4

Estimation and Statistical Inference

4.1 INTRODUCTION

So far our discussion of models for longitudinal data has been very general, with no mention of methods for estimating the regression coefficients or the covariance among the repeated measures. In Chapters 5 through 8 we will consider models for longitudinal data where the response variable is continuous and assumed to have a conditional distribution that is approximately multivariate normal. In these chapters the main focus is on various aspects of modeling longitudinal data, with particular emphasis on models for the mean and covariance. All of the models presented in Chapters 5 through 8 can be expressed in terms of a general linear regression model for the mean response vector

$$E(Y_i|X_i) = X_i\beta, \tag{4.1}$$

where the response vector, Y_i, is assumed to have a conditional distribution that is multivariate normal with covariance matrix

$$\text{Cov}(Y_i|X_i) = \Sigma_i = \Sigma_i(\theta), \tag{4.2}$$

where θ is a $q \times 1$ vector of covariance parameters. For example, with balanced longitudinal data ($n_i = n$), where an "unstructured" covariance matrix has been assumed, the elements of θ are simply the n variances and $\frac{n(n-1)}{2}$ pairwise covariances stacked in a single $q \times 1$ vector (where $q = \frac{n(n+1)}{2}$). On the other hand, if the covariance is assumed to have a "compound symmetry" pattern, then $q = 2$ and the two elements

Applied Longitudinal Analysis, Second Edition. By Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. Copyright © 2011 John Wiley & Sons, Inc.

of θ represent the common value of the variances and common value of the pairwise covariances. In this section we consider a framework for estimation of the unknown parameters, β and θ (or equivalently, Σ_i).

4.2 ESTIMATION: MAXIMUM LIKELIHOOD

Given that full distributional assumptions have been made about the vector of responses, Y_i, since the multivariate normal distribution is entirely specified by the mean vector and covariance matrix, a very general approach to estimation is the method of *maximum likelihood* (ML). The fundamental idea behind ML estimation is really quite simple and is conveyed by its name: use as estimates of β and θ the values that are most probable (or most "likely") for the data that have actually been observed. The maximum likelihood estimates of β and θ are those values of β and θ that maximize the joint probability of the response variables evaluated at their observed values. The probability of the response variables evaluated at the fixed set of observed values, and regarded as functions of β and $\Sigma_i(\theta)$, is known as the *likelihood function.* Thus estimation of β and θ proceeds by maximizing the likelihood function. In a certain sense the method of maximum likelihood chooses values of β and θ that best explain the observed data. The values of β and θ that maximize the likelihood function are called the *maximum likelihood estimates* of β and $\Sigma_i(\theta)$, and are usually denoted $\widehat{\beta}$ and $\widehat{\Sigma}_i$ (or $\Sigma_i(\widehat{\theta})$).

Before we present any more details concerning maximum likelihood estimation of β and θ, it will be informative to consider this method of estimation in the simpler case where all observations can be assumed to be independent, that is, in the standard linear regression model with independent (and hence uncorrelated) errors that are assumed to have a univariate normal distribution.

Independent Observations

Suppose that the data arise from a series of cross-sectional studies that are repeated at n different occasions. At each occasion, data are obtained on a sample of N individuals. Here it is reasonable to assume that the observations are independent of one another, since each individual is measured at only one occasion. Also, for ease of exposition, we assume that the variance is constant, say σ^2. The mean response is related to the covariates via the following linear regression model:

$$E(Y_{ij}|X_{ij}) = X_{ij}'\beta.$$

To obtain maximum likelihood estimates of β, we must find the values of the regression parameters that maximize the joint normal probability density function of all the observations, evaluated at the observed values of the response, and regarded as a function of β (and σ^2). Recall from Section 3.2 that the univariate normal (or Gaussian) probability density function for Y_{ij} given X_{ij} can be expressed as

$$f\left(y_{ij}\right) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(y_{ij} - \mu_{ij}\right)^2/\sigma^2\right\},$$

where $-\infty < y_{ij} < \infty$. When all the responses are independent of one another, the likelihood function is simply the product of the individual univariate normal probability density functions for Y_{ij} given X_{ij},

$$\prod_{i=1}^{N}\prod_{j=1}^{n} f\left(y_{ij}\right).$$

It is more common to work with the log-likelihood function, which will involve sums, rather than products, of the individual univariate normal probability density functions for Y_{ij}. Note that maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood; the latter is denoted by l. Hence the goal is to maximize

$$l = \log\left\{\prod_{i=1}^{N}\prod_{j=1}^{n} f\left(y_{ij}\right)\right\} = -\frac{K}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}\left(y_{ij} - X_{ij}'\beta\right)^2/\sigma^2,$$

evaluated at the observed numerical values of the data, with respect to the regression parameters, β. Here $K = n \times N$, the total number of observations. Note that β does not appear in the first term in the log-likelihood; as a result this term can be ignored when maximizing the log-likelihood with respect to β. Furthermore, since the second term has a negative sign, maximizing the log-likelihood with respect to β is equivalent to minimizing the following function:

$$\sum_{i=1}^{N}\sum_{j=1}^{n}\left(y_{ij} - X_{ij}'\beta\right)^2.$$

Maximizing or minimizing a function is a common mathematical problem that can be solved using calculus. Specifically, the maximum likelihood estimate of β can be obtained by equating the derivative of the log-likelihood, often called the score function, to zero and finding the solution to the resulting equation. However, in the example considered here, there is no real need to resort to calculus. Obtaining the maximum likelihood estimate of β is equivalent to finding the ordinary least squares (OLS) estimate of β, that is, the value of β that minimizes the sum of the squares of the residuals. Using vector notation, the least squares solution can be written as

$$\widehat{\beta} = \left\{\sum_{i=1}^{N}\sum_{j=1}^{n}\left(X_{ij}X_{ij}'\right)\right\}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{n}\left(X_{ij}y_{ij}\right).$$

This least squares estimate is the value produced by any standard statistical software for linear regression (e.g., PROC GLM or PROC REG in SAS, the `lm` function in R and S-Plus, and the `regress` command in Stata). In the next section we consider how these ideas can be extended to the setting of correlated data. Also the alert reader might have noticed that we have thus far only focused on estimation of β, ignoring estimation of σ^2; in the next section we also consider estimation of the covariance matrix.

Correlated Observations

When there are n_i repeated measures on the same individual, it cannot be assumed that these repeated measures are independent. As a result we need to consider the joint probability density function for the vector of repeated measures. Note, however, that the vectors of repeated measures are assumed to be independent of one another. Thus the log-likelihood function, l, can be expressed as a sum of the individual multivariate normal probability density functions for Y_i given X_i.

To find the maximum likelihood estimate of β in the repeated measures setting, we first assume that Σ_i (or θ) is *known* (and therefore does not need to be estimated); later we will relax this very unrealistic assumption. To obtain the maximum likelihood estimate of β, we must find the value of β that maximizes the log-likelihood function. Given that $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ is assumed to have a conditional distribution that is multivariate normal, we must maximize the following log-likelihood function:

$$l = -\frac{K}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \log|\Sigma_i| - \frac{1}{2} \left\{ \sum_{i=1}^{N} (y_i - X_i\beta)' \Sigma_i^{-1} (y_i - X_i\beta) \right\},$$

where $K = \left(\sum_{i=1}^{N} n_i \right)$ is the total number of observations. Note that β does not appear in the first two terms in the log-likelihood; as a result these two terms can be ignored when maximizing the log-likelihood with respect to β. Furthermore, since the third term has a negative sign, maximizing the log-likelihood with respect to β is equivalent to minimizing

$$\sum_{i=1}^{N} (y_i - X_i\beta)' \Sigma_i^{-1} (y_i - X_i\beta). \tag{4.3}$$

The estimator of β that minimizes this expression is known as the *generalized least squares* (GLS) estimator of β and can be expressed as

$$\widehat{\beta} = \left\{ \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}X_i) \right\}^{-1} \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}y_i). \tag{4.4}$$

Recall that so far we have made the somewhat unrealistic assumption that Σ_i, or θ, is *known*. Before considering how to proceed when we must relax this assumption, it is worth discussing some of the properties of the GLS estimator of β when Σ_i is known. The first very notable property is that for any choice of Σ_i, the GLS estimate of β is unbiased; that is,

$$E(\widehat{\beta}) = \beta.$$

In addition, in large samples (or asymptotically), the sampling distribution of $\widehat{\beta}$ can be shown to have a multivariate normal distribution with mean, β, and covariance,

$$\text{Cov}(\widehat{\beta}) = \left\{ \sum_{i=1}^{N} (X_i'\Sigma_i^{-1}X_i) \right\}^{-1}. \tag{4.5}$$

This is true exactly when Y_i has a conditional distribution that is multivariate normal, and true in large samples even when the conditional distribution of Y_i is not multivariate normal. (By "large samples" we mean that the sample size, N, grows larger while the number of repeated measures and model parameters remains fixed.) Thus an important property of the GLS estimator of β, derived under the assumption of a multivariate normal distribution for Y_i given X_i, is that it provides a valid estimate of β even when the multivariate normal distribution assumption does not hold. Also note that if Σ_i is assumed to be a diagonal matrix, with constant variance σ^2 along the diagonal (i.e., the correlations are zero and the variances are constant), the GLS estimator reduces to the ordinary least squares (OLS) estimator considered earlier. Finally, although the GLS estimator of β is unbiased for any choice of Σ_i, it can be shown that the most efficient GLS estimator of β (i.e., the estimator having smallest variance or greatest precision) is the one that uses the true value of Σ_i.

