\theta 0j}\nu_{\beta 0j}^T & \frac{4(2+\kappa^*)}{(2+\kappa)^2}\tau_{\theta 0j}\tau_{\theta 0j}^T \end{pmatrix}.$$

Thus, using the definitions above, it is straightforward to conclude that $\boldsymbol{C}_n^* \xrightarrow{L} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}^*)$, where $\boldsymbol{B}^* = \lim_{n\to\infty} \boldsymbol{B}_n^*$ and

$$\boldsymbol{B}_n^* = n^{-1}\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \frac{4\sigma_0^2(2+\kappa^*)}{(2+\kappa)^2}\boldsymbol{R}^T\boldsymbol{R} & \frac{2\zeta^*}{2+\kappa}\boldsymbol{X}^T\boldsymbol{W}^{1/2}\boldsymbol{Q} + \frac{4\sigma_0(2+\kappa^*)}{(2+\kappa)^2}\boldsymbol{R}^T\boldsymbol{Q} \\ \quad + \frac{2\sigma_0\zeta^*}{2+\kappa}\{\boldsymbol{X}^T\boldsymbol{W}^{1/2}\boldsymbol{R} + \boldsymbol{R}^T\boldsymbol{W}^{1/2}\boldsymbol{X}\} & \\ \\ \frac{2\zeta^*}{2+\kappa}\boldsymbol{Q}^T\boldsymbol{W}^{1/2}\boldsymbol{X} + \frac{4\sigma_0(2+\kappa^*)}{(2+\kappa)^2}\boldsymbol{Q}^T\boldsymbol{R} & \frac{4(2+\kappa^*)}{(2+\kappa)^2}\boldsymbol{Q}^T\boldsymbol{Q} \end{pmatrix}. \tag{10.11}$$

Now consider inversion of $\boldsymbol{A}_n^*$. This calculation is facilitated by the following well known matrix result.

INVERSE OF PARTITIONED MATRIX: Consider a generic $(k \times k)$ matrix

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix},$$

where the $\boldsymbol{C}_{ij}$ are submatrices such that $\boldsymbol{C}_{11}$ is $(k_1 \times k_1)$ and $\boldsymbol{C}_{22}$ is $(k_2 \times k_2)$ such that $k = k_1 + k_2$, and $\boldsymbol{C}_{11}^{-1}$ and $\boldsymbol{C}_{22}^{-1}$ exist, as do all other inverses below. Then

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{D}_{12} \\ \boldsymbol{D}_{21} & \boldsymbol{D}_{22} \end{pmatrix},$$

where

$$\begin{aligned} \boldsymbol{D}_{11} &= (\boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21})^{-1} \\ \boldsymbol{D}_{22} &= (\boldsymbol{C}_{22} - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12})^{-1} = \boldsymbol{C}_{22}^{-1} + \boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{D}_{11}\boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1} \\ \boldsymbol{D}_{12} &= -\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12}\boldsymbol{D}_{22} = -\boldsymbol{D}_{11}\boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1} \\ \boldsymbol{D}_{21} &= -\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}\boldsymbol{D}_{11}. \end{aligned}$$

Write

$$\boldsymbol{A}_n^{*-1} = n^{-1} \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \frac{4\sigma_0^2}{2+\kappa}\boldsymbol{R}^T\boldsymbol{R} & \frac{4\sigma_0}{2+\kappa}\boldsymbol{R}^T\boldsymbol{Q} \\ \frac{4\sigma_0}{2+\kappa}\boldsymbol{Q}^T\boldsymbol{R} & \frac{4}{2+\kappa}\boldsymbol{Q}^T\boldsymbol{Q} \end{pmatrix}^{-1} = n^{-1} \begin{pmatrix} \boldsymbol{A}_n^{*11} & \boldsymbol{A}_n^{*12} \\ \boldsymbol{A}_n^{*12T} & \boldsymbol{A}_n^{*22} \end{pmatrix}.$$

From the form of $\boldsymbol{A}_n^*$ and the partitioned matrix result, we obtain

$$\begin{aligned} \boldsymbol{A}_n^{*11} &= \{\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \frac{4\sigma_0^2}{2+\kappa}\boldsymbol{R}^T\boldsymbol{R} - \frac{4\sigma_0^2}{2+\kappa}\boldsymbol{R}^T\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{R}\}^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \frac{4\sigma_0^2}{2+\kappa}\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1}, \quad \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T. \\ \boldsymbol{A}_n^{*12} &= -\sigma_0(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \frac{4\sigma_0^2}{2+\kappa}\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1} \\ &= -\sigma_0\boldsymbol{A}_n^{*11}\boldsymbol{S}, \quad \boldsymbol{S} = \boldsymbol{R}^T\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{Q})^{-1} \end{aligned}$$

Now it is straightforward to verify that the upper left $(p \times p)$ submatrix of $n^{-1}\boldsymbol{A}_n^{*-1}\boldsymbol{B}_n^*\boldsymbol{A}_n^{*-1}$ (note we have multiplied by n^{-1} in anticipation of the asymptotic distribution result) is given by

$$\boldsymbol{A}_n^{*11}\boldsymbol{B}_{n11}^*\boldsymbol{A}_n^{*11} + \boldsymbol{A}_n^{*12}\boldsymbol{B}_{n12}^{*T}\boldsymbol{A}_n^{*11} + \boldsymbol{A}_n^{*11}\boldsymbol{B}_{n12}^*\boldsymbol{A}_n^{*12T} + \boldsymbol{A}_n^{*12}\boldsymbol{B}_{n22}^*\boldsymbol{A}_n^{*12T},$$

which, using the expressions above, may be rewritten as

$$\boldsymbol{A}_n^{*11}(\boldsymbol{B}_{n11}^* - \sigma_0\boldsymbol{S}\boldsymbol{B}_{n12}^{*T} - \sigma_0\boldsymbol{B}_{n12}^*\boldsymbol{S}^T + \sigma_0^2\boldsymbol{S}\boldsymbol{B}_{n22}^*\boldsymbol{S}^T)\boldsymbol{A}_n^{*11}. \tag{10.12}$$

where the matrices $\boldsymbol{B}_{n11}^*$, $\boldsymbol{B}_{n12}^*$, and $\boldsymbol{B}_{n22}^*$, are the obvious submatrices of $\boldsymbol{B}_n^*$.

Using (10.11) and the definition of $\boldsymbol{S}$, the middle term in (10.12) may be seen to simplify to, by tedious algebra,

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \frac{4\sigma_0^2(2+\kappa^*)}{(2+\kappa)^2}\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R} + \frac{2\sigma_0\zeta^*}{2+\kappa}(\boldsymbol{X}^T\boldsymbol{W}^{1/2}\boldsymbol{P}\boldsymbol{R} + \boldsymbol{R}^T\boldsymbol{P}\boldsymbol{W}^{1/2}\boldsymbol{X}).$$

Combining all of this, we obtain the final result.

