

9 The “folklore” theorem and optimality of GLS

9.1 Introduction

In Section 8.3, we used standard M-estimator arguments to examine the large sample properties of estimators for β in the situation where we specify a set of constant weights to represent potentially nonconstant variance and estimate β by WLS; i.e., a linear estimating equation. This of course includes specifying weights all equal to one, so that the estimation procedure is that of OLS. The main insights were that

- Even if the weights are chosen incorrectly, the estimators will still be consistent. So, for example, using OLS when variance is nonconstant yields a consistent estimator for the true value of β .
- If the weights are correctly specified, i.e., chosen so that they equal w_j , where in truth $\text{var}(Y_j|\mathbf{x}_j) = \sigma^2 w_j^{-1}$, then the resulting estimator for β is more efficient than others using incorrect weights. Thus, it is the “optimal” choice among all such estimators solving linear estimating equations with some specified, fixed weights.

In fact, it was notable that we did not need to make any assumptions on the distribution of $Y_j|\mathbf{x}_j$ beyond that of the first two true moments (and those necessary to invoke the central limit theorem).

Of course, it would be very unusual in practice to be able to specify a set of known weights, as we have noted previously. However, as we have discussed at length, it is often feasible to adopt a variance *model* that provides a good representation of the form of the variance, e.g., as a function of the level of mean response.

In this Chapter, we will consider the model

$$E(Y_j|\mathbf{x}_j) = f(\mathbf{x}_j, \beta), \quad \text{var}(Y_j|\mathbf{x}_j) = \sigma^2 g^2(\beta, \theta, \mathbf{x}_j), \quad (9.1)$$

and investigate the same issues as in Section 8.4, focusing on estimation of β via solution of *linear estimating equations*. In particular, we will consider estimation of β by the GLS approach, where, in fact, we might have to also estimate unknown variance parameters θ . The GLS scheme may be viewed as “WLS with estimated weights,” so it is natural to wonder if the results in the case of “fixed weights” in Section 8.4 extend somehow to the “estimated weights” situation.

To be more specific, we will investigate the following:

- Suppose we have *correctly specified* the form of the variance function g , and we estimate β using the three-step GLS algorithm, possibly estimating additional variance parameters θ at step (ii). What are the properties of the resulting estimator $\hat{\beta}_{GLS}$?
- Suppose we specify a model of the form (9.1) with the correct model f for the conditional mean, but we *misspecify* the variance function, postulating instead a model that does not accurately characterize the true variance? What are the properties of the GLS estimator for β ? How do they compare to those of the GLS estimator where the variance function is correctly specified?

Throughout this Chapter, we will assume that the mean model in (9.1) is correctly specified; our focus will be on the variance function.

Once we have addressed these issues, we will move to practical considerations of how to use the results to obtain approximate standard errors for the elements of the estimator for β , how to construct confidence intervals and hypothesis tests about the true value β_0 , and so on.

Finally, in Section 9.6, we will address another motivation for the GLS approach: its role as the “optimal” (in a large sample sense) estimator within the class of linear estimating equation estimators (when the variance function is correctly specified). This development will in fact provide insight into “optimality” of more general estimation schemes.

9.2 The “folklore” theorem of GLS

The result we refer to as the “folklore” theorem establishes the asymptotic normality of the GLS estimator of β in the model (9.1) when the form of the variance function $g(\beta, \theta, \mathbf{x}_j)$ is *correctly specified*; that is, the functional form describes accurately the form of the variance.

WHY THE CUTE NAME? The result we are about to demonstrate has been known for decades, shown by rigorous proof or deduced by informal arguments. Econometricians were among the first to realize the result in the 1970s, and biostatisticians determined it again in the 1980s. It was “discovered” over and over again to the point that, now, it is so well known that it has achieved the status of “statistical folklore.” See Carroll and Ruppert (1988, Sections 2.2 and 7.3.1) for more.

We will state and demonstrate the result in the case where θ is unknown and estimated by a “well behaved” estimator $\hat{\theta}$. We will in fact show in Chapter 12 that the estimators for θ we discussed in Chapter 6 (based on transformations of absolute residuals) are all well behaved under reasonable conditions in this sense.

The result for the case where $\boldsymbol{\theta}$ is known will be a simpler, special case of the general result. We will use the same basic M-estimator argument as in Section 8.2, although we will derive it explicitly so that some interesting features may be highlighted.

Consider the three-step GLS algorithm outlined in Section 6.3. Suppose the algorithm is carried out for C iterations (recall that $C = \infty$ means iterate to “convergence”). We are interested in the properties of the GLS estimator for $\boldsymbol{\beta}$ obtained from step (iii) after C iterations, where $\boldsymbol{\beta}$ is estimated initially at step (i) by, for example, OLS, ignoring nonconstant variance, and $\boldsymbol{\theta}$ is estimated at step (ii) based on residuals and predicted values from the current estimate of $\boldsymbol{\beta}$. For simplicity, we will refer to the GLS estimator for any C at the end of step (iii) as $\hat{\boldsymbol{\beta}}$, suppressing the subscript “GLS.”

RECALL: As mentioned in Section 8.4, we will carry out arguments conditional on the $\mathbf{x}_1, \dots, \mathbf{x}_n$ and make use of the assumption that the conditional distribution of $Y_j | \mathbf{x}_1, \dots, \mathbf{x}_n$ is the same as that of $Y_j | \mathbf{x}_j$.

CONSISTENCY: To fix ideas, suppose that $\boldsymbol{\beta}$ is estimated initially by OLS at step (i) and that $\boldsymbol{\theta}$ is estimated by PL at step (ii). To check consistency of the GLS estimator, it makes sense to consider the entire process of estimating $\boldsymbol{\beta}$ at step (iii) along with estimation at the previous steps. Define $\boldsymbol{\beta}_0$, σ_0 , and $\boldsymbol{\theta}_0$ to be the true values of $\boldsymbol{\beta}$ ($p \times 1$), σ , and $\boldsymbol{\theta}$ ($q \times 1$) in (9.1) generating the data. (Recall that we are assuming that the models for mean and variance are both correct.)

In Section 8.3, we argued that the OLS estimator is consistent even when the true variance is nonconstant. Although in (9.1) we do not consider “fixed weights,” from a theoretical standpoint, we may still think of the “true weights” $w_j = g^{-2}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0, \mathbf{x}_j)$; conditional on \mathbf{x}_j , the value of g evaluated at the true parameter values and \mathbf{x}_j may be regarded as a constant. The OLS estimator solves

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

clearly, under (9.1), $E\{Y_j - f(\mathbf{x}_j, \boldsymbol{\beta}_0) | \mathbf{x}_j\} = 0$, so that it is still reasonable to assume that $\hat{\boldsymbol{\beta}}_{OLS}$ is consistent.

Suppose that we are at the beginning of iteration C , and $\hat{\boldsymbol{\beta}}^*$ is the current estimator for $\boldsymbol{\beta}$. Then we would solve

$$\sum_{j=1}^n \left[\frac{\{Y_j - f(\mathbf{x}_j, \hat{\boldsymbol{\beta}}^*)\}^2}{\hat{\sigma}^2 g^2(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\theta}}, \mathbf{x}_j)} - 1 \right] \tau_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\theta}}, \mathbf{x}_j) = \mathbf{0} \quad (9.2)$$

to obtain the updated estimator $(\hat{\sigma}, \hat{\boldsymbol{\theta}}^T)^T$ at step (ii) and then solve for $\hat{\boldsymbol{\beta}}$ at step (iii)

$$\sum_{j=1}^n g^{-2}(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\theta}}, \mathbf{x}_j) \{Y_j - f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})\} f_{\boldsymbol{\beta}}(\mathbf{x}_j, \hat{\boldsymbol{\beta}}) = \mathbf{0}. \quad (9.3)$$

$\hat{\beta}$ is the resulting estimator for iteration C . We could modify (9.3), replacing the $\hat{\beta}^*$ in the “weights” by $\hat{\beta}$ (so that solving (9.3) would be carried out by IRWLS for the fixed value $\hat{\theta}$), but as we will see shortly, this will not matter. In the case $C = \infty$, then $\hat{\beta}^* = \hat{\beta}$ in both (9.2) and (9.3).

Thus, to check consistency of $\hat{\beta}$, we need to consider the entire set of equations, as $\hat{\beta}^*$ and $\hat{\theta}$ are involved in the weights (and solve their own equations). As an example, suppose $C = 1$, and $\hat{\beta}^*$ is the OLS estimator. Then, to obtain $\hat{\beta}$, we are really solving the entire system of equations given by

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

(9.2), and (9.3). For this entire system of $(2p + q + 1)$ equations, the parameter being estimated is really $(\beta^T, \sigma, \theta^T, \beta^T)^T$. Note that all three equations are clearly *unbiased*. Thus, it is reasonable to be assured that the entire process leads to consistent estimators for the true values all elements of this extended parameter vector, and hence of the last p elements solved by the GLS equation (9.3).

This perspective would of course extend to any C , where $\hat{\beta}^* =$ previous GLS estimator. In the case $C = \infty$, then $\hat{\beta}^* = \hat{\beta}$ in both (9.2) and (9.3).

Based on these observations, we are willing to assume that the GLS estimator for any C is consistent, as is the previous estimator for β and those for the variance parameters σ and θ .

We now focus on the estimating equation (9.3) and apply the M-estimator argument to deduce the large sample distribution of $n^{1/2}(\hat{\beta} - \beta_0)$. As in previous chapters, it will be again convenient to define

$$\epsilon_j = \frac{Y_j - f(\mathbf{x}_j, \beta_0)}{\sigma_0 g(\beta_0, \theta_0, \mathbf{x}_j)},$$

so that $E(\epsilon_j | \mathbf{x}_j) = 0$ and $\text{var}(\epsilon_j | \mathbf{x}_j) = 1$.

THE “FOLKLORE” THEOREM: Assume that the model (9.1) is correctly specified and that $\hat{\beta}^*$ and $\hat{\theta}$ are preliminary estimators for β and θ such that

$$n^{1/2}(\hat{\beta}^* - \beta_0) = O_p(1), \quad n^{1/2}(\hat{\theta} - \theta_0) = O_p(1). \quad (9.4)$$

Under suitable regularity conditions, all GLS estimators $\hat{\beta}$ for any choice of C , including $C = \infty$, satisfy

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \Sigma_{WLS}), \quad (9.5)$$

where

$$\Sigma_{WLS}^{-1} = \lim_{n \rightarrow \infty} n^{-1} \sum_{j=1}^n w_j f_{\beta}(\mathbf{x}_j, \beta_0) f_{\beta}(\mathbf{x}_j, \beta_0) = \lim_{n \rightarrow \infty} n^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X},$$

$w_j = g^{-2}(\beta_0, \theta_0, \mathbf{x}_j)$, $\mathbf{W} = \text{diag}(w_1, \dots, w_n)$, and $\mathbf{X} = \mathbf{X}(\beta_0) = \{f_{\beta}(\mathbf{x}_1, \beta_0), \dots, f_{\beta}(\mathbf{x}_n, \beta_0)\}^T$.

Before we carry out the argument, we make a few remarks:

- The condition in (9.4) says that these quantities do not “blow up” but are “well behaved” for large n .

