INFERENCE FROM THE INCOMPLETE LONGITUDINAL DESIGN UNDER AN ARMA COVARIANCE STRUCTURE

by

James Rochon

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1489

June 1985
INFERENCE FROM THE INCOMPLETE LONGITUDINAL DESIGN UNDER AN ARMA COVARIANCE STRUCTURE

by

James Rochon

A Dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1985

Approved by:

[Signatures]

Advisor

[Signature]

Reader

[Signature]

Reader
ABSTRACT

JAMES ROCHON. Inference from the Incomplete Longitudinal Design under an ARMA Covariance Structure (Under the direction of RONALD W. HELMS).

A stochastic model is presented for the analysis of the longitudinal design, appropriate when some of the response variables are missing. The general linear model is used to relate these dependent variables to other variables which are thought to account for inherent variation. An ARMA time series representation is used to model disturbance terms, resulting in a characteristic structure in the covariance matrix among the repeated measures.

Maximum likelihood estimation procedures are performed. An unbiased estimator for the general linear model parameter vector is derived, and this estimator is observed to enjoy favourable large-sample properties. There is no convenient expression for the ARMA parameter estimator, and non-linear optimization procedures are prescribed to iterate to a maximizing value.

Two broad classes of hypotheses among the indigenous parameters are considered, and procedures to test these hypotheses are derived. The first examines the covariance structure among the repeated measures in an effort to verify the ARMA model. The second considers more substantive hypotheses among the linear model parameters.

Two datasets are studied in detail to illustrate the procedures. The first was generated at random from a known stochastic mechanism, while the second was derived from a paper published in the statistical literature. The underlying stochastic patterns are largely verified by the methodology.
It is concluded that while the assumptions underpinning the ARMA covariance models may be somewhat restrictive for many practical situations, they nevertheless offer a wide variety of covariance structures, which can have a salutary effect, particularly in the presence of missing values.
ACKNOWLEDGEMENTS

I gratefully acknowledge the guidance and support of my dissertation advisor, Dr. Ron Helms. As well, I would like to thank the other members of my dissertation committee: Prof. James Grizzle, Dr. Keith Muller, Prof. Forrest Young, and especially Dr. Dave Delong, for their help and advice over the course of this project.

A vote of thanks also goes to Mr. Ron Haines of Campbell Laser Printing, for his help in defining the printing programme in the first instance, and printing the various drafts of the paper. I knew that I could always count on Ron for fast turn-around and a good clean copy.

I would also like to thank a very special person for her uncompromising support throughout the course of my studies at the Ph.D. level in general, and especially through the dissertation itself. This project was essentially a joint effort, and while the scientific input was entirely mine, it really belongs to both of us.
<table>
<thead>
<tr>
<th>Chapter</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. REVIEW OF THE LITERATURE</td>
<td>5</td>
</tr>
<tr>
<td>Univariy Regression Model</td>
<td>5</td>
</tr>
<tr>
<td>Numerical Methods Techniques</td>
<td>17</td>
</tr>
<tr>
<td>Longitudinal Data Designs</td>
<td>23</td>
</tr>
<tr>
<td>Missing Data</td>
<td>32</td>
</tr>
<tr>
<td>III. MAXIMUM LIKELIHOOD ESTIMATION</td>
<td>39</td>
</tr>
<tr>
<td>The Model</td>
<td>39</td>
</tr>
<tr>
<td>Maximum Likelihood Equations</td>
<td>47</td>
</tr>
<tr>
<td>Properties of ML Estimators</td>
<td>51</td>
</tr>
<tr>
<td>IV. NON-LINEAR OPTIMIZATION PROCEDURES</td>
<td>62</td>
</tr>
<tr>
<td>Introduction</td>
<td>62</td>
</tr>
<tr>
<td>Iteration Sequence</td>
<td>63</td>
</tr>
<tr>
<td>Initial Estimates of the Parameters</td>
<td>69</td>
</tr>
<tr>
<td>Computational Considerations</td>
<td>73</td>
</tr>
</tbody>
</table>
V. HYPOTHESIS TESTING ........................................ 77
   Introduction ............................................. 77
   Hypotheses Concerning the Covariance Structure .......... 82
   Hypotheses Concerning the Linear Model Parameters ....... 90

VI. EXAMPLES ................................................... 97
   Introduction ............................................. 97
   Example 1 ................................................. 100
   Example 2 ................................................. 114
   Discussion ............................................... 124

VII. SUMMARY AND CONCLUSIONS ............................... 128

BIBLIOGRAPHY .................................................. 132

Appendix

A. LISTING OF THE DATASET USED IN EXAMPLE 1 .......... 139

INDEX OF REFERENCES CITED ................................. 142
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.2.1</td>
<td>Frequency of Missing values by Group and Time</td>
<td>102</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Time-Specific and Overall Sample Autocorrelations, with Approximate Standard Errors in Parentheses, Compared Against Theoretical Population Values</td>
<td>104</td>
</tr>
<tr>
<td>6.2.3</td>
<td>ML Estimate of θ Under the General Covariance Model, with the Corresponding Population Values in Parentheses</td>
<td>105</td>
</tr>
<tr>
<td>6.2.4</td>
<td>ML Estimates of θ and σ² for Various ARMA Models, with the Corresponding Covariance Sequence, Compared Against the Theoretical Population Values</td>
<td>107</td>
</tr>
<tr>
<td>6.2.5</td>
<td>Results of Hypothesis Testing for Various ARMA Models</td>
<td>108</td>
</tr>
<tr>
<td>6.2.6</td>
<td>ML Estimate of the Linear Model Parameter Matrix B, with the Asymptotic Standard Errors in Parentheses</td>
<td>109</td>
</tr>
<tr>
<td>6.2.7</td>
<td>Results of Hypothesis Testing Among the Linear Model Parameters</td>
<td>109</td>
</tr>
<tr>
<td>6.2.8</td>
<td>OLS Estimates of the Linear Model Parameters From the Piecemeal Univariate Analyses, with the Corresponding Standard Errors in Parentheses</td>
<td>113</td>
</tr>
<tr>
<td>6.3.1</td>
<td>ML Estimate of θ under the General Covariance Model</td>
<td>116</td>
</tr>
<tr>
<td>6.3.2a</td>
<td>Sample Autocorrelations Specific to Observations Taken at Hours 0.5, 1.0, 1.5, with Approximate Standard Errors in Parentheses</td>
<td>118</td>
</tr>
<tr>
<td>6.3.2b</td>
<td>Sample Autocorrelations Specific to Observations Taken at Hours 2, 3, 4, with Approximate Standard Errors in Parentheses</td>
<td>118</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Results of Hypothesis Testing on the Covariance Structure</td>
<td>120</td>
</tr>
<tr>
<td>6.3.4</td>
<td>ML Estimates of θ and σ² for Various ARMA Models, with the Corresponding Covariance Sequence</td>
<td>121</td>
</tr>
<tr>
<td>6.3.5</td>
<td>ML Estimate of the Linear Model Parameter Matrix B, with the Asymptotic Standard Errors in Parentheses</td>
<td>122</td>
</tr>
<tr>
<td>6.3.6</td>
<td>Results of Hypothesis Testing Among the Linear Model Parameters</td>
<td>123</td>
</tr>
</tbody>
</table>
6.3.7. Hypothesis Testing for Control vs. Obese Subpopulations from Piecemeal Univariate Analyses ............... 124
LIST OF FIGURES

Figure                                      page
6.2.1. Plot of the Time-Specific First-Order Sample  103
       Autocorrelations, Plus and Minus Two
       Standard Errors
6.2.2. Plot of Estimated Cell Means Over Time for  111
       the Two Subpopulations
6.3.1. Plot of the Adjusted Cell Means Over Time for  115
       the Control and Obese Subpopulations
Chapter I
INTRODUCTION

Epidemiologists, medical researchers, and other investigators generally recognize two approaches to drawing inferences concerning a population. The first, called a "cross-sectional" study, takes observations from a sample (or, "cross-section") of individuals at a fixed point in time. It is a portrait of the population at that particular moment, and in view of the fact that a number of units has been observed, it provides a broad perspective of the underlying characteristics of the population. A second approach is to follow an observational unit over time. The advantage of this approach is that the natural evolution of the underlying mechanism can be made apparent.

A longitudinal study attempts to combine the salutary features of both these approaches into an overall inference strategy. A representative subset of population elements is identified, and each observational unit is followed over a period of time. This is a comprehensive strategy, for not only does it reveal the broad characteristics of the population, but it is also possible to chart the natural evolution of these characteristics. The set of response measures may therefore be summarized in a two-dimensional array indexed by both the individual and the time point at which it had been observed.

Longitudinal designs enjoy a wide constituency, and appear under a number of aliases in a variety of scientific disciplines. A psychologist might refer to this strategy as a "repeated measures" analysis; an econometrician, as a "time series of cross-sections"; a biostatistical application, which anticipates systematic change over time, is
called "growth curve analysis". Thus, one might be interested in analyzing monthly cholesterol readings from a set of subjects enrolled in a clinical trial, or the hourly demand for electricity in a sample of households, or a periodic personality characteristic from a set of individuals under treatment for schizophrenia, and so on.

It is worth noting specifically at this point that the underlying dimension indexing the sequence in which observations are manifested need not necessarily be time, and in general, observations may be gathered along any continuum. An obvious example is distance, and an experiment could be mounted to measure the fuel economy of a fleet of vehicles at equally-spaced intervals of distance.

There are several statistical tools which can utilized for this model. When observations are captured at equally-spaced intervals of time, "time series analysis" offers a useful environment in which to model the underlying evolutionary mechanism. The most general time series model is called the "Autoregressive-Moving Average (ARMA)" representation, and it is introduced in Chapter II. The ARMA model results in a particular covariance structure among the repeated measures from each observational unit. The main advantage is that this covariance structure is dependent upon a small number of parameters, generally no more than three or four, and this appeals to one's sense of parsimony. Moreover, the general linear model is available to relate this phenomenon to other variables which are thought to "explain" or influence its behaviour. The marriage of these two data-analytic devices therefore provides a powerful inferential vehicle.

Missing data are a common occurrence in applied statistical research, and they exert a strong influence in this environment as well. Beyond the computational complexity introduced by incomplete data, there are the broader and more important statistical issues. Missing data cast doubt on the representativeness of the sample observed and the exter-
nal validity of the inferences, leaving conclusions vulnerable to charges of bias and statistical inefficiency. Nevertheless, while every effort may be made to retrieve all the information, it frequently happens that the data are not available or have not been gathered in the proper format.

It is against this backdrop therefore that the current study is being undertaken. The main research effort establishes a unified inference strategy for the analysis of the incomplete longitudinal design under an ARMA covariance structure. The issue of immediate concern is the joint estimation of the indigenous parameters. A review of the literature (Chapter II) recommends the Maximum Likelihood (ML) estimation criterion. ML estimators are observed to enjoy favourable large-sample properties, and a substantial body of evidence is introduced to show that they can be recommended on the basis of their small-sample properties as well.

Maximum likelihood estimation for the current model is considered in Chapter III. The first-order ML equations are presented, with particular attention paid to the linear model parameter vector. The ML estimator of this quantity is shown to result in an unbiased estimator under certain assumptions. Moreover, the familiar large-sample properties of the ML estimator are upheld under this model.

ML estimation under an ARMA covariance model is not a straightforward analytic procedure. There is no closed-form expression for the ML estimator of the ARMA time series parameter vector, and one must resort to algorithmic, numerical methods techniques to iterate to a solution. The mechanics necessary to implement these calculations are detailed in Chapter IV.

A second and equally important component of this inference strategy is to specify procedures for testing hypotheses among the indigenous parameters. This is the topic of Chapter V. There are two broad categories which are studied
in depth. The first examines the covariance structure among the repeated measures in an effort to verify that the ARMA covariance model is consistent with the data. The second considers the more substantive hypotheses among the linear model parameters, for assessing differences across subpopulations, for patterns over time, and for the influence of concomitant variables. Large-sample likelihood ratio tests are presented, together with their asymptotic $\chi^2$ test statistics.

