Chapter 4

Semi-Parametric Estimation in the Linear Model for Correlated Data

In this chapter we consider inference for the parameters of the LMCD without the assumption of multivariate normal errors. We will present a general theory for weighted least squares (WLS) estimators of β, which are consistent and asymptotically normal (CAN), using only the assumption that $E(Y_i) = X_i\beta$. The estimators will be asymptotically efficient as well, provided a consistent estimate of $\text{var}(Y_i)$ is available. We discuss the asymptotic distribution of WLS (and also ML) estimators when $\text{var}(Y_i)$ is misspecified. For simplicity, we will first consider the case where any imbalance is by design so that the expectation $E(Y_i) = X_i\beta$ holds for the observed data and each n_i is fixed. We will discuss imbalance due to missingness at the end of this chapter. We also assume that each $\text{var}(Y_i) = \Sigma_i$ is a function of a finite parameter vector θ. Furthermore, to simplify the presentation we restrict attention to the setting in which the covariates are stochastic, although most of the results described in this section are valid under appropriate regularity conditions when the covariates are assumed to be fixed numbers.

4.1 Weighted Least Squares Estimators of β

A weighted least squares estimator $\widehat{\beta}$ is defined as the value of β that minimizes the objective function

$$Q_W(\beta) = \sum_{i=1}^{N} (Y_i - X_i\beta)^T W_i (Y_i - X_i\beta)$$

where $W_i = W(X_i)$, $i = 1, \ldots, N$, is an arbitrary positive definite and symmetric $n_i \times n_i$ matrix chosen by the investigator. If a minimum of $Q_W(\beta)$ exists, it must solve

$$\frac{\partial Q_W(\beta)}{\partial \beta} = -2 \sum_{i=1}^{N} X_i^T W_i (Y_i - X_i\beta) = 0.$$

This equation has a unique solution at

$$\widehat{\beta} = \left(\sum_{i=1}^{N} X_i^T W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i^T W_i Y_i. \tag{4.1}$$

Because each W_i is positive definite, $Q_W(\beta)$ has a positive definite Hessian at $\widehat{\beta}$, and $Q_W(\beta)$ achieves its global minimum at $\widehat{\beta}$.

The structure of $\widehat{\beta}$ simplifies for special choices of weight matrices. For example, if $W_i \propto I$ for any constant σ, then

$$Q_I(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{n} (Y_{ij} - X_{ij}\beta)^2$$

and therefore $\widehat{\beta}$ minimizing $Q_I(\beta)$ is the ordinary least squares estimator of β based on the data $(Y_{ij}, X_{ij}), i = 1, \ldots, N$, $j = 1, \ldots, n_i$. If in addition, $X_i = X$ for all i, such as for example in the polynomial growth curve model of Examples 3 and 4 in Section 1.2, then $\widehat{\beta}$ simplifies to

$$\widehat{\beta} = (X^T X)^{-1} X^T \overline{Y}$$

where $\overline{Y} = \Sigma Y_i / N$. Thus, $\widehat{\beta}$ agrees with the ordinary least squares estimates of the parameters in the regression of the sample means at each occasion on the rows of X.

4.2 Properties of the Weighted Least Squares Estimator

The estimator $\widehat{\beta}$ is unbiased for any choice of weight function because its conditional expectation given $X_1, \ldots, X_N$ satisfies

$$E\left(\widehat{\beta}\right) = \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i^T W_i X_i \beta = \beta. \tag{4.2}$$

Furthermore, by the Gauss–Markov Theorem the weighted least squares estimator that uses $W_i = \Sigma_i^{-1}$ has the smallest conditional variance given the covariates (in the positive definite sense) among the conditional variances of all estimators that are solutions of equations of the form (4.1) for arbitrary choices of W_i. That is, letting $\widehat{\beta}(W)$ and $\widehat{\beta}\left(\Sigma^{-1}\right)$ denote the solutions of the estimating equation (4.1) with W_i arbitrary and with $W_i = \Sigma_i^{-1}$ respectively, we have that

$$\text{var}\left\{\widehat{\beta}\left(\Sigma^{-1}\right)\right\} \leq \text{var}\left\{\widehat{\beta}(W)\right\} \tag{4.3}$$

where for any squared matrices A and B, $A \leq B$ if and only if $B - A$ is semipositive definite. The proof of this is straightforward; it involves using the fact that $\text{var}[\widehat{\beta}(\Sigma^{-1}) - \widehat{\beta}(W)] \geq 0$.