As we discussed on page 195, such a result follows if the estimator is asymptotically normal with mean equal to the true value (in this case). Thus, we may regard this condition as simply stating that $\hat{\beta}^*$ and $\hat{\theta}$ are “usual” M-estimators. (Actually, any estimators $\hat{\beta}^*$ and $\hat{\theta}$ satisfying (9.4) are sufficient to demonstrate the theorem.) Note that if $\hat{\beta}^*$ is the OLS estimator, for example, the condition is satisfied, as we showed the asymptotic normality of $\hat{\beta}_{OLS}$ in the last chapter.

- Comparing (9.5) to (8.21), we see that the large sample distribution of the GLS estimator has the same form as for WLS in the case where the weights w_j are known and correctly specified! As we noted above, conditional on the \mathbf{x}_j , the values w_j are just a set of constants. Thus, (9.5) seems to say that, whether we actually know the constants w_j themselves or simply know the function $g(\beta, \theta, \mathbf{x}_j)$, we end up with the *same* large sample distribution. We will elaborate on this after we complete the argument.

We now carry out the argument. Consider the GLS estimating equation. By the usual Taylor series expansion and under suitable conditions, we have

$$\begin{aligned}
\mathbf{0} &= n^{-1/2} \sum_{j=1}^n g^{-2}(\hat{\beta}^*, \hat{\theta}, \mathbf{x}_j) \{Y_j - f(\mathbf{x}_j, \hat{\beta})\} f_{\beta}(\mathbf{x}_j, \hat{\beta}) \\
&\approx n^{-1/2} \sum_{j=1}^n g^{-2}(\beta_0, \theta_0, \mathbf{x}_j) \{Y_j - f(\mathbf{x}_j, \beta_0)\} f_{\beta}(\mathbf{x}_j, \beta_0) \\
&\quad + \left[n^{-1} \sum_{j=1}^n \{Y_j - f(\mathbf{x}_j, \beta_0)\} f_{\beta\beta}(\mathbf{x}_j, \beta_0) \right. \\
&\quad \quad \left. - n^{-1} \sum_{j=1}^n g^{-2}(\beta_0, \theta_0, \mathbf{x}_j) f_{\beta}(\mathbf{x}_j, \beta_0) f_{\beta}^T(\mathbf{x}_j, \beta_0) \right] n^{1/2}(\hat{\beta} - \beta_0) \\
&\quad + \left[-2n^{-1} \sum_{j=1}^n g^{-3}(\beta_0, \theta_0, \mathbf{x}_j) \{Y_j - f(\mathbf{x}_j, \beta_0)\} f_{\beta}(\mathbf{x}_j, \beta_0) g_{\beta}^T(\beta_0, \theta_0, \mathbf{x}_j) \right] n^{1/2}(\hat{\beta}^* - \beta_0) \\
&\quad + \left[-2n^{-1} \sum_{j=1}^n g^{-3}(\beta_0, \theta_0, \mathbf{x}_j) \{Y_j - f(\mathbf{x}_j, \beta_0)\} f_{\beta}(\mathbf{x}_j, \beta_0) g_{\theta}^T(\beta_0, \theta_0, \mathbf{x}_j) \right] n^{1/2}(\hat{\theta} - \theta_0)
\end{aligned}$$

Note that what we have done is expand $(\hat{\beta}^T, \hat{\beta}^{*T}, \hat{\theta}^T)^T$ about $(\beta_0^T, \beta_0^T, \theta_0^T)^T$, although we have elected to write the linear term in the expansion as three separate pieces rather than keeping it “stacked” for reasons that shall become obvious momentarily.

Using the definitions of ϵ_j , ν_{β} , and ν_{θ} given previously, we may write this succinctly as

$$\mathbf{0} \approx \mathbf{C}_n + (\mathbf{A}_{n1} + \mathbf{A}_{n2})n^{1/2}(\hat{\beta} - \beta_0) + \mathbf{D}_n n^{1/2}(\hat{\beta}^* - \beta_0) + \mathbf{E}_n n^{1/2}(\hat{\theta} - \theta_0). \quad (9.6)$$

Here,

$$\begin{aligned} \mathbf{C}_n &= \sigma_0 n^{-1/2} \sum_{j=1}^n w_j^{1/2} f_\beta(\mathbf{x}_j, \boldsymbol{\beta}_0) \epsilon_j, \\ \mathbf{A}_{n1} &= \sigma_0 n^{-1} \sum_{j=1}^n w_j^{1/2} f_{\beta\beta}(\mathbf{x}_j, \boldsymbol{\beta}_0) \epsilon_j, \quad \mathbf{A}_{n2} = -n^{-1} \sum_{j=1}^n w_j f_\beta(\mathbf{x}_j, \boldsymbol{\beta}_0) f_\beta^T(\mathbf{x}_j, \boldsymbol{\beta}_0), \\ \mathbf{D}_n &= -2\sigma_0 n^{-1} \sum_{j=1}^n w_j f_\beta(\mathbf{x}_j, \boldsymbol{\beta}_0) \nu_\beta^T(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0, \mathbf{x}_j) \epsilon_j, \\ \mathbf{E}_n &= -2\sigma_0 n^{-1} \sum_{j=1}^n w_j f_\beta(\mathbf{x}_j, \boldsymbol{\beta}_0) \nu_\theta^T(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0, \mathbf{x}_j) \epsilon_j. \end{aligned}$$

We now deduce the behavior of each of these terms. As $E(\epsilon_j | \mathbf{x}_j) = 0$, clearly, by the weak law of large numbers, all of \mathbf{A}_{n1} , \mathbf{D}_n , and \mathbf{E}_n converge in probability to zero; we will discuss the consequences of this shortly. From the latter two results, we may rewrite (9.6) as

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approx -\mathbf{A}_n^{-1} \mathbf{C}_n.$$

Now, conditional on the \mathbf{x}_j , the term \mathbf{A}_{n2} depends only on constants. By the assumptions of the theorem, this term satisfies

$$\mathbf{A}_{n2} \rightarrow -\boldsymbol{\Sigma}_{WLS}^{-1}. \quad (9.7)$$

Now, combining the results for \mathbf{A}_{n1} and \mathbf{A}_{n2} , we have

$$\mathbf{A}_n \xrightarrow{p} -\boldsymbol{\Sigma}_{WLS}^{-1}.$$

Moreover, we may apply the multivariate central limit theorem to \mathbf{C}_n . Clearly, $E\{w_j^{1/2} f_\beta(\mathbf{x}_j, \boldsymbol{\beta}_0) \epsilon_j | \mathbf{x}_j\} = \mathbf{0}$ and

$$\text{var}\{w_j^{1/2} f_\beta(\mathbf{x}_j, \boldsymbol{\beta}_0) \epsilon_j | \mathbf{x}_j\} = w_j f_\beta(\mathbf{x}_j, \boldsymbol{\beta}_0) f_\beta^T(\mathbf{x}_j, \boldsymbol{\beta}_0)$$

(recall that w_j is a function of \mathbf{x}_j). Thus,

$$\mathbf{C}_n \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Sigma}_{WLS}^{-1}).$$

Applying these results to (9.7) using Slutsky's theorem, we thus may conclude that

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

as claimed in (9.5).

REMARKS: Carrying out the argument explicitly rather than just “plugging in” to the generic M-estimator calculations allows us to make several important observations.

- As we noted previously, the result here is identical to that in the case of WLS with the weights w_j regarded as known constants that have been correctly specified. The implication is that, as long as we specify the functional form of the weights through the variance function g correctly and estimate the unknown β and θ that fully characterize them (conditional on \mathbf{x}_j), we can estimate β_0 using GLS *as well as if we in fact knew the weights* $w_j = g^{-2}(\beta_0, \theta_0, \mathbf{x}_j)$!
- That is, having to estimate the weights (by substituting the current estimate $\hat{\beta}^*$ and $\hat{\theta}$) rather than knowing them exactly *no penalty* in the sense of (large sample) precision.
- In the argument, the terms \mathbf{D}_n and \mathbf{E}_n corresponding to the “effects” of $\hat{\beta}^*$ and $\hat{\theta}$, respectively, are $o_p(1)$. This implies that the estimators we substitute for β and θ in the weights play *no role* in determining the large sample properties of the resulting GLS estimator $\hat{\beta}$ and leads to the “no penalty” phenomenon noted above.

This feature is the main “folklore” message – the (large sample) precision is unaffected not only by the need to estimate the parameters in the weights, but *how* these parameters are estimated (as long as they are estimated “sensibly.”)

- In the case $C = \infty$, $\hat{\beta}^* = \hat{\beta}$, the term \mathbf{D}_n also corresponds to a contribution from estimation of $\hat{\beta}$ in (9.6), but it is negligible, as $\mathbf{D}_n = o_p(1)$. This shows that, even in the case where we iterate the GLS algorithm to convergence, the properties of $\hat{\beta}$ are unaffected by the appearance of $\hat{\beta}$ itself in the weights.
- In fact, the result holds for *any* C . Nowhere in the argument does C appear. All we required was that $\hat{\beta}^*$ be the current estimate and $\hat{\theta}$ be the estimator for θ based on it. The important implication is that *all* GLS estimators (any C) have the *same* large sample distribution.

Thus, the theory provides no insight into whether or not it is necessary to iterate to convergence ($C = \infty$) or how large C should be.

- Because the effect of $\hat{\theta}$ also is negligible, as $\mathbf{E}_n = o_p(1)$, the result also implies that how one estimates θ does not matter in determining the properties of the resulting GLS estimator. Thus, whether we use PL or some other approach (e.g., other transformation of absolute residuals), as long as the estimator for θ satisfies (9.4), the properties of $\hat{\beta}$ are unaltered.

Of course, it is important to recognize that the folklore theorem is a large sample, approximate result. Intuition suggests that the implications, although theoretically interesting, might be a bit optimistic in practice (i.e., for small sample sizes n).

That is, whether $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in the weights are known or estimated might in fact have some effect on the properties of the resulting $\hat{\boldsymbol{\beta}}$ in sample sizes where that are not sufficiently large for the large sample result to have “kicked in.”

As we will discuss later, this is often the case. In Chapter 11, we will offer some more refined theoretical arguments that support this observation.

STANDARD ERRORS FOR THE COMPONENTS OF $\hat{\boldsymbol{\beta}}$: As we have discussed, a main byproduct of a large sample distributional result like (9.5) is a way to construct approximate estimates of uncertainty.

We may express the folklore result as

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}[\boldsymbol{\beta}_0, \sigma_0^2 \{ \mathbf{X}^T(\boldsymbol{\beta}_0) \mathbf{W}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \mathbf{X}(\boldsymbol{\beta}_0) \}^{-1}], \quad (9.8)$$

where $\mathbf{X}(\boldsymbol{\beta})$ is defined as before and $\mathbf{W}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \text{diag}\{g^{-2}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}_1), \dots, g^{-2}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}_n)\}$. The covariance matrix in (9.8) may also be written as

$$\sigma_0^2 \left\{ \sum_{j=1}^n g^{-2}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0, \mathbf{x}_j) f_{\boldsymbol{\beta}}(\mathbf{x}_j, \boldsymbol{\beta}_0) f_{\boldsymbol{\beta}}^T(\mathbf{x}_j, \boldsymbol{\beta}_0) \right\}^{-1}.$$

From now on and in subsequent chapters, we will move between writing this type of matrix in matrix or summation notation without comment.