Before the reader becomes exasperated, we must address the nagging concern that we usually do not know Σ_i (or θ). Instead, we typically must estimate $\Sigma_i(\theta)$ from the data at hand. Maximum likelihood estimation of θ proceeds in the same way as with estimation of β. That is, the maximum likelihood estimate of θ is obtained by maximizing the log-likelihood with respect to θ. As mentioned earlier, the problem of maximizing a function is a common mathematical problem that can be solved using calculus. Specifically, the maximum likelihood estimate of θ can be obtained by equating the derivative of the log-likelihood with respect to θ, also known as the score function, to zero and finding the solution to the resulting equation. However, in general, this equation is non-linear, and it is not possible to write down simple, closed-form expressions for the ML estimator of θ. Instead, the ML estimate must be found by solving these equation using an iterative technique. Fortunately, computer algorithms have been developed to find the solution. Once the ML estimate of θ has been obtained, we then simply substitute the estimate of $\Sigma_i(\theta)$, say $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$, into the generalized least squares estimator of β given by (4.4) to obtain the maximum likelihood (ML) estimate of β:

$$\widehat{\beta} = \left\{ \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}_i^{-1} X_i \right) \right\}^{-1} \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}_i^{-1} y_i \right). \tag{4.6}$$

Interestingly, in large samples (or asymptotically), the resulting estimator of β that substitutes the ML estimate of Σ_i has all of the same properties as when Σ_i is actually known (the case we first considered). That is, in large samples:

1. $\widehat{\beta}$ is a consistent estimator of β; this property can be loosely interpreted to mean that there is very high probability that $\widehat{\beta}$ is close to the population regression parameters β for increasing sample size N. If the distribution of the errors, e_i, is assumed to be normal, or even under the weaker assumption that the distribution of e_i is symmetric, then $\widehat{\beta}$ is also an unbiased estimator of β,

$$E(\widehat{\beta}) = \beta.$$

2. The sampling distribution of $\widehat{\beta}$, when Σ_i is estimated from the data, is approximately multivariate normal with mean, β, and covariance

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

Furthermore these properties of $\widehat{\beta}$ hold in large samples even when the assumption that Y_i has a multivariate normal distribution is not valid, provided the data are complete. Thus an important property of the ML estimator of β, derived under the assumption of a multivariate normal distribution for Y_i given X_i, is that it provides a valid estimate of β even when the multivariate normal distribution assumption does not hold. Moreover this appealing property of the ML estimator of β, and of any GLS estimator of β (recall, the ML estimator of β is also the GLS estimator with the ML estimate of $\Sigma_i(\theta)$ substituted), extends to the incomplete data setting when certain assumptions about missingness hold.

Thus, in terms of properties of the sampling distribution of $\widehat{\beta}$, there is no penalty for actually having to estimate Σ_i from the longitudinal data at hand. However, as comforting as this result may appear to be, it must be kept in mind that this is a large sample (i.e., as N approaches infinity) property of $\widehat{\beta}$. With sample sizes of the magnitude often encountered in many fields of application, the properties of the sampling distribution of $\widehat{\beta}$ can be expected to be adversely influenced by the estimation of a very large number of covariance parameters. This is an important issue that we will return to in Chapter 7.

4.3 MISSING DATA ISSUES

Although most longitudinal studies are designed to collect data on every individual in the sample at each time of follow-up, many studies have some missing observations. In longitudinal studies in the health sciences, missing data are the rule, not the exception. Missing data have three important implications for longitudinal analysis. First, when longitudinal data are missing, the data set is necessarily unbalanced over time since not all individuals have the same number of repeated measurements at a common set of occasions. As a result methods of analysis need to be able to handle the unbalanced data without having to discard data on individuals with any missing data. This feature of missingness will not be of any concern for the methods described in later chapters of the book. Second, when there are missing data, there will be a loss of information and a reduction in the precision with which changes in the mean response over time can be estimated. This reduction in precision is directly related to the amount of missing data and will also be influenced to a certain extent by how the analysis handles the missing data. For example, using only the complete cases (i.e., those individuals with no missing data) will usually be the least efficient method. Finally, when there are missing data, the validity of any method of analysis will require that certain assumptions about the reasons for any missingness, often referred to as the *missing data mechanism*, are tenable. Consequently, when data are missing we must carefully consider the reasons for missingness.

In this section we review two general types of missing data mechanisms. The two mechanisms differ in terms of assumptions concerning whether missingness is related to responses that have been observed. The distinctions between different types of missing data mechanisms and alternative methods for handling missingness in longitudinal studies will be discussed in greater detail in Chapters 17 and 18.

The missing data mechanism can be thought of as a model that describes the probability that a response is observed or missing at any occasion. We make an important distinction between missing data mechanisms that are referred to as *missing completely at random* (MCAR) and *missing at random* (MAR). The distinction between these two mechanisms determines the appropriateness of maximum likelihood estimation under the assumption of a multivariate normal distribution for the responses and GLS without requiring assumptions about the shape of the distribution.

Data are said to be missing completely at random (MCAR) when the probability that responses are missing is unrelated to either the specific values that, in principle, should have been obtained (the *missing* responses) or the set of observed responses. That is, longitudinal data are MCAR when missingness in Y_i is simply the result of a chance mechanism that does not depend on either observed or unobserved components of Y_i. The essential feature of MCAR is that the observed data can by thought of as a random sample of the complete data. As a result the moments (e.g., the means, variances, and covariances), and indeed, the distribution of the observed data do not differ from the corresponding moments or distribution of the complete data.