Notice that the conditional variance of $\widehat{\beta}$ given $X_1, \ldots, X_N$ is

$$\text{var}\left(\widehat{\beta}\right) = \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i^T W_i \Sigma_i W_i X_i\right) \left(\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1};$$

it reduces to

$$\text{var}\left(\widehat{\beta}\right) = \left(\sum_{i=1}^{N} X_i^T X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i^T \Sigma_i X_i\right) \left(\sum_{i=1}^{N} X_i^T X_i\right)^{-1}$$

when $W_i = I$ and it reduces to

$$\text{var}\left(\widehat{\beta}\right) = \left(\sum_{i=1}^{N} X_i^T \Sigma_i^{-1} X_i\right)^{-1}$$

when $W_i = \Sigma_i^{-1}$.

By (4.2), the estimator that uses $W_i = I$, i.e., the ordinary least squares estimator of β, is unbiased. However, with correlated data, its variance no longer agrees with the one obtained under independence of

the repeated observations made on each subject and it depends on the Σ_i's. Calculating the variance of $\widehat{\beta}$ incorrectly assuming independence can, of course, result in Wald tests of β that do not preserve their nominal level. The Gauss Markov Theorem implies that the choice $W_i = I$ may not be optimal. The optimal choice depends on the unknown covariance matrices Σ_i. This suggests that for optimality purposes one could replace the unknown optimal weight Σ_i by an estimator of it and warrants the study of weighted least squares estimators calculated using weight functions that are computed from the data.

4.3 Weighted Least Squares with Data-Dependent Weight Functions

Suppose that the dependence of the weights W_i on X_i can be written as

$$W_i = W(\theta; X_i)$$

for some fixed $q \times 1$ parameter vector θ. In Section 4.2 we have assumed that W_i was a known function of X_i so that θ was fixed and known. In this section we investigate the properties of weighted least squares estimators that use data-dependent weight functions where θ is replaced by a value $\widehat{\theta}$ computed from the sample. Notice that this formulation includes, but is not restricted to, estimators that use as weight W_i an estimator $\widehat{\Sigma}_i^{-1}$ of Σ_i^{-1}. This particular application is obtained by taking $W(\theta; X_i)$ a constant function of X_i and taking θ to be equal to the unknown parameters indexing the model for the covariance of Y_i.

Henceforth, let $\widehat{\beta}$ be the solution to the equation

$$\sum_{i=1}^{N} X_i^T \widehat{W}_i (Y_i - X_i\beta) = 0 \tag{4.4}$$

where $\widehat{W}_i = W(\widehat{\theta}; X_i)$. (Occasionally, when needed to stress the dependence of the estimator on the weights we will denote the solution of (4.4) by $\widehat{\beta}(\widehat{W})$.) The estimator $\widehat{\beta}$ is no longer necessarily unbiased because of the dependence of $\widehat{W}_i$ on the entire sample. However, under mild regularity conditions that include that $\widehat{\theta}$ has a probability limit θ^*, $W(\theta^*; X_i)$ is positive definite and $W(\theta; x)$ is a smooth function of θ and x, $\widehat{\beta}$ is a consistent estimator of β. Furthermore, $\sqrt{N}(\widehat{\beta} - \beta)$ converges as N goes to ∞ to a normal distribution with mean zero and variance C_W given by

$$C_W = \Gamma_W^{-1} \Omega_W \Gamma_W^{-1} \tag{4.5}$$

where

$$\Gamma_W = E\left\{X_i^T W\left(\theta^*; X_i\right) X_i\right\}$$

and

$$\Omega_W = E\left\{X_i^T W\left(\theta^*; X_i\right) \Sigma_i W\left(\theta^*; X_i\right) X_i\right\}.$$