As $\boldsymbol{\beta}_0$, σ_0 , and $\boldsymbol{\theta}_0$ are unknown, it is natural to substitute estimated values. For σ_0^2 , the obvious estimator, given $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ from the final iteration of GLS, is the “bias-corrected” estimator

$$\hat{\sigma}^2 = (n - p)^{-1} \sum_{j=1}^n g^{-2}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \mathbf{x}_j) \{Y_j - f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})\}^2.$$

Thus, to obtain estimated, approximate standard errors in practice for the components of $\hat{\boldsymbol{\beta}}$, one would take the square roots of the diagonal elements of the matrix

$$\hat{\sigma}^2 \{ \mathbf{X}^T(\hat{\boldsymbol{\beta}}) \mathbf{W}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \mathbf{X}(\hat{\boldsymbol{\beta}}) \}^{-1}. \quad (9.9)$$

The standard errors in the output in Sections 3.7 and 6.8 are in fact derived from formulæ like (9.9). The SAS proc `nlin` and R/Splus `nls()` software provide standard error estimates in the case where known weights are involved using the asymptotic covariance matrix for WLS given in (8.20). As the form of this covariance matrix and the one for GLS is the same, and the programs have been given estimated weights computed at the final values of the estimators, they simply use these estimated weights as if they were known to compute the standard errors.

In the case of fixed C , then, the formula (9.9) is used in a slightly different form; in particular, as $\hat{\beta}^*$ in the weights would be treated as fixed at step (iii), the standard errors emerging from the final call to the nonlinear regression program would have $\hat{\beta}^*$ rather than the final estimate $\hat{\beta}$ in the weight matrix \mathbf{W} . Presumably, if $\hat{\beta}^*$ and $\hat{\beta}$ are similar, the standard errors calculated this way should be very close to those that would be obtained if $\hat{\beta}$ were used instead, so most people do not bother to update them to do so.

In the case $C = \infty$, of course, at convergence, $\hat{\beta}^* = \hat{\beta}$, and the standard errors from the final invocation of step (iii) would be effectively derived from (9.9).

In the case of IRWLS in SAS `proc nlin`, where β in the weights is estimated (with a possibly estimated value for θ treated as fixed), the standard errors are calculated after the estimated value $\hat{\beta}$ is determined, so again, (9.9) is used as is with $\hat{\theta}$ equal to the current estimate.

As noted above, standard errors based on the folklore theorem may sometimes be “optimistic;” that is, they may be smaller than the true sampling variance (this has been gauged through simulation).

This is because the uncertainty due to estimation of β and θ in the weights is not taken into account. The folklore theorem says this uncertainty is negligible, but in finite samples, it may not be. Thus, it is usually advisable in problems where the sample size is not large to interpret these standard errors with caution.

9.3 Misspecification of the variance function and GLS

In Section 8.3, we saw that, in the case where the weights used in estimation of β are chosen to be a set of fixed constants, incorrect choice of these constants does not lead to an *inconsistent* estimator for β , as the estimating equation is still *unbiased*. However, such an incorrect choice will lead to potentially *inefficient* estimation (relative to the precision that could be achieved using the correct weights).

In the GLS approach, the “weights” are dictated by the choice of variance function. Thus, by analogy, it is natural to consider what happens in the event that the variance function has not been correctly specified, which would presumably lead to incorrect “estimated weights” at each \mathbf{x}_j .

An obvious complication when the variance function is not correctly specified is that the incorrect model may depend on parameters for which values may not be specified, so that these parameters must be estimated.

Because these parameters appear in a model that does not characterize the truth, it is not even clear what they represent or what is being “estimated” when the incorrect model is fitted, for example, by the PL approach!

To formalize, suppose that, *in truth*, the data follow the mean-variance model

$$E(Y_j|\mathbf{x}_j) = f(\mathbf{x}_j, \boldsymbol{\beta}), \quad \text{var}(Y_j|\mathbf{x}_j) = \sigma^2 g^2(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}_j) \quad (9.10)$$

but we *incorrectly specify* the model as

$$E(Y_j|\mathbf{x}_j) = f(\mathbf{x}_j, \boldsymbol{\beta}), \quad \text{var}(Y_j|\mathbf{x}_j) = \tau^2 h^2(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x}_j). \quad (9.11)$$

Thus, although the data come from a model of the form (9.10), we fit a model where the variance function dictates some relationship other than the true one, g . In (9.11), the parameters are τ and $\boldsymbol{\gamma}$, where τ is a scale parameter. In the correct model (9.10), σ and $\boldsymbol{\theta}$ represent parameters in the model that fully and correctly characterize the variance, with true values σ_0 and $\boldsymbol{\theta}_0$ at which evaluation of the variance model yields exactly the value for conditional variance at \mathbf{x}_j (along with $\boldsymbol{\beta}_0$).

The parameters τ and $\boldsymbol{\gamma}$, on the other hand, do not have “true values” in the sense that the variance model in (9.11) evaluated at these values will give precisely the true variance at any \mathbf{x}_j . It may well be that $\boldsymbol{\gamma}$ is of a different dimension, r , say, from that of $\boldsymbol{\theta}$ (q). In fact, it is not even clear how $\boldsymbol{\beta}$ in the mean model enters into the picture.

Suppose we have available under these conditions an estimator $\hat{\boldsymbol{\beta}}^*$ satisfying the condition (9.4), $n^{1/2}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0) = O_p(1)$, where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$ in (9.10). We know immediately of one such estimator, the OLS estimator.

Thus, suppose we estimate $\boldsymbol{\beta}$ in (9.11) by OLS to obtain $\hat{\boldsymbol{\beta}}^*$ – as this does not involve the (incorrect) variance function, we know that this estimator will be consistent. Then, (unknowingly) taking the misspecified variance model in (9.11) as correct, suppose at step (ii) of the GLS algorithm we decide to estimate τ and $\boldsymbol{\gamma}$ by the PL approach. From Chapter 6, we would find $\hat{\tau}$ and $\hat{\boldsymbol{\gamma}}$ solving

$$\sum_{j=1}^n \left[\frac{\{Y_j - f(\mathbf{x}_j, \hat{\boldsymbol{\beta}}^*)\}^2 - \hat{\tau}^2 h^2(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}, \mathbf{x}_j)}{h^2(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}, \mathbf{x}_j)} \right] \boldsymbol{\xi}_\gamma(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\gamma}}, \mathbf{x}_j) = \mathbf{0}, \quad (9.12)$$

where $\boldsymbol{\xi}_\gamma(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x}_j)$ is the $(r + 1 \times 1)$ vector whose first element is equal to one and the remaining r elements are the derivatives of $\log\{h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{x}_j)\}$ with respect to the elements of $\boldsymbol{\gamma}$.

What does solving (9.12) yield?

To gain insight, consider the general situation of M-estimation.

In Section 8.2, we considered the situation of solving an equation of the form

$$\sum_{j=1}^n \Psi_j(\mathbf{Z}_j, \boldsymbol{\eta}) = \mathbf{0}.$$

We assumed implicitly that the model underlying the estimation is correctly specified and that the parameter $\boldsymbol{\eta}$ is the parameter of interest such that the distribution F_j of \mathbf{Z}_j depends on the true value $\boldsymbol{\eta}_0$. Recall that $\hat{\boldsymbol{\eta}}$ will be consistent for $\boldsymbol{\eta}_0$ under regularity conditions if

$$E\{\Psi_j(\mathbf{Z}_j, \boldsymbol{\eta}_0)\} = \mathbf{0},$$

where expectation is with respect to the true distribution of \mathbf{Z}_j .

It turns out that, in fact, this may be relaxed somewhat. If instead we have only that

$$\sum_{j=1}^n E\{\Psi_j(\mathbf{Z}_j, \boldsymbol{\eta}_0)\} = \mathbf{0} \quad (9.13)$$

(so that each summand does not necessarily have mean zero), then under regularity conditions, it still holds in general that $\hat{\boldsymbol{\eta}} \xrightarrow{p} \boldsymbol{\eta}_0$, and the argument leading to the asymptotic normality of the estimator for $\boldsymbol{\eta}$ in Section 8.2 goes through unchanged except for one modification. In particular, the covariance matrix of $\Psi_j(\mathbf{Z}_j, \boldsymbol{\eta}_0)$ is no longer equal to

$$E\{\Psi_j(\mathbf{Z}_j, \boldsymbol{\eta}_0)\Psi_j^T(\mathbf{Z}_j, \boldsymbol{\eta}_0)\},$$

so that the definitions of the matrices \mathbf{B}_n and \mathbf{B} in the argument must be changed; e.g., $\mathbf{B}_n \approx n^{-1} \sum_{j=1}^n \text{var}\{\Psi_j(\mathbf{Z}_j, \boldsymbol{\eta}_0)\}$ instead.

Now suppose that we have an incorrect model with parameter $\boldsymbol{\gamma}$, which leads us to solve instead the estimating equation

$$\sum_{j=1}^n \Psi_j^*(\mathbf{Z}_j, \boldsymbol{\gamma}) = \mathbf{0},$$

where Ψ_j^* is some other function of $\boldsymbol{\gamma}$ and the data \mathbf{Z}_j dictated by this incorrect model. Now Ψ_j^* is some function of the random vector \mathbf{Z}_j . Usually, there exists a value $\boldsymbol{\gamma}^*$ such that

$$\sum_{j=1}^n E\{\Psi_j^*(\mathbf{Z}_j, \boldsymbol{\gamma}^*)\} = \mathbf{0}, \quad (9.14)$$

where expectation here is still with respect to the *true distribution* of \mathbf{Z}_j .

By analogy to (9.13), it turns out that, under some conditions, if (9.14) holds, solving the “incorrect” estimating equation will result in an estimator $\hat{\boldsymbol{\gamma}}$ that satisfies

$$\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*.$$

Of course, γ^* does not necessarily have any meaning as representing a quantity that pertains to the true mechanism generating the data. However, it is a fixed quantity, dictated by (9.14). Thus, it is often said, perhaps misleadingly, that $\hat{\gamma}$ is “consistent” for γ^* (see page 188). One may in fact go on to pursue an argument exactly like that in Section 8.2 to establish that $n^{1/2}(\hat{\gamma} - \gamma^*)$ converges in distribution to a mean-zero multivariate normal random vector with a covariance matrix that may be derived. It thus follows that

$$n^{1/2}(\hat{\gamma} - \gamma^*) = O_p(1).$$

TERMINOLOGY: The value γ^* may be thought of as the value that “tries to get closest” to representing the truth within the confines of a misspecified model. It has consequently sometimes been called the *least false parameter*. The important conceptual point is that, even with a misspecified model, if we estimate a parameter in the model, we may still deduce the behavior of the estimator, even if the parameter has no real meaning.