An MCAR mechanism has important consequences for the analysis of longitudinal data. In particular, any method of analysis that yields valid inferences in the absence of missing data will also yield valid inferences when missing data are MCAR and the analysis is based on all available data, or even when it is restricted to the "completers" (i.e., those with no missing data). Given that valid estimates of the means, variances, and covariances can be obtained, GLS provides valid estimates of β without requiring any distributional assumptions for Y_i. The GLS estimator of β is valid provided the model for the mean response has been correctly specified; it does not require any assumptions about the joint distribution of the longitudinal responses. The maximum likelihood (ML) estimator of β, under the assumption that the responses have a multivariate normal distribution, is also the GLS estimator (with the ML estimate of $\Sigma_i(\theta)$, e.g., $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$, substituted). Thus in this setting the ML and GLS estimators have exactly the same properties regardless of the true distribution of Y_i.

In contrast to MCAR, data are said to be missing at random (MAR) when the probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that, in principle, should have been obtained. Put another way, if subjects are stratified on the basis of similar values for the responses that have been observed, missingness is simply the result of a chance mechanism that does not depend on the values of the unobserved responses. However, because the missingness mechanism now depends on observed responses, the distribution of Y_i in each of the distinct strata defined by the patterns of missingness is not the same as the distribution of Y_i in the target population. This has important consequences for analysis. One is that an analysis restricted to the "completers" is not valid. Put another way, the "completers" are a biased sample from the target population. Furthermore the distribution of the observed components of Y_i, in each of the distinct strata defined by the patterns of missingness, does not coincide with the distribution of the same components of Y_i in the target population. Therefore the sample means, variances, and covariances based on either the "completers," or the available data are biased estimates of the corresponding parameters in the target population. As a result GLS no longer provides valid estimates of β without making correct assumptions about the joint distribution of the longitudinal responses. On the other hand, ML estimation of β is valid when data are MAR provided that the multivariate normal distribution has been correctly specified. This requires correct specification of not only the model for the mean response but also the model for the covariance among the responses. In a sense, ML estimation allows the missing values to be validly "predicted" or "imputed" using the observed data and a correct model for the joint distribution of the responses.

To summarize, we have distinguished between two types of missing data mechanisms that are referred to as *missing completely at random* (MCAR) and *missing at random* (MAR). The MAR assumption is far less restrictive than MCAR. The distinction between these two mechanisms determines the appropriateness of maximum likelihood estimation under the assumption of a multivariate normal distribution for the responses and GLS without requiring assumptions about the shape of the distribution. The general properties of GLS described in the previous section require that either the data are complete or that any missing data are MCAR. If data are MAR, GLS based only on the means, variances, and covariances of the available data can yield biased estimates of β. In contrast, ML estimation yields valid estimates of β when data are MCAR or MAR, but for the latter mechanism, at the cost of requiring that the joint distribution of the responses is correctly specified. A more detailed discussion of missing data mechanisms, with concrete examples, and the implications of different types of missing data mechanisms for analysis, is presented in Chapter 17.

4.4 STATISTICAL INFERENCE

Next we consider how to make inferences about β. In particular, we consider the construction of confidence intervals and tests of hypotheses. To construct confidence intervals and tests of hypotheses about β, we can make direct use of the ML estimate $\widehat{\beta}$, and its estimated covariance matrix

$$\widehat{\text{Cov}}(\widehat{\beta}) = \left\{ \sum_{i=1}^{N} \left(X_i' \widehat{\Sigma}_i^{-1} X_i \right) \right\}^{-1},$$

where Σ_i in (4.5) is replaced by $\widehat{\Sigma}_i$, the ML estimate of Σ_i. For example, for any single component of β, say β_k, a natural method for constructing 95% confidence limits is by taking $\widehat{\beta}_k$ plus or minus 1.96 times the standard error of $\widehat{\beta}_k$. Note that different confidence limits (e.g., 90%) can be obtained by choosing appropriate multiples of the standard error, based on the standard normal distribution. The standard error of $\widehat{\beta}_k$ is simply the square-root of the diagonal element of $\widehat{\text{Cov}}(\widehat{\beta})$ corresponding to $\widehat{\beta}_k$,

$$\sqrt{\widehat{\text{Var}}(\widehat{\beta}_k)}.$$

Similarly a test of the null hypothesis, H$_0$: $\beta_k = 0$ versus H$_A$: $\beta_k \neq 0$, can be based on the following Wald statistic:

$$Z = \frac{\widehat{\beta}_k}{\sqrt{\widehat{\text{Var}}(\widehat{\beta}_k)}},$$

where $\widehat{\text{Var}}(\widehat{\beta}_k)$ denotes the diagonal element of $\widehat{\text{Cov}}(\widehat{\beta})$ corresponding to $\widehat{\beta}_k$. This test statistic can be compared with a standard normal distribution.