See, for example, Newey and McFadden (1994), Theorems 6.1 and 6.2. The last result implies, in particular, that if $\widehat{\beta}(W)$ denotes the solution of (4.1) using $W_i = W(\theta^*; X_i)$, then the estimators $\widehat{\beta}(\widehat{W})$ and $\widehat{\beta}(W)$, appropriately normalized, have the same asymptotic distribution. This has the following interesting consequence. By (4.3), we have that

$$\text{var}\left[\sqrt{N}\left\{\widehat{\beta}\left(\Sigma\right) - \beta\right\}\right] \leq \text{var}\left[\sqrt{N}\left\{\widehat{\beta}\left(W\right) - \beta\right\}\right]$$

or equivalently,

$$\left(\frac{1}{N}\sum_{i=1}^{N} X_i^T \Sigma_i^{-1} X_i\right)^{-1}$$
$$\leq \left(\frac{1}{N}\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} X_i^T W_i \Sigma_i W_i X_i\right) \left(\frac{1}{N}\sum_{i=1}^{N} X_i^T W_i X_i\right)^{-1}.$$

Now when $W_i = \Sigma_i^{-1}$, $C_{\Sigma_i^{-1}} = E\{X_i^T \Sigma^{-1} X_i\}$. Thus, taking probability limit as N goes to infinity on both sides of the last inequality, and invoking the Law of Large Numbers we obtain

$$C_{\Sigma^{-1}} \leq C_W.$$

We conclude that the choice of weights $W_i = \Sigma_i^{-1}$ gives an asymptotically most efficient weighted least squares estimator, i.e., an estimator with asymptotic variance no greater than the asymptotic variance of any weighted least squares estimator. Furthermore, since $\widehat{\beta}(\widehat{W})$ and $\widehat{\beta}(W)$ have the same asymptotic distribution, then the estimator that uses as weights a consistent estimator of $\text{var}(Y_i)$ is also an asymptotically efficient weighted least squares estimator.

This theory insures that for sufficiently large N, the effect of estimating the weight parameter θ is negligible. In particular, the weighted least squares estimate of β has the same asymptotic variance regardless of whether or not the weights are estimated or they are fixed known constants. However, in practice, if N is not too large, estimation of the weights may have an important effect on the actual variance of $N^{1/2}(\widehat{\beta} - \beta)$. Many theoretical studies for the case of Y_i a scalar outcome support this assertion. See, for example, Freedman and Peters (1984),

Carroll, Wu and Ruppert (1988), and reference therein. Carroll *et al.* (1988) examined in detail a special case in which $W(\theta; X_i)$ is the inverse of a correctly specified model for the variance function, $W(\theta; X_i)$ that does not depend on the mean, the errors have a symmetric distribution and θ is estimated by maximum likelihood assuming normal errors. They showed that in this case, $\widehat{C}_W$ is an underestimate of the actual variance of $N^{1/2}(\widehat{\beta} - \beta)$ if the kurtosis of the error distribution is less than or equal 2 but it is an overestimate otherwise. This is in accordance to the results of Freedman and Peters (1984) who noted the decrease in variance for heavy tail error distributions. Carroll *et al.* (1988) showed that the bootstrap variance estimate provides a more refined approximation to the actual variance of $N^{1/2}(\widehat{\beta} - \beta)$.

REMARK 1. Notice that the ML estimate takes exactly the same form as $\widehat{\beta}(\widehat{W})$, where $\widehat{\theta} = \widehat{\theta}_{\mathrm{ML}}$, and $\mathrm{var}(Y_i) = W_i(\widehat{\theta}, X_i)^{-1}$. Hence $\widehat{\beta}_{\mathrm{ML}}$ is asymptotically optimal, provided $\mathrm{var}(Y_i)$ is correctly specified. Otherwise $\widehat{\beta}_{\mathrm{ML}}$ is an ordinary WLS estimate with estimated weight matrix, and the asymptotic normal distribution previously given holds, i.e.,

$$\sqrt{N}\left(\widehat{\beta}_{\mathrm{ML}} - \beta\right) \sim N\left(0, C_W\right),$$

where C_W is given in (4.5).

REMARK 2. If the conditions given in Kackar and Harville (1981) hold, i.e., the variance-covariance component estimators are even and translation invariant and the error distributions for $(Y_i - X_i\beta)$ are symmetric, then WLS estimates with data-dependent weight functions are also unbiased for β in small samples.