Finally, two datasets are studied in Chapter VI to illustrate these procedures. The first consists of data generated at random from a known stochastic mechanism. The second is derived from a paper published in the statistical literature. The structures inherent in these datasets are largely upheld by the methodology, representing a vote of confidence in the techniques. Moreover, this chapter also serves a useful pedagogical purpose, illustrating model-building procedures under this particular model.
Chapter II
REVIEW OF THE LITERATURE

2.1 UNIVARIATE REGRESSION MODEL
2.1.1 Time Series of Observations

Let $Y_t$ be a random variable which denotes the outcome or response measure at the t-th time point for a particular process or experiment under observation ($t=1,2,\ldots,T$). Assume that T such observations have been gathered, and that they have been collected at equally-spaced intervals of time, namely, $t_0+\delta$, $t_0+2\delta$, $t_0+T\delta$, for some time origin $t_0$ and time increment $\delta$. Define the random vector $Y' = [Y_1, Y_2, \ldots, Y_T]$ to represent the collective set of measurements arising from the experiment, and $Y' = [y_1, y_2, \ldots, y_T]$ to represent the realizations from these random variables. The set of random variables $\{Y_t\}$ may be considered as a "time series" of observations.

Time series have typically arisen in econometric applications. For example, the Gross National Product of a country might be considered over a period of years, or the monthly unemployment rates may be analyzed over a period of time, and so on. In a public health setting, a researcher may wish to follow hourly blood pressure readings, or cholesterol levels may be charted over the course of a clinical trial. The salient features of this design revolve around the repeated sampling from an on-going process at equally-spaced intervals of time.

The observed time series $\{Y_1, Y_2, \ldots, Y_T\}$ may be regarded as having been generated by some stochastic mechanism. It represents a particular realization from an infinite number of such realizations that could have been generated. The mechanism itself may be thought of as extending over the en-
time domain, although only those observations sampled at $t_0 + \delta, \ldots, t_0 + T\delta$ are actually available for analysis.

In order to interpret the behaviour of these observations, the broad theory of the general linear model is introduced. Consider the model

$$Y_t = \mathbf{x}_t' \mathbf{\beta} + u_t \quad (t=1,2,\ldots,T) \quad (2.1.1)$$

where $\mathbf{x}_t$ is a ($Q$ by $1$) vector of explanatory (or "exogenous") variables whose behaviour are thought to influence variation in $Y_t$, $\mathbf{\beta}$ is a ($Q$ by $1$) vector of unknown coefficients whose values measure the contributions of the corresponding explanatory variables and for whose values estimates are desired. The term $u_t$ is an unobserved random variable called the "disturbance" term, or error term, representing variation in the response variable unaccounted for by the independent variables. Gathering these observations, the usual linear model arises:

$$\mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{u} \quad (2.1.2)$$

where $\mathbf{y}$ and $\mathbf{\beta}$ have been defined above, $\mathbf{X}' = [\mathbf{x}_1' \mathbf{x}_2' \ldots \mathbf{x}_T']$ is a ($T$ by $Q$) design matrix, with full column rank, of realizations from the explanatory variables, and $\mathbf{u}$ is a ($T$ by $1$) vector of disturbance terms.

In light of the time series nature of the data, the ordinary least squares (OLS) assumption that the $\{Y_t\}$ are independent random variables is no longer tenable. With systematic variation accounted for by the deterministic component $\mathbf{X}\mathbf{\beta}$, it is assumed that the error terms $\{u_t\}$ observe some "time series" pattern. Box & Jenkins (1970) provided a useful taxonomy of time series models. An autoregressive process of order $p$, AR($p$), is written as:

$$u_t = \phi_1 u_{t-1} + \cdots + \phi_p u_{t-p} = \epsilon_t \quad (2.1.3)$$

where $\phi_p \neq 0$. That is, the current value of the error term $u_t$ is modelled as an explicit weighted function of the $p$ previous error terms ("autoregression"), plus a "random shock",
\[ u_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \ldots - \theta_q \epsilon_{t-q} \]  
\[ \text{(2.1.4)} \]

where \( \theta_q \neq 0 \). While \( u_t \) is, in a sense, a "weighted average" of the \( q \) previous shocks, the weights are neither necessarily positive nor do they sum to 1, however, the usage is widely accepted.

These two models are special cases of the most general time series model, the mixed Autoregressive - Moving Average process of order \((p,q)\), i.e., \(\text{ARMA}(p,q)\). It takes the form:

\[ u_t - \varphi_1 u_{t-1} - \ldots - \varphi_p u_{t-p} = \epsilon_t - \theta_1 \epsilon_{t-1} - \ldots - \theta_q \epsilon_{t-q} \]  
\[ \text{(2.1.5)} \]

where \( \varphi_p, \theta_q \neq 0 \). Thus, the pure AR and pure MA models are special cases of this model where \( q = 0 \) and \( p = 0 \) respectively.

Box & Jenkins defined the "backshift" operator, \( B \), with definition:

\[ B^k(u_t) = u_{t-k} \quad (k=0,1,2,...) \]  
\[ \text{(2.1.6)} \]

The \(\text{ARMA}(p,q)\) model may be expressed as:

\[ \phi(B) u_t = \theta(B) \epsilon_t \]  
\[ \text{(2.1.7)} \]

where \( \phi(B) = 1 - \varphi_1 B - \ldots - \varphi_p B^p \) and \( \theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q \) are polynomials of order \( p \) and \( q \) (respectively) in \( B \). Excepting special econometric applications, \( p \) and \( q \) rarely take on values greater than 2.

The set of random variables \( \{ \epsilon_t \} \) are called "random shocks" and are collectively referred to as the "white noise" process. It is assumed that the \( \{ \epsilon_t \} \) are i.i.d. random variables with mean 0 and common variance \( \sigma^2 \) \((0 < \sigma^2 < \infty)\). Letting \( \epsilon' = [\epsilon_1 \ldots \epsilon_T] \), this implies that

\[ E(\epsilon) = 0, \]  
\[ \text{(2.1.8)} \]

and

\[ \text{Var}(\epsilon) = \sigma^2 I_T \]  
\[ \text{(2.1.9)} \]
where \( I_T \) is the \((T \times T)\) identity matrix.

It is assumed that all first-order moments \( \mu = E(u_t) \) exist for \( t=1,2,\ldots,T \), and are independent of the time parameter \( t \), i.e., \( E(u_t) = E(u_s) \) for all \( s,t=1,2,\ldots,T \). Similarly, all second-order moments \( \gamma_j = \text{Cov}(u_t, u_{t+j}) \) are assumed to exist for \( t=1,2,\ldots,T \), and positive integers \( j \), and are invariant to shifts along the time axis. It can easily be shown from this assumption that \( \gamma_j = \gamma_{-j} \). These two conditions define "weak" stationarity. A stronger condition, "strict" stationarity, occurs when the probability density of any non-empty subset of these random variables is invariant to shifts along the time axis. Inasmuch as the normal distribution is completely specified by its first- and second-order moments, weak stationarity with normality is equivalent to strict stationarity.

Box & Jenkins demonstrated that a necessary and sufficient condition for the series to be stationary is for all the roots of \( \phi(B) = 0 \) (some of which may be complex) to lie outside the unit circle in the complex plane. A related condition is called "invertibility", and a necessary and sufficient condition for the process to be invertible is for the roots of the equation \( \theta(B) = 0 \) to lie outside the unit circle in the complex plane. It is assumed explicitly that the process is stationary and invertible; moreover, to avoid spurious results, it is further assumed that \( \phi(B) \) and \( \theta(B) \) have no roots in common.

The time series process among the elements of \( u \) imposes a particular covariance structure upon them. For example, considering the MA(1) process, \( u_t = \varepsilon_t - \theta \varepsilon_{t-1} \), it can be shown that \( E(u_t) = 0 \), (for \( t=1,2,\ldots,T \)), and

\[
\text{Cov}(u_t, u_{t+j}) = \begin{cases} 
\sigma^2 (1 + \theta^2) & \text{if } j=0 \\
-\sigma^2 \theta & \text{if } j=\pm 1 \\
0 & \text{otherwise}
\end{cases} \quad (2.1.10)
\]
Thus, \(E(\mathbf{u})=0\) and \(\text{Var}(\mathbf{u})=\sigma^2 \mathbf{G}\), where

\[
\mathbf{G} = \begin{bmatrix}
1+\theta^2 & -\theta & 0 & \ldots & 0 & 0 \\
-\theta & 1+\theta^2 & -\theta & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & -\theta & 1+\theta^2 \\
\end{bmatrix} \tag{2.1.11}
\]

and \(\mathbf{G}=\mathbf{G}(\theta)\) is an explicit matrix function of the MA parameter \(\theta\). Note that all the elements along any diagonal of \(\mathbf{G}\) are the same; a matrix with property is called a "Toeplitz" matrix. For the MA(2) case, it can similarly be shown that \(E(\mathbf{u})=0\), and that the covariance matrix may be expressed as \(\text{Var}(\mathbf{u})=\sigma^2 \mathbf{G}(\theta_1, \theta_2)\), where

\[
\mathbf{G} = \begin{bmatrix}
1+\theta_1^2+\theta_2^2 & -\theta_1+\theta_1\theta_2 & -\theta_2 & 0 & \ldots & 0 \\
-\theta_1+\theta_1\theta_2 & 1+\theta_1^2+\theta_2^2 & -\theta_1+\theta_1\theta_2 & -\theta_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
-\theta_1+\theta_1\theta_2 & -\theta_1+\theta_1\theta_2 & \ldots & 1+\theta_1^2+\theta_2^2 \\
\end{bmatrix} \tag{2.1.12}
\]

(sym)

For the AR(1) process, \(u_t=\phi u_{t-1}+\epsilon_t\), it also follows that \(E(\mathbf{u})=0\). In this case, \(\text{Var}(\mathbf{u})=\sigma^2 \mathbf{G}(\phi)\), with

\[
\mathbf{G} = (1-\phi^2)^{-1} \begin{bmatrix}
1 & \phi & \phi^2 & \ldots & \phi^{T-1} \\
\phi & 1 & \phi & \ldots & \phi^{T-2} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \ldots & 1 \\
\end{bmatrix} \tag{2.1.13}
\]

This is a Toeplitz matrix function of the AR parameter \(\phi\). Box & Jenkins provided similar expressions for other time series models.

In general, for any ARMA(p,q) process, it can be shown that \(E(\mathbf{y})=0\), and \(\text{Var}(\mathbf{y})=\sigma^2 \mathbf{G}(\phi)\) where \(\mathbf{G}\) is a Toeplitz function exclusively of the time series parameter vector \(\psi' = [\phi' \ \theta']\). It follows from these results that \(E(\mathbf{y})=\mathbf{X}\phi\), and \(\text{Var}(\mathbf{y})=\sigma^2 \mathbf{G}\). A succinct way therefore of summarizing these
results is to say that \( y \) satisfies the conditions of the General Linear Univariate Model, i.e., \( GLUM-\mathbf{FR}[y, X\theta, \sigma^2 \mathbf{G}(\psi)] \).

2.1.2 Estimators of the Parameters

To derive maximum likelihood estimators, it is incumbent to specify a form for the distribution of the random variables. The usual assumption of normality is retained, and it is specifically assumed that the \( \{\epsilon_i\} \) are i.i.d. \( N(0, \sigma^2) \) random variables. Since each \( u_t \) can be written as a linear combination of the white noise components, it follows that \( u \) must observe the \( T \)-variate multinormal distribution.