More generally, it may be of interest to construct confidence intervals and tests of hypotheses about certain linear combinations of the components of β. Let L denote a vector or matrix of *known* weights, and suppose that it is of interest to test H$_0$: $L\beta = 0$. The linear combination of the components of β, $L\beta$, represents a contrast of scientific interest. For example, suppose that $\beta = (\beta_1, \beta_2, \beta_3)'$ and let $L = (0, 0, 1)$, then H$_0$: $L\beta = 0$ is equivalent to H$_0$: $\beta_3 = 0$. Alternatively, if $L = (0, 1, -1)$, then H$_0$: $L\beta = 0$ is equivalent to H$_0$: $\beta_2 - \beta_3 = 0$ or H$_0$: $\beta_2 = \beta_3$. A natural estimate of $L\beta$ is given by $L\widehat{\beta}$. Moreover it can be shown that the sampling distribution of $L\widehat{\beta}$ is multivariate normal with mean, $L\beta$, and with covariance matrix, $L\text{Cov}(\widehat{\beta})L'$.

Note that in the two examples considered earlier, L is a single, 1 × 3 row vector, $L = (0, 0, 1)$ or $L = (0, 1, -1)$. If L is a single row vector, then $L\,\text{Cov}(\widehat{\beta})L'$ is a single value (or scalar) and its square-root provides an estimate of the standard error for $L\widehat{\beta}$. Thus an approximate 95% confidence interval for $L\beta$ is given by

$$L\widehat{\beta} \pm 1.96\sqrt{L\,\widehat{\text{Cov}}(\widehat{\beta})L'}.$$

Similarly, in order to test H$_0$: $L\beta = 0$ versus H$_A$: $L\beta \neq 0$, we can use the Wald statistic,

$$Z = \frac{L\widehat{\beta}}{\sqrt{L\,\widehat{\text{Cov}}(\widehat{\beta})L'}},$$

and compare this test statistic to a standard normal distribution. Recall that if Z is a standard normal random variable, then Z^2 has a χ^2 distribution with 1 degree of freedom (df), denoted χ^2_1. Thus an identical test of H$_0$: $L\beta = 0$ versus H$_A$: $L\beta \neq 0$, uses the statistic

$$W^2 = (L\widehat{\beta})\{L\,\widehat{\text{Cov}}(\widehat{\beta})L'\}^{-1}(L\widehat{\beta}),$$

and compares W^2 to a χ^2 distribution with 1 degree of freedom. This latter observation helps to motivate how the Wald test readily generalizes when L has more than one row, thereby allowing simultaneous testing of a single multivariate hypothesis. For example, suppose that $\beta = (\beta_1, \beta_2, \beta_3)'$ and it is of interest to test the equality of the three regression parameters. The null hypothesis can be expressed as H$_0$: $\beta_1 = \beta_2 = \beta_3$. Letting

$$L = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix},$$

this null hypothesis can also be expressed as H$_0$: $L\beta = 0$, since if

$$\begin{aligned} L\beta &= \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \\ &= \begin{pmatrix} \beta_1 - \beta_2 \\ \beta_1 - \beta_3 \end{pmatrix} = 0, \end{aligned}$$

then

$$\begin{pmatrix} \beta_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix},$$

or, equivalently, $\beta_1 = \beta_2 = \beta_3$. In general, suppose that L has r rows (e.g., representing r contrasts of scientific interest), then a simultaneous test of H$_0$: $L\beta = 0$ versus H$_A$: $L\beta \neq 0$ is given by

$$W^2 = (L\widehat{\beta})'\{L\widehat{\text{Cov}}(\widehat{\beta})L'\}^{-1}(L\widehat{\beta}),$$

which has a χ^2 distribution with r df. The latter test is often referred to as a multivariate Wald test.

One alternative to the Wald test is the *likelihood ratio test*. The likelihood ratio test of H$_0$: $L\beta = 0$ versus H$_A$: $L\beta \neq 0$ is obtained by comparing the maximized log-likelihoods for two models, one model that incorporates the constraint that $L\boldsymbol{\beta} = 0$ (e.g., $\beta_3 = 0$ or $\beta_2 = \beta_3$), the other model unconstrained (i.e., without the constraint, $L\boldsymbol{\beta} = 0$). The latter is referred to as the "full" model and the former is referred to as the "reduced" model. Note that these two models are *nested*, in the sense that the "reduced" model is a special case of the "full" model. That is, when the reduced model is *nested* within the full model, it is a particular version of the full model, so that when the reduced model holds, the full model must necessarily hold.

The likelihood ratio test for two nested models can be constructed by comparing their respective maximized log-likelihoods, say $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$, for the full and reduced models, respectively. The former is at least as large as the latter. The larger the difference between $\widehat{l}_{\text{full}}$ and $\widehat{l}_{\text{red}}$, the stronger the evidence is that the reduced model is inadequate. A formal statistical test is obtained by taking twice the difference in the respective maximized log-likelihoods,

$$G^2 = 2(\widehat{l}_{\text{full}} - \widehat{l}_{\text{red}}),$$

and comparing the statistic to a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the full and reduced models. This test is called the *likelihood ratio test*. We caution that the likelihood ratio test can only be used when the number of observations for the "full" and "reduced" models is the same. For example, when there are some missing data on a covariate that belongs to the "full" model only, the "full" and "reduced" models are no longer nested. Instead, to ensure the models are nested, the likelihood ratio test must be based on a comparison of the "full" and "reduced" models for the same subset of observations with no missing values on that covariate.