4.4 Estimation of the Optimal Weight Function

Suppose we have complete data and $\Sigma_i = \Sigma$. A consistent estimator of the optimal weight Σ^{-1} can be obtained as follows. Suppose that $\widetilde{\beta} = \widehat{\beta}(W)$ is a weighted least squares estimator of β for an arbitrary (but fixed and known) choice of weight function. For example, $\widetilde{\beta}$ is the ordinary least squares estimator obtained from the choice $W_i = I$. Then, under mild regularity conditions on the weight function

$$\widehat{\Sigma} = S = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - X_i\widetilde{\beta}\right)\left(Y_i - X_i\widetilde{\beta}\right)^T \tag{4.6}$$

is a consistent estimator of Σ (see, e.g., Newey and McFadden, 1994, Lemma 4.3). Note that $\widehat{\Sigma}$ is a one-step EM estimate, away from $\widehat{\Sigma}^0 = I$. See equation (3.11) and following.

The estimator $\widehat{\Sigma}$ separately estimates $n(n-1)/2$ non-diagonal and n diagonal elements. If n is large then a large sample size N is typically required for the asymptotic distribution of $\widehat{\beta}(\widehat{\Sigma})$ to be a good approximation of its finite sample distribution, and $\widehat{\beta}(\widehat{\Sigma})$ may have poor small sample behavior. Asymptotically efficient weighted least squares estimators with better small sample properties can be obtained under a parsimonious model for $\mathrm{cov}(Y_i)$, i.e., when Σ is a known function $\Sigma(\theta)$ of a parameter vector θ of dimension $q \ll n(n+1)/2$. We now describe consistent estimators of θ under other models $\Sigma(\theta)$ that were introduced in Section 1.3. In what follows $\widetilde{\beta}$ is the weighted least squares estimator calculated using an arbitrary, but fixed and known, weight function.

Compound Symmetry. Recall from Section 1.3, Example 1, that Σ has a compound symmetry structure when $\Sigma = \sigma_1^2 I + \kappa 1 1^T$. Thus, letting $\theta_1 = \sigma_1^2 + \kappa$ and $\theta_2 = \kappa$, $\theta = (\theta_1, \theta_2)$, $\Sigma = \Sigma(\theta)$ is a symmetric matrix with all its diagonal elements equal to θ_1 and all its non-diagonal elements equal to κ. A consistent estimator of θ_1 is obtained by averaging the diagonal elements of $\widehat{\Sigma}$ defined in (4.6). Similarly, a consistent estimator of κ is obtained by averaging its non-diagonal elements. Thus,

$$\widehat{\theta}_1 = \frac{1}{n}\left[\sum_{j=1}^{n} S_{jj}\right]$$

and

$$\widehat{\theta}_2 = \frac{2}{n(n-1)}\left[\sum_{1 \le j < k \le n} S_{jk}\right],$$

where S_{jj} and S_{jk} are elements of S defined in (4.6).

Banded. Recall from Section 1.3 that Σ has a banded structure if its (j,k)th and (j',k')th entries are equal when $j' - j = k' - k$. Thus, here $\Sigma = \Sigma(\theta)$ where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ are the entries of Σ that vary freely. Using the notation of Section 1.3, $\theta_1 = \sigma^2$ and $\theta_{l+1} = \sigma^2 \rho_l$, $1 \le l \le n-1$. These can be consistently estimated by averaging the elements of $\widehat{\Sigma}$ defined in (4.6) over the entries known to be identical under the banded structure. Thus, for example, $\widehat{\theta}_1$ is the same as for the compound symmetry model, while for $1 \le l \le n-1$,

$$\widehat{\theta}_{l+1} = \frac{2\widehat{\theta}_1}{(n-l)}\left[\sum_{k=1}^{n-l}\left(S_{k(l+k)}\right)\right] \tag{4.7}$$

REMARK. Maximum likelihood or REML estimates of θ will also be consistent for θ in this setting, even without the assumption of normality. They can be computed as described in 3.