Now consider the particular situation of (9.12). We may interpret this as a case where the “mean” $\tau^2 h^2(\beta, \gamma, \mathbf{x}_j)$ of the “response” $\{Y_j - f(\mathbf{x}_j, \beta)\}^2$ has been misspecified (and, of course, some “weighting” is also taking place). Now if $\hat{\beta}^*$ is a consistent estimator for the true value of β_0 (e.g., OLS) and held fixed when the solution is found, then we would expect that the same conclusions in the generic case of M-estimation with a misspecified model above. Clearly, it is no longer the case that the expectation of “response” – “mean” (conditional on \mathbf{x}_j) under the *true variance model* (9.10) at some value of τ and γ is zero, even if the correct value β_0 were substituted. It is usually said that such an estimating equation is *biased*. However, it is not far-fetched to think that there are values τ^* and γ^* that make things “average out” to zero over n (conditional on all n \mathbf{x}_j).

Viewing the problem as one of M-estimation with a misspecified model, then, we may conclude that solving the *biased* estimating equation (9.12) will yield estimators such that $n^{1/2}(\hat{\tau} - \tau^*) = O_p(1)$ and

$$n^{1/2}(\hat{\gamma} - \gamma^*) = O_p(1) \tag{9.15}$$

for some values τ^* and γ^* ; see page 195.

Of course, we have used PL estimation as an example here. It should be clear that the same sort of argument would apply in the case of other estimators for variance parameters. We will discuss the properties of the general class of variance parameter estimators in Chapter 12.

Now return to considering the GLS algorithm. Suppose that we solve for $\hat{\beta}$ in step (iii) the GLS equation

$$\sum_{j=1}^n h^{-2}(\hat{\beta}^*, \hat{\gamma}, \mathbf{x}_j) \{Y_j - f(\mathbf{x}_j, \hat{\beta})\} f_{\beta}(\mathbf{x}_j, \hat{\beta}) = \mathbf{0}. \tag{9.16}$$

As in the previous section, and using the above discussion, if we consider solving both the OLS equation, the PL equation (9.12), and the GLS equation (9.16), we have that $\hat{\beta}^* \xrightarrow{p} \beta_0$, the true value of β in (9.10), and $\hat{\gamma} \xrightarrow{p} \gamma^*$. Thus, if we focus on (9.16) evaluated at these values, we have

$$E\{h^{-2}(\beta_0, \gamma^*, \mathbf{x}_j)\{Y_j - f(\mathbf{x}_j, \beta_0)\}f_{\beta}(\mathbf{x}_j, \beta_0)|\mathbf{x}_j\} = \mathbf{0},$$

so that we may conclude that (9.16) is indeed an *unbiased* estimating equation.

Thus, despite the use of the “wrong” variance function, we still expect the GLS estimator $\hat{\beta}$ solving (9.16) to be consistent for β_0 . In fact, it should be clear that repeating the process with $\hat{\beta}^*$ as a (consistent) GLS estimator using the incorrect model will yield an update at step (iii) that is also consistent. Thus, heuristically, solving the GLS equation (9.16) with the “wrong” variance function should result in a consistent estimator for β for any C .

Assuming consistency of $\hat{\beta}$, $n^{1/2}(\hat{\beta} - \beta_0) = O_p(1)$, and $n^{1/2}(\hat{\gamma} - \gamma^*) = O_p(1)$ as in (9.15), we have the same situation as in the folklore theorem. We may thus expand (9.16) in $(\hat{\beta}^{*T}, \hat{\gamma}^T, \hat{\beta}^T)^T$ about $(\beta_0^T, \gamma^*, \beta_0^T)^T$ as in that argument. The steps are identical, so we do not repeat them all here.

Define

$$\epsilon_j = \frac{Y_j - f(\mathbf{x}_j, \beta_0)}{\sigma_0 g(\beta_0, \theta_0, \mathbf{x}_j)};$$

note that this involves the *true* mean and variance functions dictated by the correct model (9.10), so that $E(\epsilon_j|\mathbf{x}_j) = 0$ and $\text{var}(\epsilon_j|\mathbf{x}_j) = 1$, where, of course, expectation is with respect to the true (conditional) distributions of the Y_j . Write as before $w_j = g^{-2}(\beta_0, \theta_0, \mathbf{x}_j)$, and let $u_j = h^{-2}(\beta_0, \gamma^*, \mathbf{x}_j)$. Then the expansion yields

$$\mathbf{0} \approx \mathbf{C}_n^* + (\mathbf{A}_{n1}^* + \mathbf{A}_{n2}^*)n^{1/2}(\hat{\beta} - \beta_0) + \mathbf{D}_n^*n^{1/2}(\hat{\beta}^* - \beta_0) + \mathbf{E}_n^*n^{1/2}(\hat{\gamma} - \gamma^*), \quad (9.17)$$

where

$$\begin{aligned} \mathbf{C}_n^* &= \sigma_0 n^{-1/2} \sum_{j=1}^n u_j w_j^{-1/2} f_{\beta}(\mathbf{x}_j, \beta_0) \epsilon_j, \\ \mathbf{A}_{n1}^* &= \sigma_0 n^{-1} \sum_{j=1}^n u_j w^{-1/2} f_{\beta\beta}(\mathbf{x}_j, \beta_0) \epsilon_j, \quad \mathbf{A}_{n2}^* = -n^{-1} \sum_{j=1}^n u_j f_{\beta}(\mathbf{x}_j, \beta_0) f_{\beta}^T(\mathbf{x}_j, \beta_0), \\ \mathbf{D}_n^* &= -2\sigma_0 n^{-1} \sum_{j=1}^n u_j^{3/2} f_{\beta}(\mathbf{x}_j, \beta_0) h_{\beta}^T(\beta_0, \gamma^*, \mathbf{x}_j) \epsilon_j, \\ \mathbf{E}_n^* &= -2\sigma_0 n^{-1} \sum_{j=1}^n u_j^{3/2} f_{\beta}(\mathbf{x}_j, \beta_0) h_{\gamma}^T(\beta_0, \gamma^*, \mathbf{x}_j) \epsilon_j, \end{aligned}$$

and h_{β} and h_{γ} represent the vectors of partial derivatives of h with respect to β and γ .

As in the argument for the folklore result, \mathbf{A}_{n1}^* , \mathbf{E}_n^* , and \mathbf{D}_n^* all converge in probability to zero; that this last term is negligible is especially interesting, as it shows that there is *no effect* of the “wrong estimator” $\hat{\gamma}$ for the incorrect variance model!

Rewriting (9.17) as

$$n^{1/2}(\hat{\beta} - \beta_0) \approx -\mathbf{A}_{n2}^{*-1}\mathbf{C}_n^*,$$

letting $\mathbf{X} = \mathbf{X}(\beta_0)$, and defining as in Section 8.3 $\mathbf{U} = \text{diag}(u_1, \dots, u_n)$, we have that

$$\mathbf{A}_{n2}^* \rightarrow -\mathbf{A}, \quad \mathbf{A} = \lim_{n \rightarrow \infty} n^{-1}\mathbf{X}^T\mathbf{U}\mathbf{X},$$

and

$$\mathbf{C}_n^* \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{B}), \quad \mathbf{B} = \lim_{n \rightarrow \infty} n^{-1}\mathbf{X}^T\mathbf{U}\mathbf{W}^{-1}\mathbf{U}\mathbf{X}.$$

Combining, we obtain that

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}),$$

so that

$$\hat{\beta} \sim \mathcal{N}\{\beta_0, \sigma_0^2(\mathbf{X}^T\mathbf{U}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{U}\mathbf{W}^{-1}\mathbf{U}\mathbf{X})(\mathbf{X}^T\mathbf{U}\mathbf{X})^{-1}\}. \quad (9.18)$$

We may compare the result in (9.18) to that obtained in Section 8.3 in the case of *known* weights. In particular, note that (9.18) is identical to the result in (8.15) when the w_j and u_j were treated as known constants. Thus, just as in the folklore result, we obtain the interesting conclusion that, even if we estimate weights (by substituting parameter estimators) rather than knowing their values, we will obtain the same large sample distribution for the estimator for β ; here, this is seen to hold in the case of an incorrect model/incorrect constants.

- Note, in fact, that the folklore theorem may be regarded as just a special case of this general result, where h and g are the same.
- It should be obvious that taking the function h to be identically equal to 1 for all j would thus yield the large sample properties of the OLS estimator when the variance is really nonconstant, which we have already derived. Here, of course, there would be no β or γ to be estimated in “weights.”
- The efficiency comparisons carried out in Section 8.4 in the case of fixed weights thus carry over *unaltered* to the setting of estimation of variance functions!

IMPLICATION: Using an incorrect variance function will not affect the consistency of the GLS estimator for β , but it *will* affect the efficiency of the resulting estimator. Using an incorrect variance function may result in a less precise estimator for β than if the correct function is used.

- The same issues discussed on page 217 carry over to this more general setting. How one estimates “weights,” even using an incorrect model, does not play a role in the large sample properties of the GLS estimator for β in a correctly-specified mean model. Of course, this may be optimistic in finite samples.

A version of this result is discussed in the case of multivariate response by Liang and Zeger (1986), and it is often attributed more generally to these authors, although it has been known for considerably longer.

9.4 Correction of standard errors

The results of the previous section indicate that, if we have incorrectly specified the variance model, the usual formula for obtaining approximate standard errors for the elements of the GLS estimator $\hat{\beta}$ is not appropriate. This could potentially result in erroneous inferences: a misleading assessment of the precision of $\hat{\beta}$ may result, and test and confidence intervals, to be discussed next in Section 9.5, could be compromised.

However, although we may be concerned that we have selected an incorrect variance model, the result in (9.18) is not really helpful – to use this result, we must know the *true* variance function (in order to deduce the values w_j appearing in the “middle” piece of the asymptotic covariance matrix). Of course, if we knew this, we would have used it for estimation of β !

Although the result is not immediately useful, it does give insight into how one might “protect against” a potentially misspecified variance model when calculating standard error estimates.

IDEA: To formalize this, consider again the situation of the last section, where the true mean-variance model is $E(Y_j|\mathbf{x}_j) = f(\mathbf{x}_j, \beta)$ and $\text{var}(Y_j|\mathbf{x}_j) = \sigma^2 g^2(\beta, \theta, \mathbf{x}_j)$ as in (9.10) but we have misspecified the variance model as in (9.11), instead assuming $\text{var}(Y_j|\mathbf{x}_j) = \tau^2 h^2(\beta, \gamma, \mathbf{x}_j)$.

If we specify this incorrect model and are unaware that we have done so, then we would presume that the folklore result holds with the weight matrix dictated by the variance model h . Thus, we would conclude that the GLS estimator based on this wrong model satisfies

$$\hat{\beta} \sim \mathcal{N}[\beta_0, \hat{\tau}^2 \{ \mathbf{X}^T(\hat{\beta}) \mathbf{U}(\hat{\beta}, \hat{\gamma}) \mathbf{X}(\hat{\beta}) \}^{-1}];$$

here, we have substituted the estimated values into the matrices \mathbf{X} and \mathbf{U} , as would be the case for obtaining estimated standard errors in practice.

As this result is based on the incorrect assumption that h is the correct variance function, the standard errors so obtained may be misleading.