This use of the likelihood can also provide confidence limits for β or $L\beta$. Rather than calculating confidence limits for β (or $L\beta$) as the maximum likelihood estimate, $\widehat{\beta}$, plus or minus an appropriate multiple of the standard errors, likelihood-based confidence intervals can be constructed. The basic idea behind likelihood-based confidence intervals is to consider all values of β (or $L\beta$) that are consistent with the data at hand. More formally, for a single component of β, say β_k, we can define a *profile log-likelihood*, $l_p(\beta_k)$, obtained by maximizing the log-likelihood over the remaining parameters while holding β_k at some fixed value. A likelihood-based confidence interval for β_k is obtained by considering values of β_k that are reasonably consistent with the data. Specifically, an approximate 95% likelihood-based confidence interval is given by the set of all values of β_k satisfying

$$2 \times \left\{ l_p(\widehat{\beta}_k) - l_p(\beta_k) \right\} \leq 3.84,$$

where the critical value on the right-hand side of the equation is obtained from a chi-squared distribution with 1 degree of freedom. More generally, confidence intervals for $L\beta$ can be obtained by inverting the corresponding test of H_0: $L\beta = 0$ in a similar way.

Although the construction of likelihood ratio tests and likelihood-based confidence intervals is more involved (e.g., requiring an additional fit of the model under the null hypothesis) than the corresponding Wald-based tests and confidence intervals, the likelihood-based tests and confidence intervals often have superior properties. This is especially the case when the response variable is discrete. For example, in logistic regression with binary data, likelihood ratio tests have better properties than the corresponding Wald tests. Thus, when in doubt, we recommend the use of likelihood-based tests and confidence intervals. However, for ease of presentation, many of the results presented in later chapters rely on Wald-based tests and confidences intervals; likelihood-based tests and confidence intervals are presented only in cases where the discrepancies might change the substantive conclusions of the analysis.

Finally, we note that likelihood ratio tests can also be used for hypotheses about the covariance parameters. However, there are some potential problems with the standard use of the likelihood ratio test for comparing nested models for the covariance; we will return to this topic in Chapter 7. In general, we do not recommend testing hypotheses about the covariance parameters using Wald tests (i.e., based on the ratio of the parameter estimate to its standard error). In particular, the sampling distribution of the Wald test statistic for a variance parameter does not have an approximate normal distribution when the sample size is relatively small and the population variance is close to zero. Because the variance has a lower bound of zero, very large samples are required to justify the normal approximation for the sampling distribution of the Wald test statistic when the variance is close to zero.

Comment on Denominator Degrees of Freedom

So far in our discussion of confidence intervals and tests of hypotheses about β we have relied on the large sample properties of the sampling distribution of the ML estimate of β. That is, we have used the standard normal and chi-squared distributions instead of t and F distributions. It can be argued that the use of the standard normal and chi-squared distributions is more "liberal" (or "anti-conservative") than the corresponding t and F distributions because there is an implicit assumption of infinite denominator degrees of freedom. By "liberal," we mean that nominal p-values may be too small and confidence intervals may be too narrow.

With large denominator degrees of freedom (e.g., due to a large sample size) estimation of θ or Σ_i does not introduce any additional uncertainty. However, with small sample sizes there is some uncertainty attached to the estimation of θ that needs to be acknowledged in our inferences about β. Ordinarily this additional source of uncertainty is recognized by use of the t and F distributions instead of the standard normal and chi-squared distributions.

A practical difficulty with the use of the t and F distributions in this setting is that the denominator degrees of freedom associated with tests and confidence intervals for components of β is not easy to determine except in certain special cases where the data are balanced and the model for the mean has a relatively simple form. To circumvent this difficulty, various approximations for the denominator degrees of freedom have been proposed. One well-known method is the Satterthwaite approximation, a somewhat tedious and computationally demanding procedure. If Satterthwaite's (1946) method is used to obtain approximate denominator degrees of freedom, say $\widehat{v}$, then an approximate 95% confidence interval for $L\beta$ is given by

$$L\widehat{\beta} \pm t_{\widehat{v},0.025}\sqrt{L\,\widehat{\text{Cov}}(\widehat{\beta})L'},$$

where $t_{\widehat{v},0.025}$ is the upper 2.5% cutoff from a t distribution with $\widehat{v}$ degrees of freedom (i.e., for the t distribution with $\widehat{v}$ degrees of freedom, 95% of the area lies between $-t_{\widehat{v},0.025}$ and $t_{\widehat{v},0.025}$). Similarly, to test H$_0$: $L\beta = 0$ versus H$_A$: $L\beta \neq 0$, we can use the Wald statistic

$$\frac{L\widehat{\beta}}{\sqrt{L\,\widehat{\text{Cov}}(\widehat{\beta})L'}}$$

and compare this test statistic to a t distribution with $\widehat{v}$ degrees of freedom. The Satterthwaite approximation can also be applied to multivariate Wald statistics, with the chi-squared distribution replaced by the F distribution (when the multivariate Wald statistic has been divided by the number of rows of the matrix L or the numerator degrees of freedom).