LARGE SAMPLE DISTRIBUTION, QUADRATIC ESTIMATING EQUATION WITH $\zeta_j \equiv 0$ *AND* $\kappa_j = \kappa$*:* If we construct quadratic estimating equations for $(\boldsymbol{\beta}^T, \sigma, \boldsymbol{\theta}^T)^T$ as in (10.8) by taking $E(\epsilon_j^3|\boldsymbol{x}_j) \equiv 0$ and $\text{var}(\epsilon_j^2|\boldsymbol{x}_j) = 2 + \kappa$ for all j, and, *in truth*, $E(\epsilon_j^3|\boldsymbol{x}_j) = \zeta^*$ and $\text{var}(\epsilon_j^2|\boldsymbol{x}_j) = 2 + \kappa^*$ for all j, then

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{L} \mathcal{N}(\boldsymbol{0}, \sigma_0^2\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}), \tag{10.13}$$

where

$$\boldsymbol{\Gamma} = \lim_{n\to\infty}\boldsymbol{\Gamma}_n = \lim_{n\to\infty} n^{-1}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \frac{4\sigma_0^2}{2+\kappa}\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R}) = \boldsymbol{\Sigma}_{WLS}^{-1} + \frac{4\sigma_0^2}{2+\kappa}\boldsymbol{\Sigma}_\beta, \tag{10.14}$$

and

$$\begin{aligned} \boldsymbol{\Delta} = \lim_{n\to\infty}\boldsymbol{\Delta}_n &= \lim_{n\to\infty} n^{-1}\left\{\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \frac{4\sigma_0^2(2+\kappa^*)}{(2+\kappa)^2}\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R} + \frac{2\sigma_0\zeta^*}{2+\kappa}(\boldsymbol{X}^T\boldsymbol{W}^{1/2}\boldsymbol{P}\boldsymbol{R} + \boldsymbol{R}^T\boldsymbol{P}\boldsymbol{W}^{1/2}\boldsymbol{X})\right\} \\ &= \boldsymbol{\Sigma}_{WLS}^{-1} + \frac{4\sigma_0^2(2+\kappa^*)}{(2+\kappa)^2}\boldsymbol{\Sigma}_\beta + \frac{2\sigma_0\zeta^*}{2+\kappa}(\boldsymbol{T}_\beta + \boldsymbol{T}_\beta^T). \end{aligned} \tag{10.15}$$

We may write this in a more "practical" way as

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \dot{\sim} \mathcal{N}(\mathbf{0}, \sigma_0^2 n^{-1} \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\Delta}_n \boldsymbol{\Gamma}_n^{-1}),$$

of course. However, it is evident that the form of the large sample covariance matrix is very complicated, and depends not only on the assumed skewness and excess kurtosis (0 and κ) but also the *true* ones, assumed to be ζ^* and κ^* here.

This result is of little practical use. Its main usefulness is in the insights that may be gleaned from it, as follows.

One immediate conclusion from the calculations and the final result is that, unlike for GLS, the properties of $\hat{\boldsymbol{\beta}}$ *do depend* on those of $\hat{\sigma}$ and $\hat{\boldsymbol{\theta}}$. This is easily deduced from inspection of (10.10); clearly, the properties of all the estimators are linked together. This is also evident in the final result: unlike the large sample covariance matrix of $\hat{\boldsymbol{\beta}}_{GLS}$, which depends only on the matrices $\boldsymbol{X}$ and $\boldsymbol{W}$ having to do with gradient of the mean function and the "weights," the large sample covariance matrix of the quadratic estimator depends also on the matrices $\boldsymbol{R}$ and $\boldsymbol{Q}$, which clearly arise from the dependence of the variance model on $\boldsymbol{\beta}$, σ, and $\boldsymbol{\theta}$.

Thus, in the above result, the properties of the quadratic estimator for $\boldsymbol{\beta}$ are obviously tied up with those of the quadratic variance parameter estimators used for σ and $\boldsymbol{\theta}$.

We now consider some special cases.

(a) *Use normal ML and the data really are normal.* If we choose $\kappa \equiv 0$ for all j in setting up the quadratic estimating equations, then solving them jointly in $(\boldsymbol{\beta}^T, \sigma, \boldsymbol{\theta}^T)^T$ corresponds to joint normal theory ML estimation of all parameters. If in fact $\zeta^* = 0$ and $\kappa^* = 0$, then the ϵ_j have first four moments corresponding exactly to those of the $\mathcal{N}(0,1)$ distribution; certainly, this would be true if the data truly are normally distributed. Note, however, that the result we are about to deduce really only depends on the true first four moments matching those of the normal.

Under these conditions, it is easy to see that (10.14) and (10.15) reduce to

$$\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{WLS}^{-1} + 2\sigma_0^2 \boldsymbol{\Sigma}_\beta = \boldsymbol{\Sigma}_{ML}^{-1} \quad \text{and} \quad \boldsymbol{\Delta} = \boldsymbol{\Sigma}_{ML}^{-1},$$

say, so that (10.13) becomes

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Sigma}_{ML}). \tag{10.16}$$

(10.16) represents the large sample behavior of the joint normal theory ML estimator for $\boldsymbol{\beta}$ when the data are really normally distributed (or at least share the first four normal moments).

Usual large sample theory for maximum likelihood implies that the covariance matrix $\sigma_0^2\boldsymbol{\Sigma}_{ML}$ is thus the "smallest" that may be attained if the model (10.1) is correct and the data truly are normal.

Recall that the GLS estimator with the "optimal" weights using the correct variance function satisfies

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2\boldsymbol{\Sigma}_{WLS}).$$

This is true *regardless* of the true distribution of the data, as the arguments in Chapter 9 required only that the first two moments of $Y_j|\boldsymbol{x}_j$ are correctly specified.

Comparing the large sample covariance matrices of the ML estimator $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{GLS}$, it is evident that

$$\boldsymbol{\Sigma}_{WLS} \text{“} \geq \text{”} \boldsymbol{\Sigma}_{ML} = (\boldsymbol{\Sigma}_{WLS}^{-1} + 2\sigma_0^2\boldsymbol{\Sigma}_\beta)^{-1},$$

assuming that $\boldsymbol{\Sigma}_{WLS}$ and $\boldsymbol{\Sigma}_\beta$ are positive definite. This may be deduced by noting that

$$(\boldsymbol{\Sigma}_{WLS}^{-1} + 2\sigma_0^2\boldsymbol{\Sigma}_\beta)^{-1} = \boldsymbol{\Sigma}_{WLS} - \boldsymbol{\Sigma}_{WLS}\{\boldsymbol{\Sigma}_{WLS} + (\sigma_0^2/2)\boldsymbol{\Sigma}_\beta^{-1})^{-1}\boldsymbol{\Sigma}_{WLS}.$$

Thus, we see that, in the case where the data truly are normal and we use the normal theory ML estimating equations for all parameters, $\hat{\boldsymbol{\beta}}_{GLS}$ is *inefficient* relative to the quadratic estimator $\hat{\boldsymbol{\beta}}$, which is the "optimal" choice under these conditions. Solving the "optimal" linear estimating equation still results in an estimator for $\boldsymbol{\beta}$ that is less precise (asymptotically) than the quadratic ML estimator.

The source of the improvement is evident from the form of $\boldsymbol{\Sigma}_{ML}$. The term

$$\boldsymbol{\Sigma}_\beta = \lim_{n\to\infty} n^{-1}\boldsymbol{R}^T\boldsymbol{P}\boldsymbol{R}$$

depends on the function $\nu_\beta(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j)$ in $\boldsymbol{R}$, which arises from the appearance of $\boldsymbol{\beta}$ in the variance function. Thus, it is apparent that this term represents taking advantage of the additional information on $\boldsymbol{\beta}$ available in g. That is, when the data truly are normal (or at least share the first four moments of the normal), the quadratic equation exploits this extra information in a way that the linear GLS equation cannot.

(b) *Use normal ML but the data are really only symmetrically distributed.* Now suppose that we again form the quadratic estimating equation using the normal skewness and kurtosis (so take $\kappa \equiv 0$), but, *in truth*, although the data are symmetrically distributed, they are more prone to extreme observations than would be expected under normality in a way that happens similarly across all j. So, formally, $\zeta^* \equiv 0$ but $\kappa^* > 0$.