The likelihood function is proportional to the joint density function of the observations, and taking logarithms, it follows that

\[
\Lambda(\theta) = C - (T/2)\ln(\sigma^2) - (1/2)\ln|\mathbf{G}| - \frac{(y-X\hat{\theta})' \mathbf{G}^{-1}(y-X\hat{\theta})}{2\sigma^2},
\]

(2.1.14)

where \( \theta' = [\theta' \ \psi \ \sigma^2] \), and \( C \) is a constant independent of the parameters. Taking the derivative with respect to \( \theta \), and setting this expression to \( 0 \), it follows that

\[
\hat{\theta}(\psi) = [X'\mathbf{G}^{-1}(\psi)X]^{-1} [X'\mathbf{G}^{-1}(\psi)y].
\]

(2.1.15)

The MLE of \( \sigma^2 \) can similarly shown to be

\[
\hat{\sigma}^2(\psi) = \frac{(y-X\hat{\theta})' \mathbf{G}^{-1}(\psi)(y-X\hat{\theta})}{T},
\]

(2.1.16)

where \( \hat{\theta} \) is the MLE of \( \theta \). Clearly, both estimators are functions of the time series parameter vector \( \psi \).

If \( \psi \) were known, the these MLE's would be calculated directly from these expressions. Since the covariance matrix \( \mathbf{G} \) is a square matrix of order \( T \) (the number of observations), evaluating \( \mathbf{G}^{-1} \) poses a formidable computational obstacle even for moderate sample sizes. One method to avoid this difficulty is to find a factorization of \( \mathbf{G}^{-1} \) which obviates the need for this inversion. For example, considering the AR(1) error process, it can be shown that \( \mathbf{G}^{-1} = \mathbf{P}'\mathbf{P} \), where
\[
\mathbf{P} = \begin{bmatrix}
(1-\phi^2)^{1/2} & 0 & 0 & \ldots & 0 & 0 \\
-\phi & 1 & 0 & \ldots & 0 & 0 \\
0 & -\phi & 1 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & -\phi & 1
\end{bmatrix}
\]  
(2.1.17)

Taking the transformations \( y^* = \mathbf{Py}, \; x^* = \mathbf{PX}, \) and \( u^* = \mathbf{Pu}, \) and observing that the Jacobian of the transformation is \( \frac{\partial y}{\partial y^*}' = |\mathbf{P}|^{-1} = |\mathbf{G}|^{1/2}, \) the log likelihood becomes

\[
\Lambda(\phi, \gamma, \sigma^2) = c - (T/2) \ln(\sigma^2) - (y^* - x^* \hat{\phi})'(y^* - x^* \hat{\phi}) / 2\sigma^2,
\]  
(2.1.18)

This is the log likelihood function arising from the GLUM-\text{FR}(y^*, x^* \hat{\phi}, \sigma^2 \mathbf{I}) with normality. The ML estimators under this AR(1) error process are therefore derived as

\[
\hat{\phi}(\gamma) = (x^* x^*)^{-1} x^* y^*
\]  
(2.1.19)

and

\[
\hat{\sigma}^2(\gamma) = (y^* - x^* \hat{\phi})'(y^* - x^* \hat{\phi}) / T.
\]  
(2.1.20)

The transformation implied by \( \mathbf{P} \) is easily programmed as

\[
y_{t^*} = \begin{cases}
(1-\phi^2)^{1/2} y_1 & \text{for } t=1 \\
y_t - \phi y_{t-1} & \text{for } t=2,3,\ldots,T
\end{cases}
\]  
(2.1.21)

and, for any \( k=1,2,\ldots,Q, \)

\[
x_{t,k^*} = \begin{cases}
(1-\phi^2)^{1/2} x_{1,k} & \text{for } t=1 \\
x_{t,k} - \phi x_{t-1,k} & \text{for } t=2,3,\ldots,T,
\end{cases}
\]  
(2.1.22)

Judge et al.(1980) provide similar expressions for the AR(2) case.

When \( \phi \) is unknown, estimation is considerably more difficult. Historically, the first estimation procedures were the "two-stage" techniques. In the first stage, a "reason-
able" estimator \( \hat{\phi} \) is substituted for \( \phi \) in (2.1.21) and (2.1.22), whereupon these quantities are used in (2.1.19) and (2.1.20). For example, the Cochrane & Orcutt (1949) method prescribed the following steps:

1. First, compute the OLS estimator \( \hat{b} = (X'X)^{-1}X'y \), and compute the corresponding residuals \( \hat{u} = y - X\hat{b} \).

2. Compute the first-order sample autocorrelation \( r = \frac{\hat{u}_t \hat{u}_{t-1}}{\hat{u}_t^2} \), and let \( \hat{\rho} = r \).

3. Then substitute this quantity into (2.1.21) and (2.1.22) and compute \( \hat{\phi}(\hat{\rho}) \) and \( \hat{\sigma}^2(\hat{\rho}) \) from (2.1.19) and (2.1.20).

There are several other minor variations on this two-step estimator, and they are detailed in Judge et al.

The Maximum Likelihood estimators for \( \phi \) and \( \sigma^2 \) consist of substituting the ML estimator for \( \hat{\phi} \) into (2.1.15) and (2.1.16). It is a very much more difficult proposition, however, to derive an analytic expression for the ML estimator of \( \phi \). It enters the likelihood function through the matrix \( G \), and is therefore present not only in the quadratic form \( (y - X\hat{\phi})'G^{-1}(y - X\hat{\phi}) \), but also through the determinant \( |G(\hat{\phi})| \). Moreover, the resulting equations form a set of non-linear simultaneous equations in \( \phi \). There is no closed-form solution to these equations, and one must resort to algorithmic numerical methods routines to iterate to the ML estimates. These techniques are discussed in some detail in the next section.

Alternatively, because the determinant in the likelihood function is dominated by the quadratic form in the exponent, especially for large \( T \), Box & Jenkins suggested that the contours of the likelihood function in the parameter space are very nearly those of the sum of squares function, \( S(\phi, \psi) = u'G^{-1}(\psi)u \), where \( u = y - X\hat{\phi} \). Estimators which minimize this function are called "Least Squares (LS) Estimators" and can be expected to provide close approximations to the maximum
likelihood estimates. Box & Jenkins showed that this function may be expressed as:

\[ S(\theta, \hat{\phi}) = u'G^{-1}u = \xi[\epsilon_t]^2, \]

(2.1.23)

with the summation extending from \(-\infty\) to \(T\), and where 
\[ \xi_t = E[\epsilon_t / u, \theta, \phi] \]
is the expectation of \(\epsilon_t\) under the model. In practice, the lower limit is replaced by \(-R\) for some suitably large integer \(R\) beyond which the \(\{\epsilon_t\}\) are expected to be negligible.

2.1.3 Theoretical and Empirical Properties of these Estimators

Theil(1971), Ch. 8] considered in detail the two-stage estimation procedure under an AR(1) error structure. \(\hat{\phi}=r\) is shown to be a consistent estimator of the autoregressive parameter, and under suitable regularity conditions, it follows that the estimators \(\hat{\phi}(\hat{\phi})\) and \(\hat{\sigma}^2(\hat{\phi})\) are asymptotically normal, unbiased and efficient.

For a pure time series model with no deterministic component (that is, \(\phi=0\) identically), the Maximum Likelihood estimator has been shown to be consistent and asymptotically normal [Whittle(1951), Ch. 7], and asymptotically efficient [Aigner(1971), p. 361]. In the absence of normality, the LS estimator of the time series parameters \(\phi\) has been proven to have the same asymptotic distribution as the ML estimator [Walker(1964) and Whittle(1962)]. This is a powerful result, and indicates that the LS and ML estimators are asymptotically equivalent.

The univariate regression model under an ARMA covariance pattern has been shown to have similar properties. Hildreth(1969) demonstrated that under the AR(1) model, the ML estimator of the linear model parameters \(\phi\) is asymptotically normal, unbiased, and efficient, and is asymptotically uncorrelated with (i.e., asymptotically independent of) the autoregressive parameter \(\phi\).
Pierce (1971) considered the LS estimator of the joint set of parameters \( \theta' = [a' \ b' \ \sigma^2] \). It was shown that the joint set of estimators is asymptotically normal, irrespective of the underlying distribution of the white noise process \( \{\epsilon_t\} \); moreover, it is asymptotically unbiased. The time series parameter estimator \( \hat{\theta} \) is asymptotically uncorrelated with (i.e., asymptotically independent of) the linear model parameter estimator \( \hat{\beta} \), and observes the same limiting distribution as that arising from the pure time series with no deterministic component. "Thus, any large-sample properties of the estimates in [the pure time series model], and associate procedures for constructing tests and confidence intervals, also hold ... in the regression model with stationary errors" (p. 311).

Magnus (1978) considered a slightly more general case where the covariance matrix \( \Sigma \) is a function of a smaller set of parameters \( \phi \), without explicitly specifying the functional form. He proved that the ML estimators \( [\hat{\theta}' \ \hat{\phi}'] \), under general conditions, are (weakly) consistent, asymptotically normally distributed, and asymptotically efficient. Don & Magnus (1980) further showed that \( \hat{\theta} \) is unbiased.

Whereas the large-sample properties of all three estimators are essentially equivalent, some important differences begin to emerge for smaller sample sizes. Considering the linear model with an AR(1) error process, Griliches & Rao (1969) suggested that the two-stage estimators represent an improvement over the OLS estimator for \(|\phi| > 0.30\). Beach & MacKinnon (1978a) however, described an algorithm for performing ML estimation under this model, and demonstrated the superiority of this estimator over the two-stage econometric estimators. Spitzer (1979) also considered this model, and found that both LS and ML estimators are superior to the two-stage techniques, and recommended ML estimation. Beach & MacKinnon (1978b) considered the AR(2) error process, and in a small simulation study, also preferred ML estimation.
Pagan & Nicholls(1976) considered the univariate regression model under a MA(1) error structure. An algorithm for evaluating the associated likelihood function was presented, and it was concluded that both LS and ML estimation techniques performed reasonably well under this model. ML estimates not only observed the invertibility conditions, but also required fewer iterations to converge.

Several Monte Carlo studies have investigated pure time series models. For example, Nelson(1974) compared the LS to the ML estimation procedure for a MA(1) process, and found that LS estimators exhibited reasonably good properties. For the general moving average process, Osborn(1976) strongly endorsed non-linear optimization routines in general and ML estimation in particular, and she provided evidence for the superiority of ML estimators under this model.

An extensive investigation into LS and ML estimation techniques for a wide range of time series models was reported by Dent & Min(1978). In each case, a time series of size T=100 was generated, and the properties of these estimators, based upon 100 replications, were considered. They concluded that the extra computational effort required for ML estimation appears to be justified. While little difference was observed among these estimators for the AR(1) model, the ML estimator was clearly superior for higher-order AR processes. For MA(1) and MA(2) models, it was difficult to discern any differences between LS and ML estimation, while the former was marginally favoured for the mixed ARMA(1,1) model.

Some studies have captured the behaviour of these estimators near boundary conditions. For the AR(1) process, for example, stationarity considerations constrain the AR(1) parameter $|\phi|<1$. Beach & MacKinnon(1978a) studied the determinant of $G$ for the AR(1) error process. It was shown that this function effectively constrains the parameter value to remain within the stationarity region, and that the statio-
narity condition is therefore "... incorporated a priori in
the likelihood function itself" (p. 57). This is not the
case for LS estimation, and provides strong prima facie evi-
dence for preferring ML estimation. A similar result for
the AR(2) model was documented in Beach & MacKinnon(1978b).

Both Osborn(1982) and Damsleth(1983) considered the theo-
retical properties of the LS and ML estimators towards the
extremes of the parameter space for the general MA case. By
considering the expected value of the sum of squares func-
tion, it was shown that the LS estimator results in a biased
estimate of \( \theta \) for true \( \theta \) near the boundary conditions. The
expected value of the likelihood function, however, suggests
that the maximum value will be achieved at the true value of
\( \theta \). Damsleth concluded that the LS estimation technique de-
tailed in Box & Jenkins may yield rather biased estimates,
and lead to numerical difficulties if the sample size is
small and/or \( \theta \) is close to the boundary. He recommended
that "...exact maximum likelihood methods should be applied
when estimating time series models" (p. 127).