Recently Kenward and Roger (1997) proposed an alternative approximation that adjusts the test statistics and provides approximate denominator degrees of freedom. Although the Satterthwaite approximation and the approximation proposed by Kenward and Roger (1997) are implemented as options in some statistical software packages (e.g., PROC MIXED in SAS), it must be emphasized that the small sample properties of these approximations in regression models for longitudinal data have not been extensively studied.

In summary, the use of the standard normal and chi-squared distributions is valid when Σ_i (or θ) is known, or when Σ_i has been estimated with a large number of degrees of freedom. Recall that there is not much practical difference between the use of the standard normal and t distributions once the degrees of freedom of the latter exceed 100. With small sample sizes, there is some uncertainty in the estimation of θ that should be accounted for and the use of the t and F distributions, with degrees of freedom approximated by the methods of Satterthwaite (1944) or Kenward and Roger (1997), is preferred. Fortunately, in many applications in the health sciences the numbers of subjects is reasonably large relative to the number of measurement occasions. As a result the unknown denominator degrees of freedom, especially for components of β that represent time trends and their interactions with covariates (e.g., group × time interactions), will be sufficiently large that the standard normal and chi-squared distributions are reasonable approximations to the corresponding t and F distributions. For the remainder of the book, we construct confidence intervals and tests of hypotheses about β using the standard normal and chi-squared distributions; we use approximations for the denominator degrees of freedom only in cases where it might change the substantive conclusions of the analysis.

4.5 RESTRICTED MAXIMUM LIKELIHOOD (REML) ESTIMATION

We conclude this chapter with a discussion of a variant on ML estimation, known as *restricted maximum likelihood* (REML) estimation. Recall that the ML estimates of β and θ (or Σ_i) were obtained by maximizing the following log-likelihood function:

$$l = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\Sigma_i| - \frac{1}{2}\left\{\sum_{i=1}^{N}(y_i - X_i\beta)'\,\Sigma_i^{-1}\,(y_i - X_i\beta)\right\}.$$

Although the ML estimates of β and $\Sigma_i(\theta)$ have desirable large sample (or asymptotic) properties, the ML estimate of Σ_i has a well-known bias in finite samples. For example, the diagonal elements of Σ_i are underestimated.

To illustrate the problem, consider the case where data arise from a series of cross-sectional studies that are repeated at n different occasions. Here we can assume that the observations are independent of one another, and for ease of exposition we also assume that the variance is constant, say σ^2. As noted earlier, the ML estimates of β

and σ^2 are obtaining by maximizing

$$-\frac{K}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}\left(y_{ij}-X_{ij}'\beta\right)^2/\sigma^2.$$

The ML estimator of β is

$$\widehat{\beta}=\left\{\sum_{i=1}^{N}\sum_{j=1}^{n}\left(X_{ij}X_{ij}'\right)\right\}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{n}\left(X_{ij}y_{ij}\right),$$

while the ML estimator of σ^2 is

$$\widehat{\sigma}^2=\sum_{i=1}^{N}\sum_{j=1}^{n}\left(Y_{ij}-X_{ij}'\widehat{\beta}\right)^2/K,$$

where $K=n\times N$. Furthermore, it can be shown that

$$E(\widehat{\sigma}^2)=\left(\frac{K-p}{K}\right)\sigma^2,$$

where p is the dimension of β. As a result, the ML estimate of σ^2 is biased in small samples and underestimates σ^2. An unbiased estimator is obtained by using $K-p$ (or the residual degrees of freedom) as the denominator instead of K,

$$\widehat{\sigma}^2=\sum_{i=1}^{N}\sum_{j=1}^{n}\left(Y_{ij}-X_{ij}'\widehat{\beta}\right)^2/(K-p).$$

This estimator for σ^2 is also known as the REML estimator. Note that the bias of the ML estimate of σ^2 is a decreasing function of the total number of observations, K.

In effect, the bias arises because the ML estimate has not taken into account the fact that β is also estimated from the data. In the estimator of σ^2 we have replaced β by $\widehat{\beta}$ but have failed to acknowledge in some sense that β was estimated from the data. If there are problems of bias with the ML estimate of σ^2 with independent observations, then it should not come as a great surprise that similar problems arise in the estimation of Σ_i (or θ) with correlated data.