Under these conditions,

$$\mathbf{\Gamma} = \mathbf{\Sigma}_{WLS}^{-1} + 2\sigma_0^2\mathbf{\Sigma}_\beta = \mathbf{\Sigma}_{ML}^{-1} \quad \text{and} \quad \mathbf{\Delta} = \mathbf{\Sigma}_{WLS}^{-1} + (2+\kappa^*)\sigma_0^2\mathbf{\Sigma}_\beta,$$

so that

$$\mathbf{\Gamma}^{-1}\mathbf{\Delta}\mathbf{\Gamma}^{-1} = \mathbf{\Sigma}_Q = (\mathbf{\Sigma}_{WLS}^{-1} + 2\sigma_0^2\mathbf{\Sigma}_\beta)^{-1}\{\mathbf{\Sigma}_{WLS}^{-1} + (2+\kappa^*)\sigma_0^2\mathbf{\Sigma}_\beta\}(\mathbf{\Sigma}_{WLS}^{-1} + 2\sigma_0^2\mathbf{\Sigma}_\beta)^{-1}.$$

How do $\hat{\boldsymbol{\beta}}_{GLS}$ and the quadratic estimator $\hat{\boldsymbol{\beta}}$ compare now? The properties of $\hat{\boldsymbol{\beta}}_{GLS}$ are unchanged. It is no longer immediately clear that $\mathbf{\Sigma}_{WLS}$" $\geq$" $\mathbf{\Sigma}_Q$. In fact, one may try very hard to establish a general ordering for any choice of f and g; however, it only seems possible to show that $\mathbf{\Sigma}_Q$ is "smaller" if $\kappa^* \leq 2$.

Thus, if we use the normal theory ML quadratic estimator for $\boldsymbol{\beta}$ but the data are only symmetrically distributed with positive excess kurtosis, then the "optimality" advantage no longer applies uniformly; in fact, it is not clear which estimator is to be preferred.

(c) *Use the quadratic estimating equation (10.8) with* $\zeta \equiv 0$ *assuming* $\text{var}(\epsilon_j|\boldsymbol{x}_j) = 2+\kappa$ *for some constant* κ *where in fact* $\zeta \equiv \zeta^* \equiv 0$ *and* $\kappa \equiv \kappa^*$. In this situation, we are willing to specify that the data are symmetrically distributed with some common fourth moment, and we are *correct*. Thus, the quadratic estimating equation is "optimal."

Under these conditions, we have

$$\mathbf{\Gamma} = \mathbf{\Sigma}_{WLS}^{-1} + \frac{4\sigma_0^2}{2+\kappa^*}\mathbf{\Sigma}_\beta \quad \text{and} \quad \mathbf{\Delta} = \mathbf{\Sigma}_{WLS}^{-1} + \frac{4\sigma_0^2}{2+\kappa^*}\mathbf{\Sigma}_\beta,$$

so that

$$\mathbf{\Gamma}^{-1}\mathbf{\Delta}\mathbf{\Gamma}^{-1} = \{\mathbf{\Sigma}_{WLS}^{-1} + \frac{4\sigma_0^2}{2+\kappa^*}\mathbf{\Sigma}_\beta\}^{-1}.$$

Again, the properties of $\hat{\boldsymbol{\beta}}_{GLS}$ are unchanged. By an argument similar to that in case (a) above, we have

$$\mathbf{\Sigma}_{WLS}\text{" } \geq \text{ "}\{\mathbf{\Sigma}_{WLS}^{-1} + \frac{4\sigma_0^2}{2+\kappa^*}\mathbf{\Sigma}_\beta\}^{-1}.$$

Thus, the quadratic estimator is "better" than the linear GLS estimator when we know the data are symmetric and are able to specify correctly a value for the excess kurtosis. (This extends to the case where we specify a function.)

This is of theoretical interest, but, in practice, it would be very unusual that one would feel comfortable being able to specify the excess kurtosis if one did not believe in the normality assumption! In general, as we have pointed out previously, being able to deduce the form of higher (than two) moments from observed data is very difficult.

(d) *g does not depend on $\boldsymbol{\beta}$.* In this case, $\boldsymbol{\Sigma}_\beta \equiv \mathbf{0}$ and $\boldsymbol{T}_\beta \equiv \mathbf{0}$, as $\nu_\beta(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j) \equiv \mathbf{0}$ so that $\boldsymbol{R} = \mathbf{0}$. Thus, if g does not depend on $\boldsymbol{\beta}$, as expected, there is no additional information on $\boldsymbol{\beta}$ to be gained from g. Then, *regardless* of the assumed or true third and fourth moments, (10.13) reduces to

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Sigma}_{WLS}),$$

as $\boldsymbol{\Gamma} = \boldsymbol{\Delta} = \boldsymbol{\Sigma}_{WLS}$ here. Of course, $\nu_\beta \equiv \mathbf{0}$ renders the estimating equation linear.

This verifies our intuition: for estimating $\boldsymbol{\beta}$, there is nothing to be gained by using a quadratic equation over a linear one (GLS) if g does not involve the regression parameter $\boldsymbol{\beta}$.

REMARKS: Examination of this special case highlights some of the general properties in the comparison between GLS and the quadratic estimator (with $(\sigma, \boldsymbol{\theta}^T)^T$ also estimated by the corresponding quadratic estimators). Assuming that the first two moments are correctly specified in (10.1):

- The large sample properties of the quadratic estimator depend on the assumed and true third and fourth moments of the data. Those of the GLS estimator do not, and are unchanged regardless of the nature of the true third and fourth moments.
- If the third and fourth moments are correctly specified, the "optimal" GLS estimator is inefficient relative to the resulting quadratic estimator for $\boldsymbol{\beta}$. If these are not correctly specified, it is no longer clear that one estimator dominates the other in terms of efficiency.
- Although we have made some comparisons, it is clear that computing the actual value of ARE for GLS relative to the quadratic estimator would depend on the problem: the nature of f, g, design, and so on. Thus, these observations are qualitative; although GLS is inefficient in some circumstances, how much of a loss really results will depend on the problem.
- Intuitively, because the performance of the quadratic estimator depends on third and fourth moment properties, it would seem to be *sensitive* to incorrect assumptions about them, whereas the performance of the GLS estimator does not depend on these moments at all. In the next section, we investigate this in a special case.
- A *trade-off* between linear and quadratic estimating equations for $\boldsymbol{\beta}$ seems to be emerging. If we are confident in our ability to specify third and fourth moments of $Y_j|\boldsymbol{x}_j$, then we can exploit that knowledge to obtain a potentially more precise estimator than could be obtained via GLS. So, for example, if we are confident that the data truly are normal, we are better off using normal theory ML in terms of efficiency.

However, if we are wrong about the third and fourth moments, it seems that going to the extra trouble of using the quadratic equation could hurt rather than help. As the properties of the GLS estimator do not depend at all on third and fourth moments, and no specification of them is required, it would be insensitive to this issue, so might be a "safer" choice.

To gain greater understanding of the trade-off, we consider misspecification of third and fourth moments for the quadratic estimating equation in the next section.

10.3 Robustness of linear and quadratic equations to misspecification of third and fourth moments

As mentioned in the last section, although we can gain general insights into the relative performance of GLS and the quadratic estimating equations, nothing we have done actually quantifies the degree to which one estimator might be inefficient relative to the other under different circumstances. One way to do this for a particular problem (models f and g, true values of parameters, distribution for $Y_j|\boldsymbol{x}_j$, and settings of $\boldsymbol{x}_j$) would be to substitute these into the expressions for fixed n or to carry out simulations.

Here, we do the former and consider the implications of the theory in a very simple model for which AREs can be calculated under different conditions. The following is taken from Carroll and Ruppert (1988, pp. 21–23).