Under these circumstances, to obtain standard errors that give an accurate assessment of uncertainty, we would rather base them on the result in (9.18). As noted above, we cannot do this directly, but we can do something close. The covariance matrix we would like to estimate is, from (9.18),

$$(\mathbf{X}^T \mathbf{U} \mathbf{X})^{-1} (\sigma_0^2 \mathbf{X}^T \mathbf{U} \mathbf{W}^{-1} \mathbf{U} \mathbf{X}) (\mathbf{X}^T \mathbf{U} \mathbf{X})^{-1}. \quad (9.19)$$

From above, we can estimate the two “end” pieces by simply substituting the estimators for β and γ from the fit of the (incorrect) model. The middle piece is the troublesome one. Note that

$$\sigma_0^2 n^{-1} \mathbf{X}^T \mathbf{U} \mathbf{W}^{-1} \mathbf{U} \mathbf{X} = n^{-1} \sum_{j=1}^n u_j^2 f_{\beta}(\mathbf{x}_j, \beta_0) f_{\beta}^T(\mathbf{x}_j, \beta_0) \{ \sigma_0^2 g^2(\beta_0, \boldsymbol{\theta}_0, \mathbf{x}_j) \},$$

as $w_j = g^{-2}(\beta_0, \boldsymbol{\theta}_0, \mathbf{x}_j)$. We do not know the w_j ; however, we *do* know that, in truth,

$$E[\{Y_j - f(\mathbf{x}_j, \beta_0)\}^2 | \mathbf{x}_j] = \sigma_0^2 g^2(\beta_0, \boldsymbol{\theta}_0, \mathbf{x}_j).$$

Thus, by the weak law of large numbers, we know that

$$n^{-1} \sum_{j=1}^n u_j^2 f_{\beta}(\mathbf{x}_j, \beta_0) f_{\beta}^T(\mathbf{x}_j, \beta_0) \{Y_j - f(\mathbf{x}_j, \beta_0)\}^2 - n^{-1} \sum_{j=1}^n u_j^2 f_{\beta}(\mathbf{x}_j, \beta_0) f_{\beta}^T(\mathbf{x}_j, \beta_0) \{ \sigma_0^2 g^2(\beta_0, \boldsymbol{\theta}_0, \mathbf{x}_j) \} \xrightarrow{p} \mathbf{0},$$

suggesting that we could substitute the squared deviations $\{Y_j - f(\mathbf{x}_j, \beta_0)\}^2$ and get something “close” to the middle term.

Of course, we do not know these deviations, but we do have a consistent estimator for β_0 , with which we may estimate them. Letting

$$r_j = Y_j - f(\mathbf{x}_j, \hat{\beta})$$

denote the unweighted GLS residual, the obvious suggestion is thus to estimate the middle matrix in (9.19) by

$$\mathbf{X}^T(\hat{\beta}) \mathbf{U}(\hat{\beta}, \hat{\gamma}) \mathbf{R} \mathbf{U}(\hat{\beta}, \hat{\gamma}) \mathbf{X}(\hat{\beta}), \quad \mathbf{R} = \text{diag}(r_1^2, \dots, r_n^2).$$

It may be shown that in fact

$$n^{-1} \sum_{j=1}^n h^{-4}(\hat{\beta}, \hat{\gamma}, \mathbf{x}_j) f_{\beta}(\mathbf{x}_j, \hat{\beta}) f_{\beta}^T(\mathbf{x}_j, \hat{\beta}) r_j^2 - n^{-1} \sum_{j=1}^n u_j^2 f_{\beta}(\mathbf{x}_j, \beta_0) f_{\beta}^T(\mathbf{x}_j, \beta_0) \{ \sigma_0^2 g^2(\beta_0, \boldsymbol{\theta}_0, \mathbf{x}_j) \} \xrightarrow{p} \mathbf{0}$$

(by expanding the first term about $\hat{\beta} = \beta_0$ and $\hat{\gamma} = \gamma^*$ and using the facts that $\hat{\beta} - \beta_0 = O_p(n^{-1/2})$ and $\hat{\gamma} - \gamma^* = O_p(n^{-1/2})$ and the weak law of large numbers).

RESULT: To protect against possible misspecification of the variance model in GLS estimation of β , it is suggested to derive standard errors for the elements of $\hat{\beta}$ by the square roots of the diagonal elements of the estimated covariance matrix

$$\{\mathbf{X}^T(\hat{\beta})\mathbf{U}(\hat{\beta}, \hat{\gamma})\mathbf{X}(\hat{\beta})\}^{-1}\mathbf{X}^T(\hat{\beta})\mathbf{U}(\hat{\beta}, \hat{\gamma})\mathbf{R}\mathbf{U}(\hat{\beta}, \hat{\gamma})\mathbf{X}(\hat{\beta})\{\mathbf{X}^T(\hat{\beta})\mathbf{U}(\hat{\beta}, \hat{\gamma})\mathbf{X}(\hat{\beta})\}^{-1}. \quad (9.20)$$

- Even if it has been derived using an incorrect variance model, $\hat{\beta}$ is still consistent. Thus, it is reasonable to use as an estimator for β (recognizing that it is inefficient). Calculation of standard errors using (9.20) is an attempt to ensure the assessments of precision are correct.

This idea is a special case of a general technique that has many different names.

- In the particular case of $h \equiv 1$, so that we use OLS when the variance may in fact be nonconstant, $\mathbf{U} = \mathbf{I}$. In the econometrics literature, the estimated covariance matrix (9.20) under this condition has been called the *heteroscedasticity-consistent* covariance matrix. This is because it is itself a consistent estimator of the true covariance matrix of $\hat{\beta}$ given in (9.19) with $\mathbf{U} = \mathbf{I}$ in the case where the true variance may be nonconstant (heteroscedasticity).

In fact, in situations where one's main interest is to estimate β and obtain realistic standard errors and where modeling the variance could be quite complicated, this approach has been advocated for simplicity: just estimate β by OLS and “fix up” the standard errors. The obvious drawback is that the OLS estimator may be very inefficient.

- More recently, (9.20) has been called the *robust sandwich* estimator of the sampling covariance matrix of $\hat{\beta}$; this term is usually attributed to Liang and Zeger (1986). “Robust” refers to the hope that, as an estimator of sampling variation, (9.20) is insensitive to misspecification of the variance model. The term “sandwich” refers to the form of the estimator: a “correction term” “sandwiched” between two copies of the covariance matrix one would naively use if one believed the variance model were correct. See also Moore and Tsiatis (1991) for an example in the univariate case.
- Indeed, it is straightforward to see that this estimator for the covariance matrix is exactly that one obtains by applying the general “sandwich” technique discussed on page 201 to the estimating equation (9.16), where the “meat” is estimated by the sample covariance matrix of the equation's summand.

WARNING: Several authors have expressed concerns over the practical performance of this approach. Most of the discussion has been with respect to the multivariate generalization of the approach, which we will discuss in Chapter 14, but the implications are similar.

- (9.20) may be shown to be a consistent estimator of the true covariance matrix of the estimator.

However, several authors have reported that, in finite samples, it may produce rather unreliable estimates of the true sampling variation (as deduced by simulation studies). See, for example, Rotnitzky and Jewell (1990) and Kauermann and Carroll (2001).

Part of this may be due to the estimator’s apparent sensitivity to “unusual observations.” In particular, using the squared residuals r_j^2 as essentially a proxy for the true variance at \mathbf{x}_j , thus basing this on a single data value, may be sensitive to an “outlying” value of Y_j , and this effect could be noticeable in small samples. Thus, “robustness” to an incorrect variance model could be offset by lack of “robustness” to unusual data values.

- Some authors have suggested that in some circumstances one might be better off just using the usual folklore result and not attempt to “correct” for possible misspecification of variance.

Alternatively, other authors advocate always using the correction for protection, as it may be optimistic to expect that one has modeled variance perfectly. In the case of univariate response that we have been discussing, it is actually quite reasonable to expect to be able to model variance “well,” so it is routine not to use the sandwich correction. However, in the case of multivariate response, which we will discuss in Chapter 14, the reasons for using the sandwich correction are much more compelling, as we will discuss.

- Recently, there have been attempts to improve the estimator via “small sample” corrections; see, for example Mancl and DeRouen (2001).

9.5 Inference for β

Once β has been estimated and standard errors obtained, we may wish to construct confidence intervals for β_0 , carry out hypothesis tests, and so on. Here, we discuss several approaches, all based on large sample theory approximations of the type used in the folklore theorem. Throughout this section, we assume that the variance model has been correctly specified unless otherwise indicated, so that in truth the model $\text{var}(Y_j|\mathbf{x}_j) = \sigma^2 g^2(\beta, \theta, \mathbf{x}_j)$ is correct.

WALD INFERENCE: A natural and easy approach to these objectives is to use the usual “Z-statistic” method that is applied in situations where exact, finite-sample results are available. From the folklore theorem, we have

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}[\boldsymbol{\beta}_0, \hat{\sigma}^2 \{\mathbf{X}^T(\hat{\boldsymbol{\beta}}) \mathbf{W}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \mathbf{X}(\hat{\boldsymbol{\beta}})\}^{-1}], \quad (9.21)$$

$$\hat{\sigma}^2 = (n - p)^{-1} \sum_{j=1}^n g^{-2}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \mathbf{x}_j) \{Y_j - f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})\}^2. \quad (9.22)$$

To form confidence intervals and test statistics regarding one of the elements of $\boldsymbol{\beta}$, β_k , say, $k = 1, \dots, p$ (a scalar), let $SE(\hat{\beta}_k)$ be the square root of the k th diagonal element of the estimated covariance matrix in (9.21).

- The usual, symmetric, equal-tailed confidence interval for the true value of β_k , with confidence coefficient $100(1 - \alpha)\%$, is

$$\hat{\beta}_k \pm c_{\alpha/2} SE(\hat{\beta}_k),$$

where $c_{\alpha/2}$ is a critical value from a symmetric distribution. Following the folklore result, $c_{\alpha/2}$ is often chosen to be the appropriate quantile of the standard normal distribution, $c_{\alpha/2} = \Phi^{-1}(1 - \alpha/2) = z_{1-\alpha/2}$.

- Recall that it has been observed that inferences based on the folklore result can be optimistic in practical, finite-sample situations. Thus, it is common to replace the standard normal critical values by something else. By analogy to the “classical” linear regression case, a routine approach is to instead choose $c_{\alpha/2} = t_{n-p, 1-\alpha/2}$, where $t_{n-p, 1-\alpha/2}$ is the quantile of the t distribution with $n-p$ degrees of freedom with area $1 - \alpha/2$ to the left. The rationale is the same as in the “classical” case: to account for the degrees of freedom lost in estimating $\hat{\boldsymbol{\beta}}$. However, it is customary to make no attempt to take into account the fact that $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ have also been replaced by estimates.
- Test statistics would be constructed in the analogous way. For example, to test $H_0 : \beta_k = \beta_{k0}$ versus a one- or two-sided alternative, the test statistic $(\hat{\beta}_k - \beta_{k0})/SE(\hat{\beta}_k)$ would be compared to the relevant standard normal or t critical value.
- Programs such as SAS `proc nlin` and R/Splus `nls()`, when used at step (iii) of the GLS algorithm, print out such *Wald statistics* for each component of $\boldsymbol{\beta}$, along with a p-value based on the normal or t critical values. See the discussion on page 218 for more on how the standard errors are calculated.