To summarize, the maximum likelihood estimator is an ob-
vious candidate for the parameters of univariate regression
model under an ARMA covariance structure. Two-stage estima-
tors and least squares estimators were largely introduced to
simplify the computations. While they share asymptotic
properties, most workers tend to favour the maximum likeli-
hood estimator on both a theoretical and empirical basis.
MLE's revealed superior small-sample properties relative to
the two-stage techniques and their least squares counter-
parts. Monte Carlo evidence suggested that they were fre-
quently better in terms of bias and MSE.
2.2 NUMERICAL METHODS TECHNIQUES

2.2.1 Non-Linear Optimization Procedures

It has been stated that differentiating the log likelihood function with respect to the time series parameter vector $\mathbf{\theta}$ results in a set of non-linear simultaneous equations. Except in the most trivial cases, no closed-form solution exists to these equations, and one must therefore resort to algorithmic, numerical methods to iterate to an optimal solution.

A thorough review of the techniques to actually perform non-linear optimization is, however, beyond the scope of this study. Fletcher(1969) and Murray(1972) both provide extensive summaries of the issues and procedures in non-linear optimization up to ten years ago. More recent compendiums of the methodology have been prepared by Bard(1974), Powell(1981), and in two volumes for unconstrained and constrained optimization by Fletcher(1980,1981). Moreover, some textbooks on the general linear model, e.g., Draper & Smith[(1981), Ch. 10], as well as more advanced econometrics textbooks, e.g., Goldfeld & Quandt[(1972), Ch. 1] and Judge et al[(1980), Ch. 17] devote a chapter towards a basic understanding of these algorithms.

Let $\mathbf{\lambda}$ be an ($r$ by 1) vector of parameters, and $\Lambda(\mathbf{\lambda})$ be the criterion function for which an optimizing value, say $\hat{\mathbf{\lambda}}$, is desired. In our case, $\mathbf{\lambda}$ would represent the vector of time series parameters. $\Lambda$ will generally represent the $-\log$ likelihood function, or some transformation of it; however, it could equally well represent the sum of squares function in a non-linear least squares estimation problem. Thus, without loss of generality, this discussion will be considering minimizing the criterion function, since maximizing a criterion function is equivalent to minimizing its negative.

Considering the second-order Taylor series expansion of $\Lambda$ about $\lambda_u$, under suitable regularity conditions, one may write
\( \Lambda(\Lambda) \approx \Lambda(\Lambda_u) + (\Lambda - \Lambda_u)' \mathbf{q}(\Lambda_u) \\
+ (1/2)(\Lambda - \Lambda_u)' \mathbf{H}(\Lambda_u)(\Lambda - \Lambda_u). \) 

(2.2.1)

\( \mathbf{q}(.) = \partial \Lambda / \partial \Lambda \) is called the "gradient vector" with typical element \( \partial \Lambda / \partial \Lambda_i \) \((i=1,2,...,r)\). \( \mathbf{H}(.) = \partial^2 \Lambda / \partial \Lambda_i \partial \Lambda_j \) is the matrix of mixed second-order partial derivatives, called the "Hessian matrix", whose \((i,j)\)th element is \( \partial^2 \Lambda / \partial \Lambda_i \partial \Lambda_j \). Both these functions are evaluated at the constant \( \Lambda_u \) in the above expression.

Taking the derivative w.r.t. \( \Lambda \) and setting this expression to \( \mathbf{0} \), and bearing in mind that the derivative of a constant is zero, it follows that

\[ \mathbf{0} = \partial \Lambda / \partial \Lambda \approx \mathbf{q}(\Lambda_u) + \mathbf{H}(\Lambda_u)(\Lambda - \Lambda_u), \]

(2.2.2)

so that

\[ \Lambda \approx \Lambda_u - \mathbf{H}^{-1}(\Lambda_u) \mathbf{q}(\Lambda_u) \]

(2.2.3)

In general, estimates are updated at each iteration by writing

\[ \Lambda^{(u+1)} = \Lambda^{(u)} - t \mathbf{P}^{-1}(\Lambda^{(u)}) \mathbf{q}(\Lambda^{(u)}), \quad (u=0,1,2,...) \]

(2.2.4)

where the scalar \( t \) is called the "step length", and \( \mathbf{P}(.) \) is a positive-definite matrix, called the "step direction" matrix, which provides a reasonable approximation to \( \mathbf{H} \).

Generally, it is preferable to make analytic expressions for the derivatives available to a computer routine, however, derivatives may be difficult to specify analytically, or they may be so complicated that they themselves become a source of error or consume a large amount of computer resources. Numerical derivatives are found by perturbing the parameter element by a small amount and observing the relative change in the criterion function. While they may be easier to program, they require a large number of function evaluations which again consumes a great deal of computer resources. Belsley(1980) found that in terms computer time
alone, the extra cost incurred in specifying analytic derivatives was amply rewarded by the ensuing computational efficiency.

It can be shown that initially the steepest descent to the global value is obtained by taking \( P = I_r \), i.e., the \( r \)-dimensional identity matrix, and \( t = 1 \). This method is therefore called the "Method of Steepest Descent". However, it fails to take into account the subtle contours of the criterion function, and while it makes good initial progress, it converges rather slowly thereafter.

The Newton-Raphson method explicitly takes \( P = H \), i.e., the Hessian matrix evaluated at \( \lambda^{(u)} \), and \( t = 1 \). This technique exhibits fast local convergence and has been used to advantage over a wide range of applications. However, in addition to the difficulty of making second derivatives available to the algorithm, the Hessian itself may only be positive-definite inside a small neighbourhood about the optimum value. A variant of the Newton-Raphson technique for maximum likelihood estimation problems is to use

\[
P = -E[\partial^2 \lambda / \partial \lambda \lambda']
\]

(2.2.5)
evaluated at \( \lambda^{(u)} \). The Hessian is approximated by its expected value, and this is known as the "Method of Scoring" [Rao(1973)].

Several techniques have been proposed to overcome difficulties inherent in the Newton-Raphson procedure. Rather than estimating the inverse Hessian at every step, one can simply update its value from iteration to iteration. One example of this approach is the Davidon-Fletcher-Powell method [Davidon(1959) and Fletcher & Powell(1963)]. The Marquardt algorithm [Marquardt(1963)] can be used to assist those procedures which do not guarantee a positive-definite direction matrix. This approach uses the fact that for any matrix \( P \), \( P + vP^* \) is always positive-definite if \( P^* \) is positive-definite and the scalar \( v \) is sufficiently large. An
obvious choice for \( P^* \) is simply the identity matrix, and the algorithm typically begins with large values for \( v \), allowing the method of steepest descent to make good initial progress, and gradually decreases its value capitalizing on the fast local convergence of the Newton-Raphson technique.

One technique which avoids computing second derivatives is the Gauss-Newton method, and it is described in Draper & Smith (1981). At each step, the non-linear function is approximated by a linear surrogate, so that estimates are expeditiously updated based on OLS theory. The technique is obviously dependent upon how successful the linear approximation is in a small neighbourhood about the global minimum. In practice, it shows good convergence. Moreover, the Marquardt strategy is frequently coupled with the Gauss-Newton approach.

Bard (1974, Sec. 5-4) considered the theoretical properties of the iteration sequence (2.2.4). Provided that \( \lambda \) has continuous first-order, and bounded second-order derivatives, the sequence of estimators will converge to a stationary point. No more definitive statement is forthcoming, however. The criterion function itself must be examined to determine whether a minimum or a maximum, global or local, or a saddle point has been reached. A global optimum value can be assured only if the criterion function has no other stationary point. In practice, a maximum likelihood algorithm will almost surely converge to at least a local maximum value.

Jennrich (1969) considered the convergence properties of the Gauss-Newton algorithm and found that provided the initial estimate of \( \lambda \) is sufficiently "close" to \( \hat{\lambda} \), the algorithm will converge to \( \hat{\lambda} \). Jennrich quoted from Hartley & Booker (1965) who suggested that iteration should begin from a consistent estimator of \( \lambda \).

Conventional wisdom suggests that the gradient procedures are reasonably robust with respect to starting values.
Belsley (1980) recommended the intuitive and expeditious OLS parameters, that is, \( b = (X'X)^{-1}X'y \), and \( \psi(0) = 0 \). Nicholls, Pagan & Terrell (1975) observed that this may lead to difficulty, especially for small sample sizes, and provided that resources are available, recommended starting from several values. Box & Jenkins described a procedure, based on a method due to Wilson (1969), for computing an initial estimate of the time series vector \( \psi \) based upon the sample autocorrelations of the observed time series.

In practice, all these non-linear optimization routines appear to be effective in estimating the parameters of a model with this particular structure. Box & Jenkins outlined a Marquardt procedure for estimating the parameters of a pure time series model, and both SAS [SAS Institute (1982b), p. 87] and BMDP [Dixon (1983), p. 690] implemented their respective versions of Box & Jenkins time series analysis using the Gauss-Marquardt approach. Ansley & Newbold (1980) also applied this technique in their Monte Carlo study, while Kang (1975) employed the Marquardt technique for the general MA(q) model. Gauss-Newton approach was adopted by Osborn (1976) for the pure MA(q) process.

These procedures appear to carry over well in the estimation of the linear model with an ARMA error structure. For the AR(1) case, Spitzer (1979) used a Gauss-Newton algorithm with derivatives evaluated analytically. Nicholls, Pagan & Terrell (1975) in their survey article, recommended either the Gauss-Newton or the Newton-Raphson algorithm. The Method of Scoring was strongly endorsed in Sandland & McGilchrist (1979) and McGilchrist et al. (1981).

2.2.2 Evaluating the Likelihood Function

The heart of any iterative maximum likelihood estimation routine is, for any specified parameter values, to evaluate the (log) likelihood function. This allows one to monitor the progress of the likelihood function towards its maximum
value, and in the absence of analytic expressions for the first- and second-order derivatives, to evaluate these quantities numerically. Moreover, since these routines are imbedded within an iterative procedure, such as those outlined in the previous section, an efficient evaluation routine will reap the greatest benefits in terms of computer costs.

The Cochrane-Orcutt transformation matrix $P$ such that $G^{-1}=P'P$ had been presented for the AR(1) error process. Savin & White(1978) presented a similar factorization matrix for the AR(1) error process when there is a gap of consecutive missing values. Pesaran(1973) studied the linear model with a MA(1) error process, and by considering the spectral decomposition of $G$, a transformation matrix was derived. This model was also considered by Belsley(1980), and in this case, a triangular decomposition was presented resulting in a recursion formula to evaluate the likelihood function. Nicholls, Pagan & Terrell(1975) and Pagan & Nicholls(1976) expanded this technique for the MA(q) error process, while the mixed ARMA(1,1) model was considered by Tiao & Ali(1971).

For higher ARMA model, however, no transformation matrices $P$ are immediately discernible. General algorithms were nevertheless presented for the pure ARMA process with no deterministic component by Newbold(1974) and Dent(1977). Ansley(1979) demonstrated that these algorithms are essentially equivalent, and described another approach, based upon the Cholesky decomposition of $G$, which results in substantial savings in computer time. This algorithm forms the basis for the ARMALIK function in SAS[SAS Institute(1982b), p. 205]. Nicholls & Hall(1979) similarly presented an algorithm to evaluate the general likelihood function of a multivariate ARMA time series.

In the past three years, a technique which is more frequently found in the engineering literature, called "Kalman
Filtering" [Kalman(1960) and Kalman & Bucy(1961)], has been applied for time series models. The Kalman Filter is an algorithm for sequentially updating estimates of a parameter vector under a particular stochastic model. Meinhold & Singpurwalla(1983) provided a useful overview of this technique from a statistical perspective.

Harvey & Phillips(1979) implemented the Kalman Filter approach for the linear model with an ARMA(p,q) error structure. A set of independent error terms whose values can be used directly in the evaluation of the likelihood function is produced. A FORTRAN subroutine, implementing this approach, was described in Gardner et al(1980).