4.5 Locally Optimal Weighted Least Squares

Once the estimator $\Sigma(\widehat{\theta})$ of Σ is obtained, estimation of β proceeds by solving the weighted least squares equation using the weight matrices $\Sigma_i(\widehat{\theta})^{-1}$. The estimation procedure is then a three stage procedure summarized as follows:

1. Obtain a preliminary weighted least squares estimator $\widetilde{\beta}$ (e.g., the ordinary least squares estimator),
2. Estimate Σ with any consistent estimate $\widehat{\Sigma}$ (as defined in (4.6) for example) or with $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$ where $\widehat{\theta}$ is a consistent estimator of θ under a model for $\text{cov}(Y_i)$,
3. Obtain the solution $\widehat{\beta}(\widehat{\Sigma})$ to the weighted least squares estimating equation (4.4) with weights $\widehat{W}_i = \widehat{\Sigma}_i$.

The estimator $\widehat{\beta}(\widehat{\Sigma})$ is consistent and asymptotically normal. We say that this estimator is *a locally optimal weighted least squares (LOWLS) estimator* because it has asymptotic variance that is equal to the lower bound for the variances of all weighted least squares estimators when the model for $\text{cov}(Y_i)$ is correctly specified and $\widehat{\beta}(\widehat{\Sigma})$ remains consistent and asymptotically normal even if the model for $\text{cov}(Y_i)$ is misspecified. Notice that the LOWLS makes no reference to the distribution of the error terms. In general, $\widehat{\beta}(\widehat{\Sigma})$ will be close to the ML estimate of β obtained by assuming normality, and the ML estimate has the same properties in this setting.

4.6 Model Based and Robust Variance Estimation

In this section we consider estimation of the asymptotic variance of $\widehat{\beta}(\widehat{\Sigma}^{-1})$. By the results of Section 4.3, $\sqrt{N}\{\widehat{\beta}(\widehat{\Sigma}^{-1}) - \beta\}$ has an asymptotic mean zero normal distribution with variance equal to $C_{\Sigma^{-1}}$ provided $\widehat{\Sigma}$ is a consistent estimator of Σ, i.e., provided the model used to estimate

$\text{cov}(Y_i)$ is correctly specified. When this is the case, a consistent estimator of $C_{\Sigma^{-1}}$, and therefore of the asymptotic variance of $\sqrt{N}\{\widehat{\beta}(\widehat{\Sigma}^{-1})-\beta\}$ is given by

$$\frac{1}{N}\sum_{i=1}^{N} X_i^T\widehat{\Sigma}^{-1}X_i. \tag{4.8}$$

If the model for $\text{cov}(Y_i)$ was incorrectly specified, then by the results of Section 2.3, $\sqrt{N}\{\widehat{\beta}(\widehat{\Sigma}^{-1})-\beta\}$ has the same asymptotic variance as $\sqrt{N}\{\widehat{\beta}(\Sigma^{*^{-1}})-\beta\}$ where Σ^* is the probability limit of $\widehat{\Sigma}$. This asymptotic variance is equal to

$$C_{\Sigma^{*-1}} = \Gamma^{-1}_{\Sigma^{*-1}}\Omega_{\Sigma^{*-1}}\Gamma^{-1}_{\Sigma^{*-1}}.$$

A consistent estimator of $\Gamma_{\Sigma^{*-1}}$ is given by

$$\widehat{\Gamma}_{\Sigma^{*-1}} = \frac{1}{N}\sum_{i=1}^{N} X_i^T\widehat{\Sigma}^{-1}X_i,$$

and a consistent estimator of $\Omega_{\Sigma^{*-1}}$ is given by

$$\widehat{\Omega}_{\Sigma^{*-1}} = \frac{1}{N}\sum_{i=1}^{N} X_i^T\widehat{\Sigma}^{-1}\left(Y_i - X_i\widehat{\beta}\right)\left(Y_i - X_i\widehat{\beta}\right)^T\widehat{\Sigma}^{-1}X_i.$$

Thus, a consistent estimator of $C_{\Sigma^{*-1}}$ is given by

$$\widehat{\Gamma}^{-1}_{\Sigma^{*-1}}\widehat{\Omega}_{\Sigma^{*-1}}\widehat{\Gamma}^{-1}_{\Sigma^{*-1}}. \tag{4.9}$$

The variance estimator (4.8) is often referred to as a *model based* variance estimator to remind the user that its consistency relies on the correct specification of the model for $\text{cov}\,(Y_i)$. The variance estimator (4.9) is often referred to as a *robust* variance estimator since it is consistent even under misspecification of the model for $\text{cov}(Y_i)$. It has also been referred to as the sandwich or empirical variance estimator (Liang and Zeger, 1986). The usual trade off between bias and variance applies to these two choices. The sandwich variance estimator is asymptotically unbiased regardless of model specification. However, even with large samples it will typically have bigger fluctuations around the true asymptotic variance than the model based variance estimator unless the assumed model is far from the true variance. Also, under correct model specification the model based variance estimator will typically have smaller mean squared error than the robust variance estimator.