For more complicated questions of interest, the above extends in the obvious way. Suppose we are interested in a confidence region for or test concerning a subset of the elements of β . Partitioning β as

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \beta_1 \ (r \times 1), \ \beta_2 \ (p - r \times 1),$$

suppose we are interested in β_2 (we can, of course, always reorder the elements of β to group those of interest together). Similarly, partition the estimated covariance matrix of $\hat{\beta}$ as

$$\hat{\Sigma} = \hat{\sigma}^2 \{ \mathbf{X}^T(\hat{\beta}) \mathbf{W}(\hat{\beta}, \hat{\theta}) \mathbf{X}(\hat{\beta}) \}^{-1} = \begin{pmatrix} \hat{\Sigma}^{11} & \hat{\Sigma}^{12} \\ \hat{\Sigma}^{12T} & \hat{\Sigma}^{22} \end{pmatrix}.$$

Thus, $\hat{\beta}_2 \sim \mathcal{N}(\beta_{20}, \hat{\Sigma}^{22})$ ($p - r \times 1$), and it follows that

$$(\hat{\beta}_2 - \beta_{20})^T (\hat{\Sigma}^{22})^{-1} (\hat{\beta}_2 - \beta_{20}) \sim \chi_{p-r}^2.$$

This approximation may be used as the basis for confidence regions and hypothesis tests in the usual way.

More generally, suppose we are interested in a linear contrast $\mathbf{L}\beta$ of the elements of β , where \mathbf{L} is ($r \times p$) and of full rank. Then, treating the large sample results as exact, $\mathbf{L}\hat{\beta} \sim \mathcal{N}(\mathbf{L}\beta_0, \mathbf{L}\hat{\Sigma}\mathbf{L}^T)$, and it follows that

$$(\hat{\beta} - \beta_0)^T \mathbf{L}^T (\mathbf{L}\hat{\Sigma}\mathbf{L}^T)^{-1} \mathbf{L}(\hat{\beta} - \beta_0) \sim \chi_r^2.$$

ADVANTAGES OF WALD INFERENCE:

- Wald inference is often the default choice because it is easy to implement and is based on familiar ideas used in simpler problems where exact finite-sample results are available.
- Another advantage is that it is straightforward to replace the “folklore” standard error estimates by the “robust sandwich” versions to protect against misspecification of the variance model.

DISADVANTAGES OF WALD INFERENCE:

- The large sample distributional results may be unreliable in practice. For instance, the asymptotic standard normal or chi-square approximations above may be poor in small samples, resulting in erroneous conclusions.
- Wald inference is not *invariant* to reparameterization of the model. Thus, if we reparameterize the mean model and then attempt to make inference on the same feature in the new parameterization, we may be led to different conclusions.

In general, the inadequacies of Wald inference are widely recognized; however, because of ease of implementation and ready availability in the output of common software packages, it is widely used.

“*LIKELIHOOD-BASED*” *APPROACHES*: There a number of alternative approaches to constructing confidence intervals and hypothesis tests that are meant to circumvent some of the disadvantages of Wald inference, at the expense of increased complexity. We do not attempt to demonstrate all of these here; rather, we give a flavor for the types of approaches that have been suggested.

For definiteness, we will consider again the problem where we may partition

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{\beta}_1 \text{ } (r \times 1), \boldsymbol{\beta}_2 \text{ } (p - r \times 1).$$

Suppose we are interested in testing $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$ for some specified value $\boldsymbol{\beta}_{20}$.

- We may wish to compare two nested generalized linear models, for example, where $\boldsymbol{\beta}_2$ contains the coefficients of the linear predictor corresponding to a group of covariates whose joint importance is to be assessed. In this case, $\boldsymbol{\beta}_{20} = \mathbf{0}$.
- For nonlinear models where the mean function is dictated by theoretical considerations, $\boldsymbol{\beta}_2$ may correspond to certain physical parameters, and we may be interested in whether there is evidence that these differ from default values (that may be different from zero).

We first consider a “normal likelihood ratio test” based on “pretending” that the GLS weights are fixed. Recall that, when the weights w_j are *known*, solving the GLS equation

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

corresponds exactly to maximum likelihood estimation of $\boldsymbol{\beta}$ under the assumption that $Y_j | \mathbf{x}_j$ is normal with $\text{var}(Y_j | \mathbf{x}_j) = \sigma^2 / w_j$ for each j . The idea is to use the normal likelihood as the basis for a test, even if the data themselves are not really normally distributed. We now show by a heuristic argument that the “likelihood ratio test” statistic derived from these considerations has a large sample chi-square distribution regardless of whether or not normality holds.

Ignoring σ^2 , the “important part” of the normal loglikelihood is, as a function of $\boldsymbol{\beta}$ and holding the weights fixed,

$$L(\boldsymbol{\beta}) = -(1/2) \sum_{j=1}^n \hat{w}_j \{Y_j - f(\mathbf{x}_j, \boldsymbol{\beta})\}^2, \quad \hat{w}_j = g^{-2}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \mathbf{x}_j).$$

Minimizing $L(\boldsymbol{\beta})$ in $\boldsymbol{\beta}$ for fixed weights gives the usual GLS estimating equation. Suppose that $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$ under H_0 . Expanding $L(\boldsymbol{\beta}_0)$ about $L(\hat{\boldsymbol{\beta}})$ yields

$$n^{-1}L(\boldsymbol{\beta}_0) \approx n^{-1}L(\hat{\boldsymbol{\beta}}) + n^{-1}L_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + (1/2)(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})^T \{n^{-1}L_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})\}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}).$$

Now $L_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, as it corresponds to the equation that we solve to obtain $\hat{\boldsymbol{\beta}}$ (for fixed weights).

Moreover,

$$n^{-1}L_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) = -n^{-1} \sum_{j=1}^n \hat{w}_j f_{\boldsymbol{\beta}}(\mathbf{x}_j, \hat{\boldsymbol{\beta}}) f_{\boldsymbol{\beta}}^T(\mathbf{x}_j, \hat{\boldsymbol{\beta}}) + n^{-1} \sum_{j=1}^n \hat{w}_j \{Y_j - f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})\} \mathbf{f}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\mathbf{x}_j, \hat{\boldsymbol{\beta}}).$$

The second term can be disregarded as negligible by the weak law of large numbers, as $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}_0$ under H_0 and the term has mean zero.

Combining all of this, we have

$$2\{L(\hat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta}_0)\} \approx \{n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\} \left\{ n^{-1} \sum_{j=1}^n \hat{w}_j f_{\boldsymbol{\beta}}(\mathbf{x}_j, \hat{\boldsymbol{\beta}}) f_{\boldsymbol{\beta}}^T(\mathbf{x}_j, \hat{\boldsymbol{\beta}}) \right\} \{n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}.$$

Now, from the argument leading to the folklore theorem, under H_0 with $\boldsymbol{\beta}_0$ the true value,

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approx \sigma_0 \boldsymbol{\Sigma}_{WLS} \left\{ n^{-1/2} \sum_{j=1}^n w_j^{1/2} f_{\boldsymbol{\beta}}(\mathbf{x}_j, \boldsymbol{\beta}_0) \epsilon_j \right\} = \sigma_0 \boldsymbol{\Sigma}_{WLS} \mathbf{C}_0,$$

say.

Thus, we have

$$\frac{2\{L(\hat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta}_0)\}}{\sigma_0^2} \approx \mathbf{C}_0^T \boldsymbol{\Sigma}_{WLS} \{n^{-1} \mathbf{X}^T(\hat{\boldsymbol{\beta}}) \mathbf{W}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \mathbf{X}(\hat{\boldsymbol{\beta}})\} \boldsymbol{\Sigma}_{WLS} \mathbf{C}_0.$$

Because $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are consistent for $\boldsymbol{\beta}_0$ and the true value $\boldsymbol{\theta}_0$ under H_0 , and the middle term is continuous in $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, it follows from the definition of $\boldsymbol{\Sigma}_{WLS}$ that the middle term converges in probability to $\boldsymbol{\Sigma}_{WLS}^{-1}$. Thus, we obtain

$$\frac{2\{L(\hat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta}_0)\}}{\sigma_0^2} \approx \mathbf{C}_0^T \boldsymbol{\Sigma}_{WLS} \mathbf{C}_0$$

under H_0 . In fact, even if we were to replace σ_0^2 by its estimator, because the estimator is consistent, the result would be unaltered.

Now consider the “restricted” estimator under H_0 , where we hold $\boldsymbol{\beta}_2$ fixed at the null value $\boldsymbol{\beta}_{20}$ and estimate the first r components $\boldsymbol{\beta}_1$ only. This is an $r < p$ dimensional problem.

If we were to use GLS to estimate β_1 , it is straightforward to realize that, applying the folklore theory treating β_2 as fixed at β_{20} we would have under H_0 (so $\beta_1 = \beta_{10}$, the true value under H_0)

$$n^{1/2}(\hat{\beta}_1 - \beta_{10}) \xrightarrow{L} \mathcal{N}(\mathbf{0}, \sigma_0^2 \Sigma_{WLS,11}),$$

where $\Sigma_{WLS,11}$ is the inverse of the upper left $(r \times r)$ submatrix $\Sigma_{WLS,11}^{-1}$ of Σ_{WLS} ($p \times p$). Moreover, from the argument leading to the folklore result, we may conclude that

$$n^{1/2}(\hat{\beta}_1 - \beta_{10}) \approx \sigma_0 \Sigma_{WLS,11} \left\{ n^{-1/2} \sum_{j=1}^n w_j f_{\beta,1}(\mathbf{x}_j, \beta_0) \epsilon_j \right\} = \sigma_0 \Sigma_{WLS,11} \mathbf{C}_{10},$$

say, where $f_{\beta,1}$ represents the $(r \times 1)$ vector of partial derivatives of f with respect to the elements of β_1 , and \mathbf{C}_{10} is equal to the first r elements of \mathbf{C}_0 .

Let $\hat{\beta}_0 = (\hat{\beta}_1^T, \beta_{20}^T)^T$ be the $(p \times 1)$ vector of the estimator under the restriction that $\beta_2 = \beta_{20}$ for H_0 .

Then $L(\hat{\beta}_0)$ is just the loglikelihood evaluated for the restricted problem, and, by an argument similar to that above, we may conclude that

$$\frac{2\{L(\hat{\beta}_0) - L(\beta_0)\}}{\sigma_0^2} \approx \mathbf{C}_{10}^T \Sigma_{WLS,10} \mathbf{C}_{10}.$$

Now consider the usual likelihood ratio test statistic

$$\begin{aligned} \frac{-2\{L(\hat{\beta}_0) - L(\hat{\beta})\}}{\sigma_0^2} &= \frac{-2\{L(\hat{\beta}_0) - L(\beta_0)\}}{\sigma_0^2} + \frac{2\{L(\hat{\beta}) - L(\beta_0)\}}{\sigma_0^2} \\ &\approx \mathbf{C}_0^T \Sigma_{WLS} \mathbf{C}_0 - \mathbf{C}_{10}^T \Sigma_{WLS,11} \mathbf{C}_{10} \\ &\approx \mathbf{C}_0^T \left\{ \Sigma_{WLS} - \begin{pmatrix} \Sigma_{WLS,11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right\} \mathbf{C}_0, \end{aligned}$$

where the last expression follows from the fact that the first r elements of \mathbf{C}_0 are \mathbf{C}_{10} .