A variation on this theme was considered by Jones(1980) for a pure time series model. In this case, Kalman Filtering yields a set of one-step ahead prediction error variances which are used directly in the evaluation of the likelihood function. Ansley & Kohn(1983) extended the Kalman Filter strategy to estimate multivariate time series models. It was claimed that the number of computations is O(T) per iteration.

The Kalman Filter methodology is also applicable for missing data problems. The algorithms described in Jones and Ansley & Kohn for the general ARMA model are specifically available.

2.3 LONGITUDINAL DATA DESIGNS
2.3.1 Expanded Model

The previous sections had been largely devoted to studying observations gathered at equally-spaced intervals of time from a single source. In this section, the model is expanded to allow for a time series of observations to be captured, not from one individual, but from a number of subjects. Thus, trends in the gross national product of a set of countries, monthly cholesterol levels from a set of subjects over the course of a clinical trial might be examined. First, the expanded model is presented. Then, the
literature in this area, particularly with regard to estimation considerations, is reviewed.

Let $Y_i(t)$ be a random variable which denotes the value observed from the $i$-th individual ($i=1,2,...,N$) at the $t$-th time point ($t=1,2,...,T$). It is assumed that observations have been gathered at equally-spaced intervals of time consistently for all observational units. These observations may be gathered into a two-dimensional array

$$
Y = \begin{bmatrix}
Y_1^{(1)} & \cdots & Y_1^{(T)} \\
\vdots & & \vdots \\
Y_N^{(1)} & \cdots & Y_N^{(T)}
\end{bmatrix} = \begin{bmatrix}
Y_1' \\
\vdots \\
Y_N'
\end{bmatrix} \quad (2.3.1)
$$

where rows refer to different observational units and columns refer to different time points, and $Y_i' = [Y_i^{(1)} \ Y_i^{(2)} \ \cdots \ Y_i^{(T)}]$ represents the set of repeated measures from the $i$-th individual ($i=1,2,...,N$).

This type of structure is sometimes referred to as a "time series of cross-sections" in the econometrics literature. That is, the set of individuals might define a representative "cross-section" of the population, each of which contributes a "time series" of observations. Other names include "longitudinal" data and "panel studies".

Hechman & Singer(1982) suggested that the distinguishing feature of longitudinal datasets in an econometrics setting is that a short time series is available for a large number of individuals. Kiefer(1980) observed that $N=5,000$ and $T=9$ characterize many currently available panel datasets. This feature would also seem appropriate in a public health setting. It would be unusual in a clinical trial to capture more than a dozen repeated measures from a single observational unit.

In order to interpret the behaviour of these random variables, the broad theory of the general linear model is again introduced. Let

$$
Y_i = Z_i b + u_i, \quad (2.3.2)
$$
where $Z_i$ is a matrix of non-stochastic explanatory variables whose values are thought to account for variation in the $Y_i(t)$; $\theta$ is an unknown vector of linear model parameters for whose values estimates are sought; and $u_i' = [u_{i1} \ldots u_{iT}]$ is a vector of unobserved error terms, representing variation in the $Y_i$ unaccounted for by the linear model.

The symbol $I$ is used for a covariance structure, and in the most general case, let

$$\text{Cov}(Y_i, Y_j) = I_{ij}, \quad (i,j=1,2,\ldots,N). \quad (2.3.3)$$

Define $Y' = [Y_1' \ldots Y_N']$ to be that (NT by 1) vector formed by successively stacking the set of repeated measures from each observational unit underneath each other. Then, one may write the model as

$$Y = Z\theta + U, \quad (2.3.4)$$

where $Z' = [Z_1' \ldots Z_N']$, and $U' = [u_1' \ldots u_N']$. Moreover,

$$\text{Var}(Y) = \Sigma, \quad (2.3.5)$$

where $\Sigma$ is a partitioned matrix whose $(i,j)$-th block is $I_{ij}$. It therefore follows that $Y$ satisfies the conditions of the GLUM-FR$(Y, Z\theta, \Sigma)$.

Both biostatistics and econometrics impose very particular structures on the first- and second-order moments of $Y$. For example, the General Linear Multivariate Model (GLMM) prescribes that different observational units are mutually uncorrelated, with a homogeneous covariance pattern across the repeated measures for all observational units. Thus,

$$\text{Cov}(Y_i, Y_j) = \begin{cases} I & \text{for } i=j \\ 0 & \text{for } i\neq j \end{cases}, \quad (2.3.6)$$

and $\Sigma$ is a block diagonal matrix with all diagonal blocks equal to $I$. The explanatory variables are assumed to be time-invariant (e.g., Sex), although their influence is allowed to vary over time. Using the matrix $Y$ of (2.3.1), the GLMM specifies that
\[ E(Y) = XB, \quad (2.3.7) \]

where \( X \) is an \((N \times Q)\) matrix of realizations from \( Q \) explanatory variables, and \( B = [\beta_j(t)] \) is a \((Q \times T)\) matrix of linear model parameters. Equivalently, one may write

\[ E(Y_i(t)) = \sum_{j=1}^{Q} x_{ij} \beta_j(t), \quad (2.3.8) \]

with the summation extending from \( j=1 \) to \( Q \). An intercept term can easily be accommodated by the GLMM by defining the first (say) column of \( X \) to be a vector of 1's. Thus,

\[ E(Y_i(t)) = \lambda(t) + \sum_{j=1}^{Q} x_{ij} \beta_j(t). \quad (2.3.9) \]

Harris[(1975), Sec. 4.9] discussed the difference between viewing repeated measures analysis in a "multivariate" framework, versus a univariate, split-plot design. Necessary and sufficient conditions for proceeding with the univariate approach were provided by Huynh & Feldt(1970). One such manifestation is the assumption of "compound symmetry" in the covariance matrix \( \Sigma \). That is, all the main diagonal elements must be equal to one value, and the off-diagonal elements must all be equal to a second value. While this is a Toeplitz matrix, it does not correspond to any of the ARMA patterns.

A particular genre of repeated measures design is the "growth curve analysis" of Grizzle & Allen(1969) among others. This strategy is particularly suited to those situations where a systematic change is expected over the repeated measures, e.g., linear, quadratic, sinusoidal, and so on. The design matrix \( X \) is analogous to the "across individuals" model of growth curve analysis, although there is no equivalent to the "within individuals" design in the longitudinal model.
2.3.2 Related Econometric Models

Detailed reviews of models indigenous to the econometrics literature have recently been published by Dielman (1983) and Hendry & Richard (1983), and a brief summary of this research is presented next. This survey will necessarily be short, since it will be seen that they observe a different focus in their outlook. Nonetheless, these differences can be used to gain a greater perspective on longitudinal data analysis.

In contrast to (2.3.9), the model postulated at the most basic level in the econometrics literature [da Silva (1975), p. 8] is defined as

$$E(Y_i(t)) = \alpha_i + \lambda^t + \Sigma x_{ij} \beta_j$$  \hspace{1cm} (2.3.10)

Since the explanatory variables are indexed by "t", they are allowed to vary over time; the regression coefficients, however, are held to be time-invariant. This model therefore strives to demonstrate underlying relationships between the response measure and the exogenous variables irrespective of the time points at which the observations happened to have been gathered. Moreover, both models postulate different intercept terms across the various time points; the econometrics model also allows for a different intercept term for each observational unit.

The econometrics literature cleaves according to the assumptions made concerning the \{\alpha_i\} and \{\lambda^t\} terms. In a fixed-effects model, it is assumed that \Sigma \alpha_i = 0 = \Sigma \lambda^t. Wallace & Hussain (1969) and Swamy & Arora (1972) described an algorithm for isolating the estimates for the intercept terms. However, interest is usually restricted to the regression parameters \{\beta_j\}, and a transformation matrix was described which effectively centers the response measure as

$$y_i^*(t) = y_i(t) - \bar{y}_i(*) - \bar{y}_(t) + \bar{y}_(*)$$  \hspace{1cm} (2.3.11)

where the dot indicates that the mean has been taken over that index. If each explanatory variable is centered in a similar manner, then OLS regression on the transformed ob-
servations results in the best linear unbiased estimator (BLUE) for $\mathbf{g}$. Kiefer(1980) exposed several pitfalls in (2.3.10) all of which can be attributed to the nuisance parameters $\{\alpha_i\}$. He proposed an estimator for each $\alpha_i$ which is a weighted (rather than a simple) average of the elements of $u_i$, and conditional upon these estimators, an estimator for $\mathbf{g}$ was described.

Balestra & Nerlove(1966) argued that the $\{\alpha_i\}$ are really of secondary interest, squandering degrees of freedom, and generally impeding the estimation of the parameters of primary interest, viz., $\mathbf{g}$. An alternative is to consider the intercept terms, together with the error terms $\{u_i(t)\}$, as mutually uncorrelated random variables with means 0 and variances $\sigma_2^2$, $\sigma_3^2$ and $\sigma_4^2$ respectively. Wallace & Hussain(1969) derived consistent estimators for these variance terms based upon the observed residuals arising from OLS regression applied to (2.3.10), and these in turn were used to specify an estimator for the regression parameter vector.

Hussain(1969) proposed a model which is a hybrid of the fixed effects and error components models. He suggested, in contrast to Kiefer(1980), that some econometric studies are characterized by data available from a small number of subjects over a long period of time. In these situations, it is reasonable to assume that the cross-sectional coefficients $\{\alpha_i\}$ are fixed effects parameters, while the time components $\{\lambda(t)\}$ are random effects parameters. Under these assumptions, an estimator of the regression parameter $\mathbf{g}$ was derived which is unbiased, consistent, and asymptotically normal, and asymptotically more efficient than the OLS estimator.

da Silva(1975) suggested that these models owe their origins to the "Seemingly Unrelated Regressions (SUR)" model pioneered by Zellner(1962). Under this model, the response vector for the $i$-th individual is modelled as

$$\mathbf{y}_i = \mathbf{X}_i \mathbf{g}_i + \mathbf{u}_i,$$  \hspace{1cm} (2.3.12)
and one may write (2.3.4) as

\[ \mathbf{v} = \mathbf{Z}\mathbf{b} + \mathbf{u}, \]  

(2.3.13)

where \( \mathbf{Z} \) is now a block diagonal matrix, whose i-th diagonal block is \( \mathbf{X}_i \), and \( \mathbf{b}' = [\mathbf{b}_1' \ldots \mathbf{b}_N'] \). It is assumed that \( \mathbf{E}(\mathbf{u}_i) = \mathbf{0} \), but that

\[ \text{Cov} (\mathbf{u}_i, \mathbf{u}_j) = \sigma_{ij} \mathbf{I}_T, \quad (i, j=1, 2, \ldots, N) \]  

(2.3.14)

Thus, the observations from any two subjects (at the same time point) are assumed to be intercorrelated, but there is no covariance structure among the repeated measures within an individual. The covariance structure is given by

\[ \mathbf{Q} = \mathbf{I} \otimes \mathbf{I}, \]  

(2.3.15)

where \( \mathbf{I} = [\sigma_{ij}] \) is the (N by N) covariance matrix among the observational units.

If \( \mathbf{I} \) were known, then the Aitken estimator would be the BLUE of \( \mathbf{b} \) and a MLE under multivariate normality. Generally, \( \mathbf{I} \) is unknown, but one may define \( \mathbf{S} \) to be the (N by N) matrix with typical element

\[ s_{ij} = (\mathbf{v}_i' - \mathbf{X}_i \mathbf{b})' (\mathbf{v}_j' - \mathbf{X}_j \mathbf{b}) / (T - Q) \]  

(2.3.16)

for \( i, j=1, 2, \ldots, N \), where \( \mathbf{b} \) is the OLS regression estimator from (2.3.13). Zellner showed that if \( \tilde{\mathbf{Q}} = \mathbf{S} \otimes \mathbf{I} \) is substituted for \( \mathbf{Q} \) in the Aitken Estimator, then

\[ \tilde{\mathbf{b}} = (\mathbf{Z}' \tilde{\mathbf{Q}}^{-1} \mathbf{Z})^{-1} \mathbf{Z}' \tilde{\mathbf{Q}}^{-1} \mathbf{v}, \]  

(2.3.17)

has the same asymptotic properties as the Aitken estimator itself. Kakwani(1967) further showed that this estimator is unbiased for \( \mathbf{b} \) provided that its expectation exists, and the distribution of the disturbance vector \( \mathbf{u} \) is continuous and symmetric.