Consider the simple mean-variance model

$$E(Y_j) = \beta, \quad \text{var}(Y_j) = \sigma^2\beta^2, \quad \beta \neq 0. \tag{10.17}$$

Here, we do not condition on covariates, as there are none. The main feature of (10.17) is that the variance depends on the mean (in fact, the Y_j have constant coefficient of variation σ under this model). As usual, denote the true values of the parameters by β_0 and σ_0; for simplicity, take $\sigma^2 = \sigma_0^2$ to be known; and let $\epsilon_j = (Y_j - \beta_0)/(\sigma_0\beta_0)$.

LINEAR ESTIMATING EQUATION (GLS): Regardless of the third and fourth moments of the Y_j, the GLS estimating equation ($C = \infty$) is

$$\sum_{j=1}^{n} \beta^{-2}(Y_j - \beta) = 0,$$

which yields

$$\hat{\beta}_{GLS} = n^{-1}\sum_{j=1}^{n} Y_j = \bar{Y}.$$

By elementary application of the weak law of large numbers and the central limit theorem, we may show explicitly the $\hat{\boldsymbol{\beta}}_{GLS}$ is consistent, as

$$\bar{Y} \xrightarrow{p} \beta_0,$$

and

$$n^{1/2}(\hat{\beta}_{GLS} - \beta_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \Sigma_{WLS}), \quad \Sigma_{WLS} = \beta_0^2.$$

Note that this is true *regardless* of the values of $E(\epsilon_j^3)$ and $\text{var}(\epsilon_j^2)$.

QUADRATIC ESTIMATING EQUATION (NORMAL ML): Consider the particular quadratic estimating equation corresponding to the assumptions of the normal skewness and kurtosis. From (10.8) with $\kappa \equiv 0$ and considering only the first p rows corresponding to estimation of β, the estimating equation is

$$\frac{Y_j - \beta}{\sigma_0^2 \beta^2} + \frac{(Y_j - \beta)^2 - \sigma_0^2 \beta^2}{\sigma_0^2 \beta^3} = 0,$$

which may be simplified to

$$\sigma_0^2 \beta^2 - \beta \bar{Y} - T_n = 0, \quad T_n = n^{-1} \sum_{j=1}^{n} Y_j^2. \tag{10.18}$$

This equation is quadratic in β so may be solved using the quadratic formula, which yields two roots

$$\hat{\beta} = \frac{-\bar{Y} \pm (\bar{Y}^2 + 4\sigma_0^2 T_n)^{1/2}}{2\sigma_0^2}.$$

It turns out that the positive root is the appropriate one, as it is the one that maximizes the corresponding normal likelihood, to which solving the estimating equation corresponds in this case. Writing $\hat{\boldsymbol{\beta}}_{ML}$ to reflect the fact that the estimator is normal theory ML, we have

$$\hat{\beta}_{ML} = \frac{-\bar{Y} + (\bar{Y}^2 + 4\sigma_0^2 T_n)^{1/2}}{2\sigma_0^2},$$

thus providing an explicit expression for the quadratic estimator.

Suppose, in truth, $E(\epsilon_j^3) = \zeta_0 \neq 0$ and $\text{var}(\epsilon_j^2) = 2 + \kappa_0$. Now of course $\bar{Y} \xrightarrow{p} \beta_0$, and thus $\bar{Y}^2 \xrightarrow{p} \beta_0^2$. Moreover, using $Y_j = \beta_0 + \sigma_0 \beta_0 \epsilon_j$, we have

$$T_n = \beta_0^2 + 2\sigma_0 \beta_0^2 n^{-1} \sum_{j=1}^{n} \epsilon_j + \sigma_0^2 \beta_0^2 n^{-1} \sum_{j=1}^{n} \epsilon_j^2 \xrightarrow{p} \beta_0^2 (1 + \sigma_0^2),$$

as $n^{-1} \sum_{j=1}^{n} \epsilon_j \xrightarrow{p} 0$ and $n^{-1} \sum_{j=1}^{n} \epsilon_j^2 \xrightarrow{p} 1$. Thus, as $\hat{\beta}_{ML}$ is a continuous function of these quantities, we may conclude that

$$\hat{\beta}_{ML} \xrightarrow{p} \frac{\beta_0(1 + 2\sigma_0^2) - \beta_0}{2\sigma_0^2} = \beta_0.$$

This shows explicitly that $\hat{\beta}_{ML}$ is consistent, which, of course, is not unexpected, as $\hat{\beta}_{ML}$ solves an unbiased estimating equation.

It is also possible to show, by a Taylor series of

$$n^{1/2}(\hat{\beta}_{ML} - \beta_0) = n^{1/2}\left\{\frac{-\bar{Y} + (\bar{Y}^2 + 4\sigma_0^2 T_n)^{1/2}}{2\sigma_0^2} - \beta_0\right\}$$

about $\bar{Y} = \beta_0$ and $T_n = \beta_0^2(1 + \sigma_0^2)$ that

$$n^{1/2}(\hat{\beta}_{ML} - \beta_0) \approx (1 + \sigma_0^2)^{-1}\sigma_0\beta_0 n^{-1/2}\sum_{j=1}^{n}\{\epsilon_j + \sigma_0(\epsilon_j^2 - 1)\},$$

from which, by application of the central limit theorem, we obtain

$$n^{1/2}(\hat{\beta}_{ML} - \beta_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2\Sigma_{ML}), \quad \Sigma_{ML} = \frac{\beta_0^2\{1 + (2 + \kappa_0)\sigma_0^2 + 2\sigma_0\zeta_0}{(1 + 2\sigma_0^2)^2}.$$

We are now in a position to calculate the ARE of $\hat{\beta}_{ML}$ relative to $\hat{\beta}_{GLS}$, which is given by

$$ARE = \frac{\sigma_0^2\Sigma_{WLS}}{\sigma_0^2\Sigma_{ML}} = \frac{(1 + 2\sigma_0^2)^2}{1 + (2 + \kappa_0)\sigma_0^2 + 2\sigma_0\zeta_0}.$$

Note that the comparison thus depends on the true CV σ_0 as well as the true skewness and excess kurtosis ζ_0 and κ_0. Table 10.1 shows the values of ARE for various distributions and σ_0, ζ_0, and κ_0.

Table 10.1: *ARE of ML to GLS for the simple model under different conditions.*

True Distribution	κ_0	ζ_0	σ_0	*ARE*
Normal	0	0	0.20	1.08
	0	0	0.30	1.18
	0	0	1.00	3.00
Symmetric	2	0	0.20	1.01
($\zeta_0 = 0$)	2	0	0.30	1.02
	2	0	1.00	1.80
	4	0	0.20	0.94
	4	0	0.30	0.90
	4	0	1.00	1.29
	6	0	0.20	0.88
	6	0	0.30	0.81
	6	0	1.00	1.00
	8	0	0.20	0.83
	8	0	0.30	0.73
	8	0	1.00	0.82
Gamma	0.24	0.40	0.20	0.93
($\zeta_0 = 2\sigma_0$, $\kappa_0 = 6\sigma_0^2$)	0.54	0.60	0.30	0.88
	0.96	0.80	0.40	0.82
	6.00	2.00	1.00	0.69

IMPLICATIONS: We may make the following observations from Table 10.1.

- If the data are truly normally distributed, the quadratic estimator, which is the normal theory ML estimator, is uniformly more precise that the GLS estimator, as expected.

  Note, however, for CVs that are relatively "small" ($\sigma_0 \leq 0.30$), the gain in efficiency for ML is not substantial and decreases with decreasing CV. A CV of 1.00 is considered pretty large; in this situation, the GLS estimator is seriously relatively inefficient.