We are now in a position to exploit the following generic result:

If $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and $\mathbf{A}\Sigma\mathbf{A} = \mathbf{A}$ with $k = \text{tr}(\Sigma\mathbf{A})$, then $\mathbf{Z}^T \mathbf{A} \mathbf{Z} \sim \chi_k^2$.

Note that $\mathbf{C}_0 \sim \mathcal{N}(\mathbf{0}, \Sigma_{WLS}^{-1})$, and it is straightforward to verify from the definition of $\Sigma_{WLS,11}$ that

$$\begin{aligned} &\left\{ \Sigma_{WLS} - \begin{pmatrix} \Sigma_{WLS,11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right\} \Sigma_{WLS}^{-1} \left\{ \Sigma_{WLS} - \begin{pmatrix} \Sigma_{WLS,11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right\} \\ &= \Sigma_{WLS} - \begin{pmatrix} \Sigma_{WLS,11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \end{aligned}$$

Moreover,

$$\text{tr} \left[\Sigma_{WLS}^{-1} \left\{ \Sigma_{WLS} - \begin{pmatrix} \Sigma_{WLS,11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right\} \right] = p - r,$$

again using the definition of $\Sigma_{WLS,11}$.

We may thus conclude that the “likelihood ratio statistic” satisfies

$$\frac{-2\{L(\hat{\beta}_0) - L(\hat{\beta})\}}{\sigma_0^2} \sim \chi_{p-r}^2.$$

This result may be used as the basis for tests and confidence regions in the usual ways, replacing σ_0^2 by the estimator $\hat{\sigma}^2$.

Variations on this theme are possible; Carroll and Ruppert (1988, p. 25) discuss constructing the “likelihood ratio test” statistic based on “profiling out” σ^2 (see page 130) and basing the statistic on the “profiled” (or “concentrated”) loglikelihood. This idea is also discussed in Seber and Wild (1989, section 5.2.3).

Modifications to the basic “likelihood ratio test” idea have been advocated to improve reliability in small samples. One such modification is to instead construct an asymptotically equivalent version of the test statistic by analogy to the ordinary F test in “classical” linear regression. In particular, an alternative F test is based on the result

$$\frac{\{L(\hat{\beta}_0) - L(\hat{\beta})\}/(p - r)}{L(\hat{\beta})/(n - p)} \sim F_{p-r, n-p}.$$

Seber and Wild (1989, section 5.3) discuss this idea further.

As we have noted, all of the aforementioned ideas treat $\hat{\theta}$ as if it were a fixed, known, constant, and thus do not attempt to “adjust” the statistics for the fact that $\hat{\theta}$ is estimated. Indeed, the presence of $\hat{\beta}$ in the weights is not accounted for, either. In Chapter 11, we will discuss the use of the bootstrap to take this extra uncertainty into account.

As mentioned in Section 4.6 in our discussion of quaslikelihood, in the case where there are no unknown parameters θ in the variance function and variance depends on β through a function of the mean response, it has been proposed to use the quaslikelihood in the same way as a loglikelihood to construct so-called “quaslikelihood ratio test” statistics. Specifically, with the quaslikelihood $L_{QL}(\beta)$ defined as in Section 4.6 (suppressing dependence on the data), it may be shown by an argument similar to the one above that

$$\frac{-2\{L_{QL}(\hat{\beta}_0) - L_{QL}(\hat{\beta})\}}{\sigma_0^2} \sim \chi_{p-r}^2.$$

See, for example, McCullagh (1983). A problem with the quasilielihood approach is that L_{QL} may be difficult to derive for general variance functions. As we saw in Section 4.6, in the case of the power-of-the-mean variance model, the derivation is relatively straightforward.

REMARKS:

- Inference based on “likelihood” approaches is thought to be more reliable in small samples than Wald inference; this has been deduced through simulations. Because every nonlinear problem is different, however, it is impossible to say that this is always the case.
- “Likelihood-based” inference is obviously more complicated to implement than Wald inference. Thus, use of these techniques is much less common in practice.

MORE COMPLICATED HYPOTHESES: For nonlinear models that arise from theoretical or empirical considerations in applications like growth analysis or pharmacokinetics, it is not uncommon for interest to focus on *nonlinear* functions of the elements of β .

For example, recall the data on the pharmacokinetics of indomethacin, most recently discussed in Section 7.5, for which, with x representing time (hours) after the dose, the biexponential model

$$f(x, \beta) = e^{\beta_1} \exp(-e^{\beta_2} x) + e^{\beta_3} \exp(-e^{\beta_4} x)$$

is reasonable; we have used the parameterization that enforces positivity here. In this model, the quantities e^{β_2} and e^{β_4} are rate constants, with units of 1/hour.

As discussed in Section 7.5, a quantity that is of some interest is the so-called *terminal half-life*. If $e^{\beta_4} < e^{\beta_2}$, then the second exponential term in the model dictates the “terminal phase” of the elimination of drug, which manifests itself as the “second part” of the decay. The “half-life” of this phase is the time it takes for the mean response in this phase to decrease by half, given by

$$t_{1/2} = \log 2 / e^{\beta_4},$$

which has units of hours. It is of general interest to estimate the terminal half-life and provide some assessment of the uncertainty in the estimate, e.g., approximate estimated standard errors and confidence intervals.

Here, then, the quantity of interest is itself a *nonlinear* function of the regression parameters. A standard approach to deriving estimates of uncertainty for such quantities is via the Wald approach.

Consider a general, real-valued nonlinear function of $\boldsymbol{\beta}$, $a(\boldsymbol{\beta})$, say; the following argument is easily extended by analogy to the case of a vector-valued function. The obvious estimator for $a(\boldsymbol{\beta}_0)$, the function evaluated at the true value of $\boldsymbol{\beta}$, is $a(\hat{\boldsymbol{\beta}})$. By a standard Taylor series expansion to linear terms, we have

$$a(\hat{\boldsymbol{\beta}}) \approx a(\boldsymbol{\beta}_0) + a_{\boldsymbol{\beta}}^T(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \quad (9.23)$$

where $a_{\boldsymbol{\beta}}(\cdot)$ is the vector of partial derivatives of $a(\cdot)$ with respect to the elements of $\boldsymbol{\beta}$. As

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}[\boldsymbol{\beta}_0, \hat{\sigma}^2 \{\mathbf{X}^T(\hat{\boldsymbol{\beta}})\mathbf{W}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})\mathbf{X}(\hat{\boldsymbol{\beta}})\}^{-1}],$$

(9.23) suggests that

$$a(\hat{\boldsymbol{\beta}}) \sim \mathcal{N}[a(\boldsymbol{\beta}_0), \hat{\sigma}^2 a_{\boldsymbol{\beta}}^T(\hat{\boldsymbol{\beta}}) \{\mathbf{X}^T(\hat{\boldsymbol{\beta}})\mathbf{W}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})\mathbf{X}(\hat{\boldsymbol{\beta}})\}^{-1} a_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})].$$

This result may be used as the basis for Wald-type confidence intervals for $a(\boldsymbol{\beta}_0)$, i.e.

$$a(\hat{\boldsymbol{\beta}}) \pm c_{\alpha/2} \hat{\sigma} [a_{\boldsymbol{\beta}}^T(\hat{\boldsymbol{\beta}}) \{\mathbf{X}^T(\hat{\boldsymbol{\beta}})\mathbf{W}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})\mathbf{X}(\hat{\boldsymbol{\beta}})\}^{-1} a_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})]^{1/2},$$

and a test statistic regarding $a(\boldsymbol{\beta})$.

See Seber and Wild (1989, Chapter 5) for more on alternative approaches to inference for nonlinear functions of parameters. As with the regression parameters themselves, Wald inference is routine in this context because of the relative ease of implementation.

9.6 Optimality of GLS and extensions

We have noted previously that

- GLS ($C = \infty$) is maximum likelihood estimation in the class of scaled exponential family distributions for $Y_j | \mathbf{x}_j$. Here, there are no additional variance parameters $\boldsymbol{\theta}$.
- GLS ($C = \infty$) estimation of $\boldsymbol{\beta}$ is equivalent to normal theory ML when the variance function *does not* depend on $\boldsymbol{\beta}$. Here, if the variance function depends on unknown parameters $\boldsymbol{\theta}$, jointly solving the GLS and PL estimating equations leads to joint normal ML for $(\boldsymbol{\beta}^T, \sigma, \boldsymbol{\theta}^T)$.

Thus, there are several situations in which the GLS approach is “optimal” in the sense that, in general, maximum likelihood estimation yields the most precise estimator (in terms of asymptotic relative efficiency). Of course, this is only true if the mean-variance model *and* the distributional assumption are *exactly* correct.

We have also argued that, even in the case where we are unwilling to make distributional assumptions, the GLS approach is “sensible,” as it weights the (linear) contributions of responses in accordance with their quality (as dictated by the assumed variance function).

It turns out that we may be more formal about this. In particular

- We will show momentarily that GLS arises naturally as the “optimal” approach among the class of all possible *linear estimating equations* for β .
- Because linear estimating equations depend on the data in a fairly simple way and are reasonably easy to solve, they are an obvious practical choice relative to more complicated quadratic or other equations. Thus, the “optimality” result ensures that using GLS will yield the most precise estimation within this practical class.

In fact, the argument we are about to carry out has broader implications for other types of equations, e.g., quadratic equations; we will discuss this in Chapter 10.

ASYMPTOTIC GAUSS-MARKOV PROPERTY: We have already derived some results related to the one we now consider, showing in Section 9.3 that GLS yields a more precise estimator for β in a large sample sense than estimators constructed with an incorrect variance model (including OLS). The following argument subsumes that one. In particular, we now show that the GLS equation leads to the estimator for β with the “smallest” large sample covariance matrix among *all* linear estimating equations.

To simplify the calculations, we will consider θ as known and thus write θ_0 ; because from the folklore theorem the effect of $\hat{\theta}$ in the weights is negligible asymptotically, replacing θ by $\hat{\theta}$ will not alter the result.

Suppose that the true variance model is $\text{var}(Y_j|\mathbf{x}_j) = \sigma^2 g^2(\beta, \theta, \mathbf{x}_j)$, and define the matrix $\mathbf{W}(\beta, \theta)$ as before. Then we may write the GLS equation with the correct variance model in matrix form as

$$\mathbf{X}^T(\beta)\mathbf{W}(\beta, \theta_0)\{\mathbf{Y} - \mathbf{f}(\beta)\} = \mathbf{0}. \quad (9.24)$$

Consider the general class of linear estimating equations for β of the form

$$\mathbf{A}^T(\beta)\{\mathbf{Y} - \mathbf{f}(\beta)\} = \mathbf{0} \quad (9.25)$$

of which it is clear that (9.24) is a special case. Let $\hat{\beta}$ denote the estimator for β satisfying (9.24), and let $\tilde{\beta}$ denote the estimator solving (9.25). Obviously, (9.25) is an unbiased estimating equation, so we expect that $\tilde{\beta} \xrightarrow{p} \beta_0$.