This model is counter-intuitive from a biostatistical viewpoint since it allows for contemporaneous correlation among the individuals, but not for serial correlation over the repeated measures. A simple interchange of indices would bring it into line with the usual GLMM approach.
Parks(1967) presented an extension to the original SUR which allows for both contemporaneous and serial correlation. Consider the AR(1) process among the disturbance terms,

\[ u_i(t) = \phi_i u_i(t-1) + \epsilon_i(t), \quad (i=1,2,\ldots,N), \quad (2.3.18) \]

where \(|\phi_i|<1\) for all \(i\). It is assumed that \(E(\epsilon_i(t))=0\) for all \(i\) and \(t\), and that

\[
\text{Cov}(\epsilon_i(t), \epsilon_j(s)) = \begin{cases} 
\sigma_{ij} & \text{for } t=s \\
0 & \text{otherwise.} \end{cases} \quad (2.3.19)
\]

Under this model, it can be shown [Judge et al.(1980), p. 328] that the covariance structure may be expressed as \(E_{ij} = \sigma_{ij} \Omega_{ij}\), where

\[
\Omega_{ij} = (1-\phi_i \phi_j)^{-1} \left[ \begin{array}{cccc}
1 & \phi_j & \phi_j^2 & \ldots & \phi_j^T-1 \\
\phi_i & 1 & \phi_j & \ldots & \phi_j^{T-2} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\phi_i^{T-1} & \phi_i^{T-2} & \phi_i^{T-3} & \ldots & 1 \\
\end{array} \right] \quad (2.3.20)
\]

is a Toeplitz matrix of the AR(1) parameters \(\phi_i\) and \(\phi_j\). The covariance matrix \(\Omega\) of the joint set of observations is that partitioned matrix with \(\sigma_{ij} \Omega_{ij}\) in the \((i,j)\)-th position.

When \(\Omega\) is unknown, Parks(1967) described a three-stage estimation procedure which results in a consistent and asymptotically efficient estimator for \(\phi\). In the first step, OLS regression procedures are applied to (2.3.13) resulting in a set of residuals \(\hat{u} = y - \mathbf{Zb}\). Each AR(1) parameter can be consistently estimated by

\[
\hat{\phi}_i = \frac{\sum \hat{u}_i(t)\hat{u}_i(t+1)}{\sum [\hat{u}_i(t)]^2} \quad (2.3.21)
\]

for \(i=1,2,\ldots,N\). In the second stage, the Cochrane-Orcutt transformation previously described for the AR(1) process is applied for each individual, and the transformed observations are then used to compute estimates of the contemporaneous covariances \(\sigma_{ij}\). Finally, in the third stage, the es-
timates of the AR(1) parameters and the contemporaneous covariance elements are substituted in an obvious way into each block of the partitioned matrix for \( \varphi \) to derive the Aitken estimator for \( \gamma \).

Kmenta[(1971), Sec. 12-2] considered a particular application of Parks' model. The assumption of contemporaneous correlation among the observational units was dropped, although the model does provide for heteroscedasticity across individuals. The AR(1) error process is retained among the repeated measures within individuals. Thus, the model may be expressed as

\[
\eta_i(t) = \varphi_i \eta_i(t-1) + \epsilon_i(t),
\]

(2.3.22)

where \( E(\epsilon_i(t)) = 0 \), and

\[
\text{Cov}(\epsilon_i(t), \epsilon_j(s)) = \begin{cases} 
\sigma & \text{for } (i,t) = (j,s) \\
0 & \text{otherwise.}
\end{cases}
\]

(2.3.23)

Estimation proceeds in a manner analogous to the three-stage procedure outlined for the Parks' model. If the covariance structure is further assumed to be homogeneous across observational units, then a consistent estimator of \( \varphi \) [Kmenta(1971), p. 512] is calculated as

\[
\hat{\varphi} = \sum \frac{\hat{u}_i(t)\hat{u}_i(t+1)}{\sum \hat{u}_i(t)^2},
\]

(2.3.24)

which is seen to be a weighted average of the individual-specific AR(1) parameter estimates.

Kmenta & Gilbert(1970) provided Monte Carlo evidence concerning the small sample properties for Parks' basic model. In general, all multi-stage procedures were found to be superior to those strategies ignoring one or both types of correlation structure. In light of the Monte Carlo evidence presented previously, it is perhaps not surprising to learn that the best performance was observed from a joint non-linear optimization procedure, of the type presented in the previous section.
Finally, the model proposed in da Silva's dissertation [da Silva(1975), Ch. 3] is a combination of the error components models, and the inter-correlation models presented here. In particular, he assumed that the time-dependent intercepts $\{\lambda(t)\}$ are random variables, but in contrast to the AR(1) error process, the general MA process was used to model the covariance structure among repeated measures within observational units. The resulting estimator was shown to be asymptotically efficient.

2.4 MISSING DATA

Missing data are a common occurrence in applied statistical research. Beyond the computational complexity introduced by incomplete data, there are the broader and more important statistical issues. Missing data cast doubt on the representativeness of the sample observed and the external validity of the inferences drawn, leaving conclusions vulnerable to charges of bias and statistical inefficiency. Nevertheless, while considerable efforts are made to capture all the data, it frequently happens that the data are simply not available or have not been gathered in the proper format.

Notably silent on the issue of incomplete data are many standard econometrics references, including Judge et al.(1980), Malinvaud(1980), and Theil(1971). A recent survey article [Newbold(1981)], moreover, did not consider missing value problems. Kmenta(1971), Sec. 9-3] does provide some assistance, but only for the simple linear regression model with i.i.d. error terms, and then only in the case where some of the values of the independent variable are missing.

The effects of missing values are perhaps more devastating in the case of the linear model with an ARMA error structure. Even in the complete data case, there is no closed-form solution to the normal equations, the likelihood function is tedious to compute, and expressions for first- and second-order derivatives are, for the most part, mathe-
matically intractable. All these problems are exacerbated when some elements of the response variable are missing. The regular structure in the covariance matrix $G$, its inverse, and determinant are no longer available to ameliorate the situation. It is therefore not surprising that there are few papers in the literature describing an attack on the problem.

2.4.1 The EM Algorithm

In recent years, a general methodology has been forthcoming for coping with the presence of missing values. Dempster, Laird & Rubin (1977) described the "EM Algorithm" for estimating parameters in models with missing data. The algorithm is an iterative procedure, with each iteration consisting of two steps. In the estimation (E) step, the conditional expectations of the (theoretical) complete-data sufficient statistics are computed, conditional upon the observed data and the current estimates of the parameter vector. The maximization (M) step then updates the maximum likelihood estimator of the parameter vector based upon the completed sufficient statistics. Dempster, Laird & Rubin showed that provided the underlying density function is a member of the exponential family of distributions, the iterations will converge to at least a local maximum of the parameters in the likelihood space. The EM algorithm was also shown to be a generalization of the "Missing Information Principle" proposed in Orchard & Woodbury (1972).

The "Achille's heel" of the EM algorithm is the M-step. It is critical that, given estimates of the sufficient statistics of the complete data vector, a procedure be readily available to estimate the parameters of the model. Indeed, as Rubin & Sztatowski (1982) pointed out, the situations most appropriate for attack by the EM algorithm are those for which the "... complete data problem has an explicit maximum likelihood estimate because then the M-step of the EM algorithm has an explicit form ..." (p. 657-8). However in the
linear model with an ARMA error structure, there is no closed-form expression for the MLE of the time series parameter vector \( \hat{\theta} \). There are expressions for \( \hat{\theta} \) and \( \hat{\sigma}^2 \), but they are both implicit functions of \( \hat{\theta} \) through the matrix \( G(\theta) \). Rubin & Szatrowski (1982) further showed that the EM algorithm is applicable when the covariance matrix has a linear pattern, i.e., \( \sum G G G \) where the \( G \)'s are linearly independent and symmetric matrices, and the \( \sigma \)'s are scalars. It is known, however, that ARMA covariance matrices do not conform to this particular structure. Consequently, there is no way under this model to capitalize on the power of the EM algorithm in the presence of missing values.

2.4.2 Incomplete Longitudinal Designs

Against the backdrop of the general linear multivariate model with an ARMA covariance structure, neither the statistical nor the econometrics literature provides a great deal of information. When some of these assumptions are relaxed, some studies are available for guidance. This is particularly true under a general covariance structure \( \Sigma \) with no time series representation. This research is reported first, followed by more particular research into this model.

Anderson (1957) considered the case of random sampling from the multivariate normal distribution. He derived expressions for the MLE's of the parameters specifically when the data are missing in a hierarchical manner, that is, those individuals from whom data are missing on one variable form a subset of those whose data are missing on another.

A random missing value pattern was considered by Hocking & Smith (1968) and Hartley & Hocking (1971) for the multivariate normal distribution. The sample was partitioned into subgroups according to the missing data structure. The overall MLE's of the parameters were shown to be functions of the subgroup-specific MLE's, and an iterative procedure was proposed to solve these equations simultaneously. The
resulting estimators were demonstrated to be maximum likelihood.

Kleinbaum (1970) considered the missing data problem without relying upon the likelihood function of the sample. The "More General Linear Model (MGLM)" was introduced which not only allows subgroups with different missing data patterns across the repeated measures, but also allows for different design matrices for the different columns of $\mathbf{y}$, analogous to the Seemingly Unrelated Regressions approach. Unbiased and consistent estimators of $\Sigma$ were identified, and these, in turn, were used to provide best asymptotic normal (BAN) estimators for the linear model parameters. Hypothesis testing was also considered under this structure.

Several procedures were examined in a Monte Carlo study by Hosking (1980). These included Hocking & Smith (1968) / Hartley & Hocking (1971), Orchard & Woodbury (1972), and Kleinbaum (1970), as well as standard listwise deletion. The study systematically varied the sample size, the proportion of missing values, and the average inter-correlation among the response measures. On balance, both Kleinbaum's method and Orchard & Woodbury's approach were judged to be superior. The former was preferred for estimating $\Sigma$, while the latter was preferred for $\mathbf{b}$. The Missing Information Principle was also endorsed in a small simulation study by Beale & Little (1975).

Several studies have attempted to incorporate particular missing data patterns into a useful analysis strategy. For example, Trawinski & Bargman (1964) partitioned the sample into $k$ groups, each of size $n$, according to the particular subset of repeated measures to be provided. The joint likelihood function was observed to be the product of the subgroup-specific likelihoods, and an iterative procedure was identified for finding parameter values which maximize this likelihood using a non-linear optimization routine. Rao & Rao (1966) considered the problem of establishing anthropome-
tric standards among Indian boys between the ages of 5 and 16, and a study design was presented, called a "linked cross-sectional" design. Rather than following each individual over the course of the entire 11-year period, it was suggested that they be measured on two or three consecutive birthdays, with different starting dates for various individuals. The optimal number of repeated measures per individual, and the relative precision of the resulting estimators were shown to be functions of the inter-correlations among the set of repeated measures. Patel & Khatri (1981) considered the case where all subjects enrolled in a clinical trial provide baseline (i.e., pre-treatment) values and at least one follow-up report. Under a peculiar covariance structure, they postulated an analysis of covariance model, and considered issues in hypothesis testing for treatment effects.