As CV reflects the "noise-to-signal" ratio in the data, we see that for "high quality" data where the "signal" dominates the "noise" (small CV), ML and GLS appear to exhibit similar performance. For "low quality" data, where the noise dominates the signal, we see that ML performs substantially better. This makes intuitive sense – as the ML estimator exploits information about β in the variance, when the variance is large (of similar magnitude to the mean), it seems likely that we would be able to gain more information about β than when the variance is of much smaller magnitude than the mean.

- If the data come from a symmetric but "heavy-tailed" distribution (relative to normality), the quadratic estimator, which assumes excess kurtosis is zero, is inefficient relative to GLS, except when the CV σ_0 gets very large. The inefficiency becomes worse as κ_0 increases. This shows, as noted previously, that there is no general ordering of the relative precision of GLS and normal theory ML in this case.

- The gamma distribution has constant CV. Recall from Chapter 4 that the linear estimator $\hat{\beta}_{GLS}$ is in fact the maximum likelihood estimator for β under the gamma distribution; hence, we would expect that GLS would be uniformly relatively more efficient when the data are truly from a gamma distribution, as seen in the table.

  In practice, it may difficult to distinguish between normal and gamma distributions (both with constant CV) if σ_0 is "small." Thus, if we mistakenly assume normality when the data really arise from a gamma distribution and use $\hat{\beta}_{ML}$ instead of $\hat{\beta}_{GLS}$, we stand to lose efficiency.

- Note further that the gamma distribution has the property that $\zeta_0 = 2\sigma_0 \rightarrow 0$ and $\kappa_0 = 6\sigma_0^2 \rightarrow 0$ as $\sigma_0 \rightarrow 0$. This demonstrates that, as the CV becomes small, the gamma and normal distributions in fact coincide in their first four moments. This explains why distinguishing between them is difficult for "small" CV. We will see more of the significance of "$\sigma_0 \rightarrow 0$" shortly.

These observations, which reflect more general phenomena, suggest that using GLS rather than normal theory ML estimation of $\boldsymbol{\beta}$ may be sensible in many circumstances. For small "noise-to-signal" in particular, it seems that possible loss of efficiency if the data are really normal is offset by possible larger gains if they are not.

10.4 Robustness of linear and quadratic equations to misspecification of the variance function

In addition to efficiency considerations, there is a compelling reason to favor GLS estimation over solving quadratic estimating equations. To illustrate, we first consider the very simple model (10.17). Suppose that we *assume* the model in (10.17),

$$E(Y_j) = \beta, \quad \text{var}(Y_j) = \sigma^2\beta^2,$$

but, *in truth*, the correct model is

$$E(Y_j) = \beta, \quad \text{var}(Y_j) = \sigma^2\beta^{2+2\theta}$$

for some $\theta > 0$. Thus, in adopting the assumed model, we have *misspecified* the variance function as $g(\beta) = \beta$ when the correct model is $g(\beta) = \beta^{1+\theta}$. Denote true values in the correct model with a "0" subscript, and suppose for simplicity we know σ_0.

Consider estimation of β under the *assumed* model.

LINEAR ESTIMATING EQUATION (GLS): To obtain the GLS estimator under our assumed model, we would again solve ($C = \infty$)

$$\sum_{j=1}^{n} \beta^{-2}(Y_j - \beta) = 0,$$

which still yields

$$\hat{\beta}_{GLS} = n^{-1}\sum_{j=1}^{n} Y_j = \bar{Y}.$$

In spite of the variance misspecification, we still have from the weak law of large numbers that $\hat{\beta}_{GLS}$ is consistent, i.e.

$$\bar{Y} \xrightarrow{p} \beta_0$$

under the true distribution of the data (under the correct model).

QUADRATIC ESTIMATING EQUATION (NORMAL ML): Again consider the normal theory ML equation. From the previous section, under the *assumed model*, we would obtain

$$\hat{\beta}_{ML} = \frac{-\bar{Y} + (\bar{Y}^2 + 4\sigma_0^2 T_n)^{1/2}}{2\sigma_0^2}.$$

However, under the *correct model*, note that $\epsilon_j = (Y_j - \beta_0)/(\sigma_0\beta_0^{1+\theta_0})$ is such that $E(\epsilon_j) = 0$, $\text{var}(\epsilon_j^2) = 1$.

Under this correct model, we still have $\bar{Y} \xrightarrow{p} \beta_0$ and $\bar{Y}^2 \xrightarrow{p} \beta_0^2$, but now, with this definition of ϵ_j,

$$T_n = n^{-1}\sum_{j=1}^{n}(\beta_0 + \sigma_0\beta_0^{1+\theta_0}\epsilon_j)^2 = \beta_0^2 + 2\sigma_0^2\beta^{2+\theta_0}n^{-1}\sum_{j=1}^{n}\epsilon_j + \sigma_0^2\beta^{2+2\theta_0}n^{-1}\sum_{j=1}^{n}\epsilon_j^2 \xrightarrow{p} \beta_0^2(1 + \sigma_0^2\beta^{2\theta_0}).$$

Combining, we find that

$$\hat{\beta}_{ML} \xrightarrow{p} \beta_0 \left\{ \frac{(1 + 4\sigma_0^2 + 4\sigma_0^4 \beta_0^{2\theta_0})^{1/2} - 1}{2\sigma_0^2} \right\}.$$

The term in large braces is equal to 1 iff $\theta_0 = 0$; that is, iff we are correct about the variance model. Otherwise, this shows that $\hat{\beta}_{ML}$ is *inconsistent*, *even if the data are truly normally distributed!*

IMPLICATION: Misspecification of the variance function g can have a profound effect on quadratic estimators for $\boldsymbol{\beta}$ in general in that the resulting estimator will be inconsistent. Our simple example demonstrates this explicitly, but that this is the case in general is easy to see from inspection of the general quadratic estimating equation (10.2):

$$\sum_{j=1}^{n} \begin{pmatrix} f_{\beta j} & 2\sigma^2 g_j^2 \nu_{\beta j} \\ \mathbf{0} & 2\sigma^2 g_j^2 \begin{pmatrix} 1/\sigma \\ \nu_{\theta j} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \sigma^2 g_j^2 & \zeta_j \sigma^3 g_j^3 \\ \zeta_j \sigma^3 g_j^3 & (2 + \kappa_j)\sigma^4 g_j^4 \end{pmatrix}^{-1} \begin{pmatrix} Y_j - f_j \\ (Y_j - f_j)^2 - \sigma^2 g_j^2 \end{pmatrix} = \mathbf{0}.$$

Assuming f is correctly specified, as we have been doing all along, note that if the variance function is misspecified, then

$$E\{(Y_j - f_j)^2 - \sigma^2 g_j^2 | \boldsymbol{x}_j\} \neq 0$$

at the true values.

Thus, if g is misspecified, the quadratic estimating equation will be *biased*. Note this is true *even if* we have correctly specified everything else. The result is that the quadratic estimator for $\boldsymbol{\beta}$ need not be *consistent*.

In contrast, as we saw in Chapter 9, misspecification of the variance function may affect efficiency of linear estimators like GLS, but it *does not* affect consistency.

Thus, we may conclude that the GLS approach yields an estimator for $\boldsymbol{\beta}$ in the general model (10.1) that is *robust* to misspecification of the variance model in that it will still be consistent even if this model is incorrectly specified. Quadratic estimators enjoy no such robustness property; a misspecified variance model may lead to inconsistent estimation.