We will now show that $\hat{\beta}$ is “best” among all possible $\tilde{\beta}$. That is, we will show that all linear functions of $\tilde{\beta}$ have asymptotic covariance at least as great as that of $\hat{\beta}$ (in the sense of nonnegative definiteness we have described previously). Thus, in a large sample sense, $\hat{\beta}$ is optimal among all estimators solving linear estimating equations in the class (9.25).

Let $\mathbf{X} = \mathbf{X}(\beta_0)$, $\mathbf{W} = \mathbf{W}(\beta_0, \theta_0)$, $\mathbf{A} = \mathbf{A}(\beta_0)$, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T = \sigma_0^{-1} \mathbf{W}^{1/2} \{\mathbf{Y} - \mathbf{f}(\beta_0)\}$. From the folklore argument, multiplying through by $n^{-1/2}$ and using matrix notation, we may represent $\hat{\beta}$ solving (9.24) as

$$\hat{\beta} - \beta_0 \approx \sigma_0 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2} \boldsymbol{\epsilon}. \quad (9.26)$$

Now $\tilde{\beta}$ satisfies

$$\mathbf{A}^T(\tilde{\beta})\{\mathbf{Y} - \mathbf{f}(\tilde{\beta})\} = \mathbf{0};$$

thus, letting $\mathbf{a}_j^T(\beta)$ denote the j th row of $\mathbf{A}(\beta)$, by a Taylor series approximation we have

$$\begin{aligned} \mathbf{0} &\approx n^{-1/2} \sum_{j=1}^n \{Y_j - f(\mathbf{x}_j, \beta_0)\} \mathbf{a}_j(\beta_0) \\ &\quad + \left[-n^{-1} \sum_{j=1}^n \mathbf{a}_j(\beta_0) f_{\beta}(\mathbf{x}_j, \beta_0) + n^{-1} \sum_{j=1}^n \{Y_j - f(\mathbf{x}_j, \beta_0)\} \mathbf{a}_{j\beta}(\beta_0) \right] n^{1/2} (\tilde{\beta} - \beta_0). \end{aligned}$$

Analogous to the folklore argument, the second part of the linear term converges in probability to zero; thus, rearranging and multiplying through by $n^{-1/2}$, we obtain in matrix notation

$$\tilde{\beta} - \beta_0 \approx \sigma_0 (\mathbf{A}^T \mathbf{X})^{-1} \mathbf{A}^T \mathbf{W}^{-1/2} \boldsymbol{\epsilon}. \quad (9.27)$$

From (9.26) and (9.27), we thus have that

$$\begin{aligned} \text{var}(\hat{\beta} - \beta_0) &\approx \sigma_0^2 (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}, \\ \text{var}(\tilde{\beta} - \beta_0) &\approx \sigma_0^2 (\mathbf{A}^T \mathbf{X})^{-1} (\mathbf{A}^T \mathbf{W}^{-1} \mathbf{A}) (\mathbf{X}^T \mathbf{A})^{-1}. \end{aligned}$$

We now would like to show that the approximation to $\text{var}(\hat{\beta} - \beta_0)$ is “smaller” than that to $\text{var}(\tilde{\beta} - \beta_0)$; that is, show that the matrix difference

$$(\mathbf{A}^T \mathbf{X})^{-1} (\mathbf{A}^T \mathbf{W}^{-1} \mathbf{A}) (\mathbf{X}^T \mathbf{A})^{-1} - (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$$

is nonnegative definite.

The argument is entirely similar to that on page 209: We wish to show that

$$\boldsymbol{\lambda}^T \{(\mathbf{A}^T \mathbf{X})^{-1}(\mathbf{A}^T \mathbf{W}^{-1} \mathbf{A})(\mathbf{X}^T \mathbf{A})^{-1} - (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}\} \boldsymbol{\lambda} \geq 0$$

for all $\boldsymbol{\lambda}$. Letting $\mathbf{d} = (\mathbf{A}^T \mathbf{X})^{-1} \boldsymbol{\lambda}$, we may write this as

$$\mathbf{d}^T \{\mathbf{A}^T \mathbf{W}^{-1} \mathbf{A} - \mathbf{A}^T \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{A}\} \mathbf{d}.$$

Defining $\mathbf{c} = \mathbf{W}^{-1/2} \mathbf{A} \mathbf{d}$ and $\mathbf{X}_* = \mathbf{W}^{1/2} \mathbf{X}$, we may rewrite this as

$$\mathbf{c}^T \{\mathbf{I} - \mathbf{X}_*(\mathbf{X}_*^T \mathbf{X}_*)^{-1} \mathbf{X}_*^T\} \mathbf{c} = \mathbf{b}^T \mathbf{b} \geq 0,$$

as the middle matrix is symmetric and idempotent.

The result is sometimes shown alternatively as follows. Writing for convenience $\mathbf{L} = (\mathbf{A}^T \mathbf{X})^{-1} \mathbf{A}^T$ and $\mathbf{f} = \mathbf{f}(\boldsymbol{\beta}_0)$, we have

$$\begin{aligned} \text{var}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\approx \text{var}\{\mathbf{L}(\mathbf{Y} - \mathbf{f})\} = \text{var}\{\mathbf{L}(\mathbf{Y} - \mathbf{f}) - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\} \\ &= \text{var}\{\mathbf{L}(\mathbf{Y} - \mathbf{f}) - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\} + \text{var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \text{cov}\{\mathbf{L}(\mathbf{Y} - \mathbf{f}) - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\}. \end{aligned}$$

Using (9.26) and (9.27), the covariance term satisfies

$$\begin{aligned} &\sigma_0 \text{cov}\{\mathbf{L} \mathbf{W}^{-1/2} \boldsymbol{\epsilon} - (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2} \boldsymbol{\epsilon}, (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2} \boldsymbol{\epsilon}\} \\ &\sigma_0^2 \{\mathbf{L} \mathbf{W}^{-1} - (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T\} \mathbf{W} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \\ &\sigma_0^2 \{\mathbf{L} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} - (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}\} \\ &= \mathbf{0} \quad \text{using } \mathbf{L} \mathbf{X} = \mathbf{I}. \end{aligned}$$

Thus,

$$\text{var}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approx \text{var}\{\mathbf{L}(\mathbf{Y} - \mathbf{f}) - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\} + \text{var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

As the first term is a covariance matrix, it must be nonnegative definite. Thus, we may conclude that

$$\text{var}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \text{ " } \geq \text{ " } \text{var}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

in the sense of nonnegative definiteness.

RESULT:

- Among all estimators solving linear estimating equations for $\boldsymbol{\beta}$, not just those involving the gradient matrix $\mathbf{X}(\boldsymbol{\beta})$ and incorrect weighting, the GLS estimator is optimal in the sense of being the most precise in a large sample sense.

- Thus, the GLS approach may be motivated based on this appealing property with no mention of distributions or an “intuitive” connection to the “sensible” idea of WLS.
- This result is consistent with what we already know. When the variance function g does not depend on β , the normal ML equation and the GLS equation are identical. Thus, if the data *truly* are (conditionally) normally distributed, then the general maximum likelihood theory tells us that solving this linear estimating equation is the optimal approach in a large sample sense. Hence, in this situation, we would expect that the optimal linear estimator would correspond to maximum likelihood, which also involves a linear equation.

It is of course important to remember that the above result is predicated on having specified the variance model correctly. Thus, the optimal linear estimating equation involves the true (conditional) variance of the response. As we have already seen, misspecification of this variance leads to inefficiency.

CONJECTURE: The GLS equation depends on the form we highlighted first in Chapter 5 of

$$\{\text{gradient of mean function}\} \times \{\text{covariance matrix}\}^{-1} \times \{\text{response} - \text{mean}\}.$$

Recall that we have cast a number of equations in this general form as

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

(see Section 6.4), which may be written more compactly in matrix notation using the definitions in Section 6.4 as

$$\mathbf{D}^T(\boldsymbol{\alpha}) \mathbf{V}^{-1}(\boldsymbol{\alpha}) \{\mathbf{s}(\boldsymbol{\alpha}) - \mathbf{m}(\boldsymbol{\alpha})\} = \mathbf{0}, \quad (9.28)$$

where $\mathbf{D}(\boldsymbol{\alpha})$ is the “gradient matrix” of the “mean vector” $\mathbf{m}(\boldsymbol{\alpha})$ for the “response vector” $\mathbf{s}(\boldsymbol{\alpha})$ with “covariance matrix” $\text{var}\{\mathbf{s}(\boldsymbol{\alpha})\} = \mathbf{V}(\boldsymbol{\alpha})$. Here, $\mathbf{s}(\boldsymbol{\alpha})$ is some function of the data and parameters; we considered in Chapters 5 and 6

$$\mathbf{s}_j(\boldsymbol{\alpha}) = \begin{pmatrix} Y_j \\ \{Y_j - f(\mathbf{x}_j, \boldsymbol{\beta})\}^2 \end{pmatrix}$$

for quadratic estimating equations.

It is natural to wonder whether a similar result holds for equations of the form (9.28) more generally. That is, with the components of (9.28) defined as above, if we considered any other equation, in obvious notation,

$$\mathbf{A}^T(\boldsymbol{\alpha}) \{\mathbf{s}(\boldsymbol{\alpha}) - \mathbf{m}(\boldsymbol{\alpha})\} = \mathbf{0},$$

it would be interesting to find that the estimator solving (9.28) is at least as precise asymptotically.

It turns out that, under certain conditions, this is indeed the case: the optimal estimating equation based on a linear combination of some function of the data $\mathbf{s}(\boldsymbol{\alpha})$ minus its mean is of the form (9.28) with the matrices $\mathbf{D}(\boldsymbol{\alpha})$ and $\mathbf{V}(\boldsymbol{\alpha})$ correctly specified.

ILLUSTRATION: Consider joint normal theory ML estimation of σ and $\boldsymbol{\beta}$ in under the mean-variance model (9.1) with $\boldsymbol{\theta}$ known. Then we identify $\boldsymbol{\alpha} = (\boldsymbol{\beta}^T, \sigma)^T$, and from (5.19) the estimating equation is, in the previous shorthand notation,

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

If the Y_j given \mathbf{x}_j are truly normally distributed, these equations yield the asymptotically optimal estimator for $\boldsymbol{\beta}$ on the basis of general maximum likelihood theory. The above argument suggests that the equations are also “optimal” among all such equations under the conditions that the mean and variance model are correct and that the third and fourth moment assumptions for the $Y_j|\mathbf{x}_j$ that appear in the “covariance matrix” are correct, even if the data are not normal.

Similarly, if we believe that the ϵ_j are i.i.d. and symmetrically distributed, but that $\text{var}(\epsilon_j^2) = 2 + \kappa$ for all j , then the argument would suggest that the above equation would lead to an inefficient estimator. The argument would suggest that the estimating equation

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

discussed on page 115 would lead to the most precise estimation under these circumstances.

In Chapter 10, we will consider these issues more carefully. A major interest will be to investigate potential gains in efficiency by using quadratic rather than linear estimating equations and to appreciate the tradeoffs involved.