The econometric model where both the \( \alpha_j \) and \( \lambda^{(t)} \) were considered as random effects parameters, was studied by Biorn (1981) for the missing data case. A specific rotating design was detailed, that is, where at each time, a fixed number of subjects is withdrawn and replaced by an equal number of new population elements. An iterative estimation procedure was discussed which is similar to the complete-data method for this error components model.

The growth curve model was generalized to accommodate incomplete data by Kleinbaum (1973). In particular, pairwise-deletion was proposed to estimate each element of the covariance matrix \( \Sigma \), although this may be refined through an iterative procedure. This estimator was shown to be consistent, and was used to produce a BAN estimator for \( \tilde{g} \). Woolson et al (1978) pointed out that these expressions form a more general representation of expressions laid out in Rao & Rao (1966). They applied the linked cross-sectional approach in the MGLM framework to derive growth norms for children between 6 and 18 years of age. Overall, this approach generated a considerable increase in precision relative to a simple cross-sectional study.
Kleinbaum(1973) recognized that the pairwise-deletion estimator of $\mathbf{I}$ was not guaranteed to be positive definite, and this may seriously degrade the underlying distribution of any test statistic. A Monte Carlo investigation was undertaken by Leeper & Woolson(1982) to study the small sample properties of this estimator, and to compare it against the listwise-deletion estimator and a "smoothed" estimator. The latter is that positive semi-definite matrix which is "closest" (in the sense of Euclidean norm) to the pairwise estimator. On balance, while smoothing eliminated negative test statistics, it also profoundly affected the distribution of the Wald statistic, with significance levels being less than the nominal level (i.e., conservative) in some cases.

LaVange(1983) adopted a different approach for the joint estimation of $\mathbf{a}$ and $\mathbf{I}$. First, an estimate of $\mathbf{I}$ was generated from the odd-numbered rows of the data matrix, and this was shown to be consistent. Then using this estimator, and the even-numbered rows, a BAN estimator for $\mathbf{a}$ was derived. Moreover, the GLMM was discussed under two time series error models. For the AR(1) case, a procedure was described which amounts to the incomplete-data analogy of Kmenta(1971)'s three-stage algorithm. As well, da Silva(1975)'s approach for the MA(q) process was extended to accommodate missing data.

Jones(1984a) outlined a one-way Analysis of Variance model for a univariate repeated measures design, applicable when observations are missing or unequally-spaced. The model included a random-effects parameter to account for variation among the observational units, and fixed-effects parameters for treatment differences. It is not clear whether the results extend to the general linear model. An AR(1) error structure was used to model the covariance structure among the repeated measures; while Jones(1984b) considered the general time series case. An application of the Kalman filter was described to evaluate the likelihood function,
and ultimately to iterate to optimal estimates of the parameters.
3.1 THE MODEL

3.1.1 Underlying Structure

Let \( Y_{i}^{(t)} \) be a random variable which denotes the outcome or response measure from a particular process or experiment under observation taken from the \( i \)-th observational unit at the \( t \)-th time point \((i=1,2,\ldots,N, \ t=1,2,\ldots,T)\). It is assumed that the repeated measures have been taken at equally-spaced intervals of time, namely \( t_0+\delta, \ t_0+2\delta, \ldots, \ t_0+T\delta \) for some time origin \( t_0 \) and time increment \( \delta \), for each observational unit \( i=1,2,\ldots,N \). These observations may be gathered into a two-dimensional array,

\[
Y^* = \begin{bmatrix}
Y_1^{(1)} & \ldots & Y_1^{(T)} \\
\vdots & & \vdots \\
Y_N^{(1)} & \ldots & Y_N^{(T)}
\end{bmatrix}
= \begin{bmatrix}
Y_1^{*'} \\
\vdots \\
Y_N^{*'}
\end{bmatrix}
\tag{3.1.1}
\]

where rows refer to different observational units and columns refer to different time points, and \( Y_1^{*'} = [Y_1^{(1)} \ Y_1^{(2)} \ Y_1^{(T)}] \) represents the set of repeated measures from the \( i \)-th individual (i.e., the \( i \)-th row of \( Y^* \)).\(^1\)

In order to interpret the behaviour of these random variables, the broad theory of the general linear model [e.g., Timm(1975)] is introduced. Consider a model of the form:

\[
y^* = X \beta + u^*, \tag{3.1.2}
\]

where \( X \) is an \((N \times Q)\) matrix of non-stochastic explanatory variables, with full column rank, whose values are thought to influence or account for variation in the \( Y_i^{(t)} \). \( \beta \) is a

\(^1\) The superscript "*" is used to distinguish the underlying complete-data model from the missing-data case introduced in the next section.
(Q by T) matrix of fixed-effect linear model parameters, whose values measure the contribution of the explanatory variables and for whose values estimators are sought; and \( U^* \) is an (N by T) matrix of unobserved error terms, with typical element \( u_i(t) \), representing variation in the \( Y_i(t) \) unaccounted for by the model. Each \( \{Y_i(t)\}_t \) represents a "time series" of observations; however, with systematic variation accounted for by the deterministic component \( XB \), it is assumed that the elements of \( U^* \) observe the ARMA time series model. Hence,

\[
\phi(B)u_i(t) = \theta(B)\epsilon_i(t),
\]

where \( \phi(B) \) and \( \theta(B) \) retain their definitions from Chapter II, and are subject to the usual stationarity and invertibility restraints. The \( \{\epsilon_i(t)\} \) is the "white noise" error process, with properties,

\[
E(\epsilon_i(t)) = 0, \quad \text{... for all } (i,t), \quad (3.1.4)
\]

and,

\[
\text{Cov}(\epsilon_i(t),\epsilon_j(s)) = \left\{ \begin{array}{ll} \sigma^2 & \text{... } (i,t)=(j,s) \\ 0 & \text{... otherwise, } \end{array} \right. \quad (3.1.5)
\]

with \( 0<\sigma^2<\infty \). Under these conditions, it follows that

\[
E(U^*) = 0 \quad (3.1.6)
\]

and if \( u_i^* \) represents the i-th row of \( U^* \) (i.e., errors associated with the i-th individual), then

\[
\text{Var}(u_i^*) = \sigma^2 G, \quad (3.1.7)
\]

where \( G \) is a (T by T) Toeplitz matrix function, and each element of \( G \) is a function exclusively of the time series parameters \( \psi' = [ \phi' \theta' ] \) as discussed in Chapter II.

It is specifically assumed that the set of repeated measures arising from different observational units are mutually uncorrelated, so that

\[
\text{Cov}(u_i^*,u_j^*) = 0 \quad \text{... for } i\neq j. \quad (3.1.8)
\]

- 40 -
Additionally, it is useful to impose the assumption of normality upon the vectors \( u_i^* \), so that each \( u_i^* \) follows the \( \text{MVN}(0, \sigma^2 G) \). A succinct way, therefore, of summarizing the underlying model is to say that \( y^* \) satisfies the conditions of the GLMM-FR\( (y^*, X^B, \sigma^2 G) \) with normality.

3.1.2 Missing Values

Suppose that at least one of the elements of \( y^* \) is missing. Let \( A \) be an \( (N \times T) \) matrix of random indicator variables, with typical element,

\[
a_{i}(t) = \begin{cases} 
1 & \text{if } Y_i^*(t) \text{ is recorded} \\
0 & \text{if } Y_i^*(t) \text{ is missing.} 
\end{cases}
\]  

(3.1.9)

If the \( \{a_{i}(t)\} \) are mutually independent random variables, both within individuals and across individuals, then this implies that observations are missing completely at random. Some correlation structure would suggest a pattern to the missing data elements, either imposed by the experimenter or by the subjects. It is specifically assumed that these elements are stochastically independent of the elements of \( y^* \), and that the parameters of the distribution of \( A \) are functionally independent of the parameters of the distribution of \( y^* \). The conditional distribution of \( y^* \) given \( A \) therefore coincides with the marginal distribution of the observed data elements, and it is upon this distribution that inference will be drawn.

For the \( i \)-th individual, suppose that only \( T_i \) of the repeated measures \( (0 < T_i \leq T) \) are actually recorded, corresponding to the time points \( t_0 + t_{i1} \delta, t_0 + t_{i2} \delta, \ldots, t_0 + t_{iT_i} \delta \), where \( 1 \leq t_{i1} < t_{i2} < \ldots < t_{iT_i} \leq T \). Let \( Y_{i}' = [Y_{i}^*(t_{i1}) \ Y_{i}^*(t_{i2}) \ldots \ Y_{i}^*(t_{iT_i})] \) represent the set of repeated measures for the \( i \)-th individual actually captured \( (i=1,2,\ldots,N) \). Equivalently, there exists an indicator matrix, \( K_i \), such that

\[
y_i = K_i y_i^*,
\]  

(3.1.10)
where $K_i$ is a $(T_i$ by $T$) matrix, whose rows are a subset of the rows of the $T$-dimensional identity matrix, and where each row captures the appropriate repeated measure. For example, if $T=4$, but $Y_i^{(2)}$ is unavailable, then

$$
Y_i = \begin{bmatrix}
Y_i^{(1)} \\
Y_i^{(2)} \\
Y_i^{(3)} \\
Y_i^{(4)} \\
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 \\
\end{bmatrix} Y_i^{*}
$$

Similarly, one may define $u_i = K_i u_i^*$ for the set of error terms corresponding to the observed data elements for the $i$-th observational unit.

Let $B_j' = [\beta_j^{(1)} \beta_j^{(2)} \ldots \beta_j^{(T)}]$ represent the $j$-th row of the parameter matrix, and define $\beta$ to be that $(OT$ by $1)$ parameter vector, where the successive rows of $B$ are first transposed and then stacked underneath each other ("rolled out by rows"), i.e.,

$$
\beta' = [B_1' \ B_2' \ldots \ B_Q'].
$$

Let $Y$ be that vector formed by stacking the set of observed measures from each observational unit underneath each other, i.e.,

$$
Y' = [Y_1' \ Y_2' \ldots \ Y_N'],
$$

$u$ may be similarly defined. If $x_i'$ represents the $i$-th row of the design matrix $X$, then it can be shown that

$$
E(Y_i) = (x_i' \otimes K_i) \beta = Z_i \beta
$$

and,

$$
\text{Var}(Y_i) = \sigma^2 (K_i G K_i') = \sigma^2 G_i.
$$

for $i=1,2,\ldots,N$. It therefore follows that the underlying model may be represented as

---

Note that the matrix $K_i$ is the transpose of the matrix $K_j$ defined in LaVange(1983), p. 21).
\[ \mathbf{y} = \mathbf{Z}_d + \mathbf{u} \]  
(3.1.15)

where

\[
\mathbf{Z} = \begin{bmatrix}
\mathbf{Z}_1' & \mathbf{K}_1 \\
\vdots & \\
\mathbf{Z}_N' & \mathbf{K}_N
\end{bmatrix} = \begin{bmatrix}
\mathbf{Z}_1 \\
\vdots \\
\mathbf{Z}_N
\end{bmatrix},
\]  
(3.1.16)

and,

\[
\mathbf{Q} = \text{Var}(\mathbf{y}) = \sigma^2 \begin{bmatrix}
\mathbf{K}_1 \mathbf{G}_1' & & \\
& \ddots & \\
& & \mathbf{K}_N \mathbf{G}_N'
\end{bmatrix} = \sigma^2 \mathbf{W}
\]  
(3.1.17)

i.e., a block diagonal matrix, with blocks \( \sigma^2 \mathbf{K}_i \mathbf{G}_i' \). Thus, \( \mathbf{y} \) satisfies the conditions of the GLUM-FR(\( \mathbf{y}, \mathbf{Z}_d, \sigma^2 \mathbf{W} \)) with normality. This configuration is referred to by LaVange (1983) as the "linear model for individual observation times (LMI)."

Summarizing, the general linear model of the form

\[ \mathbf{y} = \mathbf{Z}_d + \mathbf{u} \]

is under consideration, subject to the following assumptions.