- The *trade-off* is thus as follows. With a quadratic estimating equation, we stand potentially to gain efficiency over a linear estimating equation for estimating $\boldsymbol{\beta}$ as long as we have modeled things correctly. This potential increase comes at the expense of possible inconsistency if the variance model is not correct! With a linear estimating equation, we may lose efficiency; how much depends on the "quality" of the data. However, we are protected against possible inconsistency.
- The result is that, in practice, most analysts prefer to use the "safer" GLS approach.

This phenonmenon is now well appreciated, and, as we will see in Chapter 14, is especially well known in the multivariate case. An early reference in the case of univariate response, as considered here, is van Houwelingen (1988).

10.5 The effect of "slight" misspecification in a special case

In the previous section, we saw that misspecification of the variance function can lead to inconsistency for quadratic estimators. In deducing this result, we took the perspective that the specified form of the variance function is simply incorrect; e.g., the functional relationship between mean and variance is wrongly modeled.

In practice, the data analyst would likely investigate the nature of variance and potential variance models via diagnostic plotting techniques as in Chapter 7 and combine this evidence with possible subject-matter considerations to arrive at a variance specification. Under these realistic conditions, if a misspecification occurs, most likely the analyst will not be "flat-out" wrong about the variance model. Rather, the specified model may be "reasonable" but only "slightly" wrong, reflecting the imperfect ability to deduce underlying true features of the data generating mechanism exactly from a finite sample.

A natural question is then how such "slight" variance misspecification might affect the properties of quadratic estimators and the comparison with linear GLS estimators. Carroll and Ruppert (1982) investigated this issue in the case of a linear mean model, but extension to the nonlinear case is straightforward. In particular, these authors compared the robustness of the GLS and the quadratic normal theory ML estimators to "slight" misspecification when the true underlying distribution of the data is normal.

The argument makes clever use of the notion of *contiguity*, as we now discuss.

"SLIGHT MISSPECIFICATION": In order to characterize "slight" misspecification, the authors took a similar approach to that of investigating the properties of hypothesis testing procedures under local alternatives to the null hypothesis. Here is the set-up.

Assume that the $Y_j|\boldsymbol{x}_j$ are *in truth* normally distributed. Thus, as long as the assumptions on both mean and variance are correct, this situation would favor the quadratic normal theory ML estimator. Restricting attention to the normal distribution allows us to focus only on the effects of misspecification, easing interpretation.

As usual, let $\boldsymbol{\beta}_0$, σ_0, and $\boldsymbol{\theta}_0$ denote the values of parameters of model parameters.

Suppose that we *assume* the following model, which we will refer to as Model 1:

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \boldsymbol{\beta}), \quad \text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j) = \sigma^2 w_j^{-1}(\boldsymbol{\beta}, \boldsymbol{\theta}).$$

However, suppose that, *in truth*, $Y_j|\boldsymbol{x}_j$ follows the model

$$E(Y_j|\boldsymbol{x}_j) = f(\boldsymbol{x}_j, \boldsymbol{\beta}), \quad \text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 w_{j,n}^{-1}(\boldsymbol{\beta}, \boldsymbol{\theta}),$$

where

$$w_{j,n}(\boldsymbol{\beta}, \boldsymbol{\theta}) = w_j(\boldsymbol{\beta}, \boldsymbol{\theta})\{1 + 2Cn^{-1/2}h_j(\boldsymbol{\beta}, \boldsymbol{\theta})\}$$

for some constant C and a function h_j (which may depend on $\boldsymbol{x}_j$) satisfying

$$n^{-1}\sum_{j=1}^{n} h_j^2(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \rightarrow \gamma, \quad 0 < \gamma < \infty.$$

We will call this Model 2.

The "true" model, Model 2, thus implies, for fixed n, the specification

$$\text{var}(Y_j|\boldsymbol{x}_j) = \sigma^2 w_j^{-1}(\boldsymbol{\beta}, \boldsymbol{\theta})\{1 + 2Cn^{-1/2}h_j(\boldsymbol{\beta}, \boldsymbol{\theta})\}^{-1}.$$

Thus, in the *assumed* model, Model 1, we are "off" in terms of specifying the variance function by a factor depending on $n^{-1/2}$.

- As $n \rightarrow \infty$, Model 1 becomes identical to the true Model 2, but, for fixed n, the models differ. Model 1 represents what we believe by our assumption, and Model 2 represents departures from the assumption, similar to a null hypothesis and alternatives to it.
- The idea is to try and characterize the realistic practical situation where the analyst bases the assumed model on the available data. As more and more data become available ($n \rightarrow \infty$), the analyst has more information with which to deduce a variance model and so becomes "better at" this modeling. The analyst will specify a fixed model that does not depend on n, of course, as in Model 1. Thus, Model 2, although obviously a technical device, has the practical interpretation as deviating slightly from this fixed, assumed model.

IDEA: The idea of the argument is to investigate the sensitivity of the GLS and normal theory ML estimators computed assuming Model 1 to the fact that, in truth, Model 2 holds. The approach is to compare the limiting distributions of $\hat{\boldsymbol{\beta}}_{ML}$ and $\hat{\boldsymbol{\beta}}_{GLS}$ computed under Model 1 when Model 2 holds.

- As the misspecification of the variance model is "slight," the hope would be that an estimator that is *robust* to "slight" misspecification would have the *same* large sample properties under both Models 1 and 2. That is, the properties would be insensitive to whether or not, in truth, Model 1 or Model 2 holds.

Rather than attack this problem directly, which is certainly possible, Carroll and Ruppert (1982) made clever use of the notion of *contiguity*. In particular, as mentioned above, it is standard to investigate the properties of hypothesis testing procedures under both the null hypothesis (assumed Model 1 here) and local alternatives (true Model 2 here), where local alternatives differ from the null by a term depending on $n^{-1/2}$. It turns out that the notion of contiguity provides a convenient theoretical device for studying this that may be exploited for our problem.

A formal treatment of the theory of contiguity we will use is given by Hájek and Sidák (1967); see also van der Vaart (1998, Section 6.2). Here, we will just state the definitions and theorems necessary for our argument.

CONTIGUITY: For each n, let p_n and q_n be probability densities. The sequence of densities $\{q_n\}$ is said to be *contiguous to* $\{p_n\}$ if, for any sequence of events A_n,

$$P_{p_n}(A_n) \rightarrow 0 \text{ as } n \rightarrow \infty$$

implies $P_{q_n}(A_n) \rightarrow 0$, where P_p denotes probability measure under the density p.

Thus, if p_n and q_n are densities under a null and alternative hypothesis, respectively, and if A_n is the critical region for a test, this says that if Type I error $= P_{p_n}(A_n) \rightarrow 0$ then power $= P_{q_n}(A_n) \rightarrow 0$.

We will use this by identifying p_n as the (normal) density of the data under Model 1 and q_n as that under Model 2.

It turns out that if one can establish that two densities corresponding to two different models are contiguous, then it is straightforward to derive limiting distribution results under the contiguous model (Model 2 here) from those that are more easily obtained assuming Model 1. A series of results due to LeCam describe how this may be accomplished; see also van der Vaart (1998, Section 7.5)..

COROLLARY TO LECAM'S FIRST LEMMA: Let $L_n = q_n/p_n$ be the likelihood ratio. If, under p_n,

$$\log L_n \xrightarrow{L} \mathcal{N}(\mu, \tau^2), \quad \mu = -\tau^2/2,$$

then q_n is contiguous to p_n.

LECAM'S THIRD LEMMA: Let S_n be any statistic. Suppose that, *under* p_n,

$$\begin{pmatrix} S_n \\ \log L_n \end{pmatrix} \xrightarrow{L} \mathcal{N}\left\{ \begin{pmatrix} \mu_1 \\ -\tau_2^2/2 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} \right\},$$

so that q_n is contiguous to p_n (by the corollary above). Then, *under* q_n,

$$S_n \xrightarrow{L} \mathcal{N}(\mu_1 + \tau_{12}, \tau_1^2).$$

Although this is stated for a univariate statistic S_n, the result extends readily to multivariate $\boldsymbol{S}_n$, say.