The variance estimators can be used to construct Wald tests of constraint null hypotheses of the form

$$H_0 : \underset{q\times p}{Q^T}\ \underset{p\times 1}{\beta} = 0.$$

Specifically, let $\widehat{\beta}$ denote $\widehat{\beta}\left(\widehat{\Sigma}^{-1}\right)$. Under the null hypothesis, we have

$$\sqrt{N}Q^T\widehat{\beta} \xrightarrow{\mathcal{L}} N_q\left(0; Q^T C_{\Sigma^{*-1}} Q\right)$$

then

$$\widehat{\beta}^T Q\left(Q^T \widehat{C}_{\Sigma^{*-1}} Q\right)^{-1} Q^T\widehat{\beta} \tag{4.10}$$

has an asymptotic χ^2_q distribution. A test that rejects when (4.10) exceeds the α point of the χ^2_q distribution is a valid asymptotic α level test of H_0.

REMARK. The results in this section can be applied straightforwardly to obtain $\text{var}(\widehat{\beta}_{\text{ML}})$ when the form of $\text{var}(Y_i)$ is unknown, and some user-specified model has been selected for $\Sigma_i(\theta)$.

4.7 Joint Estimation of β and θ

The development in Sections 4.1-4.6 was aimed at inference about the parameter β. In Section 4.4 the emphasis was on estimation of $\text{cov}(Y_i)$ for improvement of the efficiency of the weighted least squares estimator of β. Notice, for example, that although easy to obtain, in the cases described in Section 4.4, we have not given formulas for consistent estimators of the variance of $\widehat{\theta}$. In fact, nowhere in the derivation of consistency and asymptotic normality of the weighted least squares estimators of β is it required that a model for cov (Y_i) be specified. That is, the WLS estimators are consistent and asymptotically normal (CAN) under the sole assumption that the conditional mean of Y_i is linear in the covariates, i.e., that

$$E\left(Y_i\right) = X_i^T\beta. \tag{4.11}$$

This has the advantage that inferences about β are robust to misspecification of the model for $\text{cov}(Y_i)$. However, even if inference about β was the sole primary goal of the analysis, the WLS approach described above has the drawback that under most error distributions, knowledge of the model for $\text{cov}(Y_i)$, when available, can be used to improve the efficiency with which β is estimated. We can obtain efficiency improvements

over the weighted least squares estimators described above by further enlarging the class of estimating equations and calculating the optimal estimator under the larger class. This can be effectively done for many distributions, by solving joint estimating equations for β and θ. In addition, as a by-product of jointly estimating β and θ, consistent variance estimator formulae for θ can be readily derived from the standard theory of estimating equations, therefore providing the possibility of conducting inference about θ as well. We postpone the discussion of semi-parametric, joint estimation of the mean and second moment parameters to Chapter 6, where we will discuss the generalized linear multivariate model.

4.8 Efficiency of OLS Estimators

It is often suggested that using $W_i = I$ (or using OLS) leads to an estimator with nearly the same efficiency as the optimal estimate using Σ_i^{-1}, in many cases. The claim can be justified by considering simple forms for Σ, eg compound symmetry, or banded, with constant variance, and/or simple forms for each X_i. Bloomfield and Watson (1975) gave a general formula for relative efficiency of $\widehat{\beta}(I)$ and $\widehat{\beta}(\Sigma^{-1})$, by comparing

$$\left(\sum_{i=1}^{N}(X_i^T X_i)\right)^{-1}\left(\sum_{i=1}^{N} X_i^T \Sigma_i X_i\right)\left(\sum_{i=1}^{N}(X_i^T X_i)\right)^{-1}$$

with

$$\left(\sum_{i=1}^{N}\left(X_i^T \Sigma_i^{-1} X_i\right)\right)^{-1}.$$

We consider some examples given in Diggle *et al.* (1994).