The theory of restricted (or residual) maximum likelihood (REML) estimation was developed to address this problem. The main idea behind REML estimation is to separate that part of the data used for estimation of Σ_i from that used for estimation of β. Estimation of Σ_i is then based only on the relevant part of the data. Thus the fundamental idea in REML estimation of Σ_i is to eliminate β from the likelihood so that it is defined only in terms of Σ_i. This can be achieved in a number of ways. One possible way to obtain the restricted likelihood is to transform the data to a set of linear combinations of observations that have a distribution that does not depend on β. For example, the residuals after estimating β by ordinary least squares (OLS) can be used as the data for estimating Σ_i (or θ). The likelihood for these residuals depends only on θ, and not on β. Thus, rather than maximizing the log-likelihood

$$-\frac{1}{2}\sum_{i=1}^{N} \log|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{N}\left(y_i - X_i\widehat{\beta}\right)' \Sigma_i^{-1}\left(y_i - X_i\widehat{\beta}\right), \tag{4.7}$$

REML maximizes the following slightly modified log-likelihood (formed from the residuals)

$$\begin{aligned} -\frac{1}{2}\sum_{i=1}^{N} \log|\Sigma_i| &- \frac{1}{2}\sum_{i=1}^{N}\left(y_i - X_i\widehat{\beta}\right)' \Sigma_i^{-1}\left(y_i - X_i\widehat{\beta}\right) \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{N} X_i'\Sigma_i^{-1}X_i\right|. \end{aligned} \tag{4.8}$$

When the residual likelihood given by (4.8) is maximized, we obtain an estimate of θ (or $\Sigma_i(\theta)$) that has made a correction for the fact that β has also been estimated. Of note, the additional term in the REML log-likelihood involves a determinant term,

$$\begin{aligned} -\frac{1}{2}\log\left|\sum_{i=1}^{N} X_i'\Sigma_i^{-1}X_i\right| &= \frac{1}{2}\log\left|\left(\sum_{i=1}^{N} X_i'\Sigma_i^{-1}X_i\right)^{-1}\right| \\ &= \log\left|\text{Cov}(\widehat{\beta})\right|^{\frac{1}{2}}, \end{aligned}$$

that can be expressed as the covariance of $\widehat{\beta}$. As a result the REML likelihood multiplies that usual ML likelihood by a factor that is the square-root of the *generalized variance* of $\widehat{\beta}$, a single number summary of the variation in the estimate of β. This makes a correction or adjustments that is analogous to the correction to the denominator in $\hat{\sigma}^2$.

We recommend the use of the REML estimator for Σ_i. In general, the REML estimator will be less seriously biased than the ML estimator for Σ_i. It should be noted that the difference between ML and REML estimation becomes less important when the sample size, N, is substantially larger than p, the dimension of β. Finally, when REML estimation is used to estimate Σ, β is estimated by the usual generalized least squares (GLS) estimator

$$\widehat{\beta} = \left\{\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}_i^{-1}X_i\right)\right\}^{-1}\sum_{i=1}^{N}\left(X_i'\widehat{\Sigma}_i^{-1}Y_i\right),$$

where $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$ is the REML estimate of Σ_i.

On a final note, while the REML log-likelihood can be used to compare nested models for the covariance (e.g., in terms of likelihood ratio tests comparing nested models for the covariance), it should not be used to compare nested regression models for the mean. The extra determinant term in the REML log-likelihood depends on the regression model specification. As a result the REML likelihoods for two nested models for the mean response are based on quite different transformations of the data (to obtain linear combinations of Y_i whose distributions do not depend on β). In short, the REML likelihoods for two nested models for the mean are based on two entirely different sets of transformed responses, making comparisons between the models meaningless. Instead, the standard ML log-likelihood should be used for constructing likelihood ratio tests that compare nested regression models for the mean.

In conclusion, we recommend the use of REML for estimation of Σ_i (with β estimated using the GLS estimator that substitutes the REML estimate, $\widehat{\Sigma}_i$, for Σ_i). The REML log-likelihood should also be used to comparing nested models for the covariance. However, the construction of likelihood ratio tests comparing nested models for the mean should always be based on the ML, not the REML, log-likelihood.

4.6 FURTHER READING

Many textbooks on statistical theory and methods include a discussion of the methods of least squares and maximum likelihood estimation. Weisberg (1985) provides a useful introduction to the method of least squares in the context of regression; Chapter 4 of Cox and Wermuth (1996) presents a concise but remarkably lucid description of least squares, generalized least squares, and maximum likelihood estimation.

Bibliographic Notes

A discussion of the properties of generalized least squares (GLS) estimators can be found in, for example, Amemiya (1985) and Newey and McFadden (1994). Kakwani (1967) and Kackar and Harville (1981) discuss the unbiasedness properties of GLS estimators when the assumption of normally distributed errors is replaced by the weaker assumption that the distribution of the errors is symmetric.

The use of REML, as an alternative to maximum likelihood, for covariance parameter estimation was originally proposed by Patterson and Thompson (1971). Special cases of REML estimation had previously been considered by Anderson and Bancroft (1952), Russell and Bradley (1958), and Thompson (1962) in the context of balanced ANOVA models. Harville (1974) presented a Bayesian interpretation of REML.