We may now exploit these results. If we can derive the joint distribution of $\boldsymbol{S}_n = n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\log L_n$ assuming Model 1 holds for each of $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{ML}$ and $\hat{\boldsymbol{\beta}}_{GLS}$, using LeCam's third lemma, we can immediately deduce the behavior of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ for Model 2. This of course requires showing contiguity.

Identifying p_n as the likelihood under Model 1 and q_n as that under Model 2, the loglikelihoods have the form

$$\log p_n = -n \log 2\pi - n \log \sigma + (1/2) \sum_{j=1}^{n} \log w_j(\boldsymbol{\beta}, \boldsymbol{\theta}) - (1/2) \sum_{j=1}^{n} w_j(\boldsymbol{\beta}, \boldsymbol{\theta}) \{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta}\}^2 / \sigma^2,$$

$$\begin{aligned} \log q_n = -n \log 2\pi - n \log \sigma + (1/2) \sum_{j=1}^{n} \log w_j(\boldsymbol{\beta}, \boldsymbol{\theta}) + (1/2) \sum_{j=1}^{n} \log\{1 + 2Cn^{-1/2} h_j(\boldsymbol{\beta}, \boldsymbol{\theta})\} \\ -(1/2) \sum_{j=1}^{n} w_j(\boldsymbol{\beta}, \boldsymbol{\theta}) \{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta}\}^2 \{1 + 2Cn^{-1/2} h_j(\boldsymbol{\beta}, \boldsymbol{\theta})\} / \sigma^2. \end{aligned}$$

First, suppose that Model 1 holds. Then $\epsilon_j = w_j(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\{Y_j - f(\boldsymbol{x}_j, \boldsymbol{\beta}_0)\}^2 / \sigma_0 \sim \mathcal{N}(0, 1)$, and, evaluated at the true values for the model, we have

$$\log L_n = \log q_n - \log p_n = (1/2) \sum_{j=1}^{n} \log\{1 + 2Cn^{-1/2} h_j(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)\} - Cn^{-1/2} \sum_{j=1}^{n} h_j(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \epsilon_j^2.$$

Using $\log(1 + x) \approx x - x^2/2$ for all x, we may approximate this as

$$\log L_n \approx -Cn^{-1/2} \sum_{j=1}^{n} h_j(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)(\epsilon_j^2 - 1) - C^2 n^{-1} \sum_{j=1}^{n} h_j^2(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0). \tag{10.19}$$

Assuming Model 1 holds, $\text{var}(\epsilon_j^2 | \boldsymbol{x}_j) = \text{var}(\epsilon_j^2) = 2$. Thus, using Slutsky's theorem and the central limit theorem applied to (10.19), we have under p_n that

$$\log L_n \xrightarrow{L} \mathcal{N}(-C^2\gamma, 2C^2\gamma).$$

Thus, by the corollary, Model 2 is contiguous to Model 1. This result explains why the authors set up Model 2 as they did.

In fact, it is straightforward to construct the joint, large sample distributions of $n^{1/2}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0)$ and $n^{1/2}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_0)$ and $\log L_n$ under Model 1.

From the folklore theory in Chapter 9, we have

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_0) \approx n^{-1/2} \sum_{j=1}^{n} \boldsymbol{d}_j \epsilon_j \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Sigma}_{WLS}),$$

where $\boldsymbol{d}_j$ may be identified as $\sigma_0 \boldsymbol{\Sigma}_{WLS} g_{0j}^{-1} f_{\beta 0j}$. Similarly, from Section 10.2, we have

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0) \approx n^{-1/2} \sum_{j=1}^{n} \{\boldsymbol{b}_{1j} \epsilon_j + \boldsymbol{b}_{2j}(\epsilon_j^2 - 1)\} \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Sigma}_{ML}),$$

where the forms of $\boldsymbol{b}_{1j}$ and $\boldsymbol{b}_{2j}$ may be deduced.

We may now use the third lemma. Taking $\boldsymbol{S}_n = n^{1/2}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_0)$, we obtain, writing $h_j(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = h_{0j}$,

$$\begin{pmatrix} n^{1/2}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_0) \\ \log L_n \end{pmatrix} \approx n^{-1/2} \sum_{j=1}^{n} \begin{pmatrix} \boldsymbol{d}_j \epsilon_j \\ -C h_{0j}(\epsilon_j^2 - 1) \end{pmatrix} + n^{-1} \sum_{j=1}^{n} \begin{pmatrix} \mathbf{0} \\ -C^2 h_{0j}^2 \end{pmatrix}.$$

Using the central limit theorem, Slutsky's theorem, and the fact that $E(\epsilon_j^3 | \boldsymbol{x}_j) = E(\epsilon_j^3) = 0$ under normality, we obtain that, under p_n,

$$\begin{pmatrix} n^{1/2}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_0) \\ \log L_n \end{pmatrix} \xrightarrow{L} \mathcal{N}\left\{ \begin{pmatrix} \mathbf{0} \\ -C^2\gamma \end{pmatrix}, \begin{pmatrix} \sigma_0^2 \boldsymbol{\Sigma}_{WLS} & \mathbf{0} \\ \mathbf{0} & 2C^2\gamma \end{pmatrix} \right\}.$$

Thus, by LeCam's third lemma, we may immediately conclude that, under Model 2 (q_n),

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Sigma}_{WLS}). \tag{10.20}$$

RESULT FOR GLS: Under the "slight" misspecification of Model 1, $\hat{\boldsymbol{\beta}}_{GLS}$ has the *same* asymptotic normal distribution as under the *correct* Model 2. Hence, GLS is *robust* to "slight" misspecification of the variance function, as defined above.

We now apply the lemma with $\boldsymbol{S}_n = n^{1/2}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0)$. We have

$$\begin{pmatrix} n^{1/2}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0) \\ \log L_n \end{pmatrix} \approx n^{-1/2} \sum_{j=1}^{n} \begin{pmatrix} \{\boldsymbol{b}_{1j} \epsilon_j + \boldsymbol{b}_{2j}(\epsilon_j^2 - 1)\} \\ -C h_{0j}(\epsilon_j^2 - 1) \end{pmatrix} + n^{-1} \sum_{j=1}^{n} \begin{pmatrix} \mathbf{0} \\ -C^2 h_{0j}^2 \end{pmatrix}.$$

Now under Model 1 (p_n), the covariance matrix of a summand in the first term on the right hand side does not have off-diagonal elements equal to zero, as was the case for GLS above, because $E\{(\epsilon_j^2 - 1)\}^2 = \text{var}(\epsilon_j^2) = 2$.