Example 1.

$$N = 10 \;\; n = 5 \;\; t_j = (-2, -1, 0, 1, 2)$$

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_j.$$

If there are no missing observations and $\Sigma = \sigma^2\left((1-\rho)I + \rho J\right)$, then $\widehat{\beta}(\Sigma^{-1}) = \widehat{\beta}(I)$ and OLS is optimal. This requires a common design, no missing data and compound symmetry. As the next example shows, efficiency of OLS in this case is still high with an autoregressive structure.

Example 2.

Suppose the design on time remains the same, but Σ is of the form

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & 1 & \rho & \rho^2 & \rho^3 \\ & & 1 & \rho & \rho^2 \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$$

(still equal spacing and equal variances on diagonal). It is straightforward to show (Diggle *et al.*, 1994) that

$$\text{var } \hat{\beta}_{\text{OLS}} = \begin{bmatrix} 10(5+8\rho+6\rho^2+4\rho^3+2\rho^4) & 0 \\ 0 & 20(5+4\rho-\rho^2-4\rho^3-4\rho^4) \end{bmatrix}$$

and

$$\text{var } \hat{\beta}(\Sigma^{-1}) = \sigma^2(1-\rho^2) \begin{bmatrix} 0.1(5-8\rho+3\rho^2)^{-1} & 0 \\ 0 & 0.05(5-4\rho+\rho^2)^{-1} \end{bmatrix}.$$

The relative efficiencies can be obtained by looking at the ratio of the diagonals. They are computed for a range of ρ's below (taken from Diggle *et al.*, 1994):

ρ	0.1	0.2	0.3	0.4	0.5
$e(\beta_0)$	0.998	0.992	0.983	0.973	0.963
$e(\beta_1)$	0.997	0.989	0.980	0.970	0.962

ρ	0.6	0.7	0.8	0.9	0.99
$e(\beta_0)$	0.955	0.952	0.956	0.970	0.996
$e(\beta_1)$	0.952	0.955	0.952	0.955	0.961

These are all very close to 1, because $n_i = n$, $X_i = X$, and $\text{var}(Y_{ij})$ is constant.

Example 3.

A crossover design where carryover may be present (Fitzmaurice *et al.*, 1993). Here $n = 3$ and there are 2 treatments; subjects are assigned in equal numbers to all possible 2^3 sequences: AAA, AAB, ABA, ABB, BAA, BAB, BBA, BBB. Here

$$\text{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij}$$

where $t_{ij} = 1$ if ith subject gets A in the jth period. Note that there are eight distinct X_i's, one for each sequence. Assuming Σ has an autocorrelation structure, Diggle *et al.* (1994) obtain the following efficiency ratios:

ρ	0.1	0.2	0.3	0.4	0.5
$e(\beta_0)$	0.993	0.974	0.946	0.914	0.880
$e(\beta_1)$	0.987	0.947	0.883	0.797	0.692

ρ	0.6	0.7	0.8	0.9	0.99
$e(\beta_0)$	0.846	0.815	0.788	0.766	0.751
$e(\beta_1)$	0.571	0.438	0.297	0.150	0.015

The efficiency for estimating the treatment effect declines substantially for $\rho > 0.2$. An intuitive explanation for this is as follows. To estimate a treatment contrast, some individuals provide more information than others; eg., a person in sequence 1 and another in sequence 8 provide the contrast $y_{i1} - y_{i'1}$ to estimate $A - B$, with variance equal to $2\sigma^2$. Now consider within subject contrasts:

$$
\begin{array}{ll}
y_{i2} - y_{i3} & \text{for sequence 2} \\
y_{i1} - y_{i2} & \text{for sequence 3} \\
y_{i1} - y_{i2} & \text{for sequence 4} \\
\text{etc.} &
\end{array}
$$

Each has variance $2\sigma^2(1 - \rho)$, so if ρ is large, these within subject contrasts have *much more* information and should be weighted more heavily. Here correlation is important because of the design, variance is less so. The OLS ignores the design and weights all possible contrasts equally.

4.9 Remarks

The asymptotic properties of extremum estimators that minimize some objective function have been widely studied, among others by Fisher (1925), Wald (1949), Huber (1967), Jenrich (1969) and Amemiya (1973, 1985). White (1980) specialized these results to obtain robust variance estimators for the LOWLS estimators of Section 2.5. The LOWLS estimators are sometimes also referred to as the Generalized Estimating Equations (GEE) estimators after the papers of Liang and Zeger (1986)

and Zeger and Liang (1986). These papers considered estimating equations for a more general class of mean models that include generalized linear models. These authors call the model for $\text{cov}(Y_i)$ a "working covariance model" as it is indeed not needed for consistency and asymptotic normality of the LOWLS estimator. The GEE methodology will be taken up in Chapter 6.

The weighted least squares estimators are *semiparametric* estimators because they are CAN without requiring a full parametric description of the law of the data. In fact, they are CAN provided only that the model (4.11) for the conditional mean of the outcome vector is correctly specified. That is, consistency and asymptotic normality is obtained without specifying any restriction on the error distribution besides the conditional mean zero restriction.

4.10 Studies with Clusters of Random Size or Missing Data

In some designs, the cluster size may not be fixed in advance. Suppose for example N units are randomly drawn from a specific population. Each unit i, $i = 1, \ldots, N$, is composed of n_i members. On each member $j, j = 1, \ldots, n_i$, of each unit i, we observe an outcome Y_{ij} and an associated vector of covariates X_{ij}. The covariates X_{ij} may be cluster-specific, i.e., X_{ij} varies across units but not across members of the same unit so that X_{ij} is the same for all j, subject-specific, i.e., X_{ij} varies across members of the same unit, or both. Thus, $Y_i = (Y_{i1}, \ldots, Y_{in_i})'$ and $X_i = (X_{i1}, \ldots X_{in_i})'$ record the full data on the ith sampling unit. As another example, suppose we randomly sample N units from the population and intend to make n observations on each subject with $n \times p$ design matrix X_i^C. Suppose in fact we are able to obtain only n_i observations for each subject, but the missing observations are MCAR, so that for the observed data vector, $E(Y_i) = X_i\beta$, where the rows of X_i are the subset of rows of X_i^C corresponding to the observed Y_i.

The LMCD (1.1.2) specifies that the conditional mean of each observation Y_{ij} depends linearly on functions of cluster-specific and subject-specific covariates. It also specifies the functional dependence of the covariance between each pair of observations on covariates. Thus, Σ_i depends on i through its dimension $n_i \times n_i$, and through its postulated, if any, dependence on the covariate matrix X_i. Often the values of n_i are unknown to the investigator prior to the collection of data. In such settings, cluster size is also a random variable. It can be shown that

the asymptotic results given for fixed n_i also hold provided we assume each $n_i <$ some fixed value n^*. Although it was not explicitly stated in section 2, when n_i is random, the LMCD is a model for the conditional mean and covariance of Y_i given the covariates X_i and the cluster size n_i. Notice, however, that as formulated, the LMCD does not allow for the mean and covariance of Y_{ij} to depend on cluster size other than through dimensionality, nor for covariate effects to change with cluster size. It is therefore inappropriate for analyzing studies where such relationships are of interest. An illustration of this scenario would be a study of birth weight in animals in which the primary sampling units are litters and the effect of litter size on birth weight is of scientific interest. The LMCD can be easily extended to allow for the dependence of the mean and covariance of Y_i on n_i beyond dimensionality but such extension will not be elaborated in this monograph.

Because the WLS estimate is only consistent when $E(Y_i) = X_i\beta$, when the lack of balance is due to missingness, an MCAR mechanism is required for the validity of the WLS estimate. Because $\widehat{\beta}_{\mathrm{ML}}$ is a WLS estimate, it too is automatically consistent for β even if the error distributions are not normal, provided we have $E(Y_i) = X_i\beta$. The same is true for ML or REML estimates of θ, i.e., normality of the error distributions is not required for consistency. With missingness due to an MAR mechanism, then it generally no longer holds that $E(Y_i) = X_i\beta$, and WLS estimates are no longer valid. ML estimates for both β and θ are consistent in the MAR setting, although now the distributional assumption of the error terms is now crucial, as is the assumption that $\mathrm{var}(Y_i) = \Sigma_i$ is correctly specified.