Thus, letting $\boldsymbol{q} = \lim_{n\to\infty} n^{-1}\sum_{j=1}^{n} h_{0j}\boldsymbol{b}_{2j}$, we obtain under Model 1 (p_n)

$$\begin{pmatrix} n^{1/2}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0) \\ \log L_n \end{pmatrix} \xrightarrow{L} \mathcal{N}\left\{ \begin{pmatrix} \mathbf{0} \\ -C^2\gamma \end{pmatrix}, \begin{pmatrix} \sigma_0^2\boldsymbol{\Sigma}_{ML} & -2C\boldsymbol{q} \\ -2C\boldsymbol{q} & 2C^2\gamma \end{pmatrix} \right\}.$$

Thus, by LeCam's third lemma, we conclude that, under Model 2 (q_n),

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0) \xrightarrow{L} \mathcal{N}(-2C\boldsymbol{q}, \sigma_0^2\boldsymbol{\Sigma}_{ML}). \tag{10.21}$$

RESULT FOR ML: Under the "slight" misspecification of Model 1, $\hat{\boldsymbol{\beta}}_{ML}$ no longer has the same asymptotic normal distribution as under the *correct* Model 2. In particular, (10.21) implies that, if we compute $\hat{\boldsymbol{\beta}}_{ML}$ assuming Model 1 is true, but Model 2 really holds, then, for large n,

$$\hat{\boldsymbol{\beta}}_{ML} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}_0 - 2n^{-1/2}C\boldsymbol{q}, n^{-1}\sigma_0^2\boldsymbol{\Sigma}_{ML}).$$

Thus, the approximate (normal) sampling distribution of $\hat{\boldsymbol{\beta}}_{ML}$ is not centered about $\boldsymbol{\beta}_0$.

IMPLICATION FOR PRACTICE: If we were to proceed naively assuming Model 1 is true and construct tests and confidence intervals for the true value $\boldsymbol{\beta}_0$ using the result under Model 1 that $\hat{\boldsymbol{\beta}}_{ML} \dot{\sim} \mathcal{N}(\boldsymbol{\beta}_0, n^{-1}\sigma_0^2\boldsymbol{\Sigma}_{ML})$, it is clear that there is a potential for erroneous inference. How erroneous would depend on the relevance of the theory for finite n and the severity of the misspecification, as measured by C and $\boldsymbol{q}$.

On the other hand, from (10.20), if we were to base such inference on $\boldsymbol{\beta}_0$ on $\hat{\boldsymbol{\beta}}_{GLS}$, even if we "slightly" misspecified the variance, we would expect the conclusions to be unaffected.

ARE of $\hat{\boldsymbol{\beta}}_{ML}$ relative to $\hat{\boldsymbol{\beta}}_{GLS}$: Because the mean of the large sample normal distribution of $n^{1/2}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0)$ is not zero, the usual measure of relative efficiency is not appropriate, as it does not take this feature into account. A standard practice in such situations is to instead base the comparison on the ratio of *mean square error* (MSE) for each estimator, where the generic definition is

$$\text{MSE} = \text{variance} + \text{bias}^2.$$

In the case of multivariate $\boldsymbol{\beta}$, one can consider the MSE corresponding to estimating a linear combination $\boldsymbol{\lambda}^T\boldsymbol{\beta}_0$ for an $\boldsymbol{\lambda}$ (so this includes estimating each element separately). From (10.20) and (10.21), we have under Model 2 that

$$nMSE(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_{GLS}) \to \sigma_0^2\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{WLS}\boldsymbol{\lambda}, \quad nMSE(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_{ML}) \to \sigma_0^2\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{ML}\boldsymbol{\lambda} + 4C^2\boldsymbol{\lambda}^T\boldsymbol{q}.$$

Now, under normality, we know that $\boldsymbol{\Sigma}_{ML}$ "$\leq$" $\boldsymbol{\Sigma}_{WLS}$ (i.e., $\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{ML}\boldsymbol{\lambda} \leq \boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{WLS}\boldsymbol{\lambda}$).

However, the MSE for $\hat{\boldsymbol{\beta}}_{ML}$ involves an additional "bias" term depending on C and $\boldsymbol{q}$. If $\boldsymbol{\lambda}$ is not orthogonal to $\boldsymbol{q}$, which seems unlikely in general, and if C and $\boldsymbol{q}$ are "large", then it could well be that the limiting $nMSE(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_{GLS}) \leq nMSE(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}_{ML})$, offsetting any advantage for the ML estimator.

REMARK: The above argument was in the context of the particular quadratic estimator for $\boldsymbol{\beta}$ corresponding to normal ML. However, all of the results we have described in that context have general implications about the sensitivity of quadratic estimators to misspecification of the variance model. The bottom line is that quadratic estimators are not robust to assumptions about the model and about the distribution (or at least about third and fourth moments). In contrast, the GLS estimator is robust to these issues. Consequently, it is no surprise that many authors warn about the use of quadratic estimators unless the analyst has a great deal of confidence in the model and assumptions, and they advocate GLS, which is also easier to compute, for routine use.

10.6 "Small σ"

We conclude this chapter by mentioning briefly a technical device that we will exploit heavily in Chapter 12.

In Section 10.3, we discussed the notion of "high quality" data with "low" "noise-to-signal." In many applications, such data arise; for example, in pharmacokinetics, the range of the mean response is often "large" relative to the magnitude of the variation over the whole range of the response. This phenomenon can be seen in Figure 1.1 for the indomethacin data. Thus, even if variance changes over the range of the mean response, it is "small" relative to the size of the mean, so that the "signal" dominates.

In the general mean-variance model (10.1), it is common for the variance to depend on $\boldsymbol{\beta}$ through some function of the mean. In fact, a routine feature that we have seen in several examples is that the variance increases as a function of the mean. For (10.1) in this situation, then, the scale parameter σ governs the magnitude of the variance relative to the range of the values taken on by the mean function. In the particular case of the constant CV model $g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{x}_j) = f(\boldsymbol{x}_j, \boldsymbol{\beta})$, σ is equal to the CV, and so represents exactly the "noise-to-signal." But even for other functions g depending on the mean, the interpretation of σ is thus similar.

A common technical device for representing this situation in (10.1) is to let not only $n \to \infty$, but also to let $\sigma \to 0$. It turns out that this device simplifies arguments, as we will see in Chapter 12, but it also has practical relevance.

It turns out that, under $\sigma \to 0$, the possible advantages of quadratic estimating equations, as well as some of their drawbacks, vanish. For definiteness, consider (10.10) in our study of the quadratic equations in Section 10.2. Recall that we rescaled the problem to be in terms of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)/\sigma_0$ and $(\hat{\sigma} - \sigma_0)/\sigma_0$ to arrive at (10.10). This was no accident, as doing so ensures no problems in letting $\sigma_0 \to 0$.

Under this condition, (10.10) reduces to

$$n^{-1} \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & \frac{4}{2+\kappa}\boldsymbol{Q}^T\boldsymbol{Q} \end{pmatrix} n^{1/2} \begin{pmatrix} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)/\sigma_0 \\ (\hat{\sigma} - \sigma_0)/\sigma_0 \\ \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \end{pmatrix}$$
$$\approx n^{-1/2} \sum_{j=1}^{n} \begin{pmatrix} g_{0j}^{-1} f_{\beta 0j} \epsilon_j \\ 2(2+\kappa)^{-1}(\epsilon_j^2 - 1)\tau_{\theta 0j} \end{pmatrix}.$$

It should be clear that this implies that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)/\sigma_0 \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{WLS})$. That is, the quadratic contribution to the estimating equation and the connection between $\hat{\boldsymbol{\beta}}$ and the variance parameter estimators is eliminated.

In fact, under this condition, the quadratic estimator has exactly the same large sample distribution as $\hat{\boldsymbol{\beta}}_{GLS}$. More formally, if $n \to \infty$ and $\sigma_0 \to 0$, then the quadratic estimator and GLS are *asymptotically equivalent*. The advantage of the additional information on $\boldsymbol{\beta}$ contained in ν_β is lost.

Recall we mentioned in Section 10.3 that the gamma and normal distributions coincide in their first four moments when $\sigma_0 \to 0$. Recall also that, for the gamma distribution, a member of the scaled exponential family class, the maximum likelihood estimator is GLS. We see that, as $\sigma_0 \to 0$, the maximum likelihood estimators for $\boldsymbol{\beta}$ under the gamma and normality coincide, as intuition would suggest.