**Assumption 3.1**
Each \( \mathbf{u}_i \) follows a multivariate normal distribution.

**Assumption 3.2**
The elements of each \( \mathbf{u}_i \) observe a particular ARMA(p,q) time series representation, where \( p \) and \( q \) are fixed and known, i.e.,

\[ \phi(B) \mathbf{u}_i(t) = \theta(B) \mathbf{e}_i(t). \]

**Assumption 3.3**
\[ \mathbf{E}(\mathbf{e}_i(t)) = 0, \quad \text{... for all } (i,t), \]

\[ \text{Cov}(\mathbf{e}_i(t), \mathbf{e}_j(s)) = \begin{cases} \sigma^2 & (i,t) = (j,s) \\ 0 & \text{otherwise}, \end{cases} \]

with \( 0 < \sigma^2 < \infty \). As an immediate consequence of this assumption, it follows that \( \mathbf{E}(\mathbf{u}_i) = 0 \), and \( \text{Var}(\mathbf{u}_i) = \sigma^2 \mathbf{K}_i \mathbf{G}_i' \), for all
i=1,2,...,N, where $G$ is a ($T$ by $T$) Toeplitz matrix function of a finite and constant number of parameters, $\phi' = [\phi' \phi']$.

**Assumption 3.4**
The ARMA process is stationary, i.e., the roots of $\phi(B)=0$ lie outside the unit circle in the complex plane; the ARMA process is invertible, i.e., the roots of $\theta(B)=0$ lie outside the unit circle in the complex plane; and, $\phi(B)$ and $\theta(B)$ have no roots in common.

**Assumption 3.5**
The set of disturbance terms arising from different observational units are mutually uncorrelated,

$$\text{Cov}(u_i, u_j) = 0, \quad \text{for } i \neq j.$$  

It therefore follows that the covariance matrix of $u$ is a block diagonal structure whose $i$-th diagonal block is $\sigma^2 K_i G K_i' = \sigma^2 G_i$.

**Assumption 3.6**
The linear model parameters $\theta$ are functionally independent of the covariance parameters $\phi$ and $\sigma^2$.

As discussed in the previous chapter, this model has been considered in a number of different forms in a variety of settings over the past twenty years. Kleinbaum(1970) described the "More General Linear Model (MGLM)", and under this formulation, there are also $T$ response measures, although not necessarily related by time. The model specifically being proposed here corresponds to Kleinbaum's "General Incomplete Model (GIM)".

Both Kleinbaum(1970) and LaVange(1933) addressed estimation and hypothesis testing issues under this model. For the most part, the general covariance structure is considered, although LaVange specifically proposed an estimator under an AR(1) and a MA(q) error process. In both studies, best asymptotic normal (BAN) estimators were proposed for the linear model parameter matrix $B$, and their asymptotic properties were discussed.
There are therefore two points of departure for the present study. This effort will be concerned with the general ARMA(p,q) error process, of which the AR(1) and MA(q) error processes are special cases. Moreover, a different estimation criterion is being proposed for this parameter set, namely, maximum likelihood (ML) estimation. The weight of evidence forthcoming from the univariate complete-data regression model highly recommends ML estimation, and is upon the strength of these studies that this technique is pursued in this setting.

3.1.3 Miscellaneous Properties

In theory there are $2^T-1$ missing-value patterns possible among the repeated measures, although in practice far fewer are generally observed. Moreover, several observational units may share the same missing-value pattern. Since one may permute the rows of $\mathbf{Y}^*$, $\mathbf{X}$, and $\mathbf{U}^*$ without disturbing inferences on the parameters, it is assumed, without loss of generality, that each set of observational units revealing the same missing-value pattern is arranged in a contiguous block of rows in the data array $\mathbf{Y}^*$.

Suppose that $A$ distinct missing-value patterns are observed among the repeated measures, indexed by the subscript "a", and let $G_a = K_a G K_a$ represent the covariance structure associated with the $a$-th missing-value pattern. If this pattern is observed by $n_a$ observational units ($a=1,2,\ldots,A$) such that $\sum n_a = N$, then the covariance matrix may equivalently be represented by the block diagonal structure

$$
\Omega = \sigma^2 \begin{bmatrix}
I_{n_1} \otimes G_1 \\
I_{n_2} \otimes G_2 \\
\vdots \\
I_{n_A} \otimes G_A
\end{bmatrix}
$$

(3.1.18)

---

3 The matrix $G$ will be subscripted by both "i" for individuals and "a" for missing-value patterns. The meaning, however, should be clear from the context.
Next, since $G$ is a Toeplitz matrix, there are at most $T$ distinct elements in $G$, and these are denoted by $g_0, g_1, \ldots, g_{T-1}$, where for the $(r,s)$-th element of $G$, $g_{r,s} = g_{r-s}$. Consider the partial derivatives of each $g_j$ with respect to any ARMA parameter. For example, for an AR(1) process, it can be shown that $g_j = \phi^j/(1-\phi^2)$ for $j=0,1,2,\ldots$. Thus,

$$\partial g_j / \partial \phi = j \phi^{j-1} / (1-\phi^2) + 2\phi^{j+1} / (1-\phi^2)^2 \quad (3.1.19)$$

for $j=0,1,2,\ldots,T-1$. The stationarity restraint requires that $|\phi|<1$, so that all these quantities will be well defined.

Similarly, for the general MA model of order $q$, it can be shown that $g_j = \sum_{k=0}^{q} \theta_k \epsilon_{k+j}$, with the summation being taken from $k=0$ to $q$, and where

$$\theta_k = \begin{cases} 
0 & \text{... for } k<0 \\
-1 & \text{... for } k=0 \\
0 & \text{... for } k>q.
\end{cases}$$

It therefore follows that

$$\partial g_j / \partial \theta_k = \theta_{k-j} + \theta_{k+j} \quad (3.1.20)$$

exists for all $j$ and $k$.

For the general ARMA model, the required partial derivatives can be found using a two-step procedure. First, Box & Jenkins (Sec. 3.4) defined the "cross-covariance" function between $\{u_i(t)\}$ and $\{\epsilon_i(t)\}$ as

$$\eta_j = \text{Cov}(u_i(t+j), \epsilon_i(t)). \quad (3.1.21)$$

Since each $u_i(t+j)$ can only be dependent upon those shocks which have occurred up to (and including) time $t+j$, it follows that $\eta_j=0$ for $j<0$. Straightforward algebraic manipulation [see, for example, McGilchrist et al.(1981), p. 2576] results in the following relationships among the cross-covariance terms

$$\sum_{i=1}^{\infty} \eta_{k-i} = \theta_k \quad \text{... } k=1,2,\ldots,c, \quad (3.1.22)$$

- 46 -
the summation being taken for i=0 to p. A general solution can be found to this system of equations, so that the \( \eta \)'s can be expressed in terms of the \( \phi \)'s and \( \theta \)'s. Simple differentiation produces their corresponding partial derivatives, \( \partial \eta_j / \partial \phi_k \).

Next, Box & Jenkins (p. 74) demonstrated that

\[
\Xi_i \phi g_{k-i} = \Xi_j \theta \eta_{j-k} \tag{3.1.23}
\]

for k=1,2,...,p. Findley (1981) detailed a Lattice version of Levinson's algorithm for the recursive solution to this system of equations; stationarity restraints ensure that a unique solution exists. Thus, each \( g_j \) can be expressed in terms of the \( \eta \)'s, \( \phi \)'s, and \( \theta \)'s. Simple differentiation produces the partial derivatives \( \partial g_j / \partial \phi_k \) for all \( j=0,1,...,T-1 \) and \( k=1,2,...,p+q \).

Thus, the quantity \( \partial G / \partial \phi_k \) exists unequivocably, so that \( \partial G_a / \partial \phi_k = G_a (\partial G / \partial \phi_k) G_a' \), and \( \partial G_a^{-1} / \partial \phi_k = -G_a^{-1} (\partial G / \partial \phi_k) G_a^{-1} \) all exist as well \( (k=1,2,...,p+q) \). Similar results hold for the mixed second-order partial derivatives. These results may therefore be summarized in the follow lemma.

**Lemma 3.1.1:**

Under Assumptions 3.1 - 3.6, the partial derivatives \( \partial G / \partial \phi_m, \partial G / \partial \phi_m \partial \phi_n \), and \( \partial G^{-1} / \partial \phi_m, \partial G^{-1} / \partial \phi_m \partial \phi_n \) all exist for \( m,n=1,2,...,p+q \).

### 3.2 Maximum Likelihood Equations

Consider the LMI representation (3.1.15 - 3.1.17). For each observational unit, \( E(\mathbf{y}_i) = Z_i \mathbf{\theta} \), and \( \text{Var}(\mathbf{y}_i) = \sigma^2 (K_i G K_i') = \sigma^2 G_i \), where \( G_i = K_i G K_i' \). The set of observed measures from each observational unit is assumed to follow the multivariate normal distribution, so that its density function may be written as:

\[
f(\mathbf{y}_i, \mathbf{\theta}, \phi, \sigma^2) = (2\pi\sigma^2)^{-T_i/2} |G_i|^{-1/2} 
\exp\left\{ -\frac{1}{2} (\mathbf{y}_i - Z_i \mathbf{\theta})' G_i^{-1} (\mathbf{y}_i - Z_i \mathbf{\theta}) / 2\sigma^2 \right\} \quad (3.2.1)
\]
Different observational units are assumed to be mutually uncorrelated, so that the joint density of the set of observations is the product of the individual-specific density functions. The likelihood function of the parameter set $\theta^* = [\theta^* \, \lambda^* \, \sigma^2]$ is proportional to the joint density function, so that taking logarithms and ignoring constant terms, it follows that

$$
\Lambda(\theta) = -\left(\frac{T^*}{2}\right) \ln(\sigma^2) - (1/2) \sum \ln|G_i| \\
- \left(\frac{1}{2\sigma^2}\right) \sum (Y_i - Z_i \theta)' G_i^{-1} (Y_i - Z_i \theta)
$$

where $T^* = \sum T_i$ is the total number of observed data elements. The ML equations are found by taking partial derivatives with respect to each of the parameters, and setting the resultant equations to zero. The main result is established in Theorem 3.2.1, but first the following useful results are presented.

**Lemma 3.2.1:**

If $G_i = K_i G K_i'$ $(i=1, 2, \ldots, N)$, then

$$
(\partial/\partial \psi_k) \left[ \ln|G_i| \right] = \text{tr} \left[ V (\partial G/\partial \psi_k) \right].
$$

where $V = \sum K_i' G_i^{-1} K_i$, and "tr" signifies the trace of the matrix.

**Proof:**

Dwyer[(1967), eq. (8.1)] showed that, for a scalar function $D(\psi_k)$, the chain rule of differentiation could be expressed as:

$$
(\partial D/\partial \psi_k) = \text{tr} \left[ (\partial D/\partial G) (\partial G/\partial \psi_k)' \right],
$$

where "tr" signifies the trace of the matrix. Trawinski & Bargman[(1964), eq. (2.11)] showed that for the general matrix $G$,

$$
(\partial/\partial G) \left[ \ln|G_i| \right] = \sum K_i' G_i^{-1} K_i = V.
$$
Bearing in mind that the post-multiplier matrix $\delta G/\delta \psi_k$ is a symmetric matrix, it therefore follows that

$$(\partial G/\partial \psi_k)[\ln |G_i|] = \text{tr}[V(\partial G/\partial \psi_k)].$$

\[ (3.2.3) \]

Q.E.D.

**Lemma 3.2.2:**

If $G_i = K_i G K_i'$ (i=1,2,...,N) and $P_i$ and $Q_i$ are any conformable matrices, each of which is independent of all $\psi_k$, then

$$(\partial G/\partial \psi_k)[\sum P_i' G_i^{-1} Q_i] = - \sum P_i' G_i^{-1} K_i [\partial G/\partial \psi_k] K_i' G_i^{-1} Q_i,$$

\[ (3.2.4) \]

for $k=1,2,...,p+q$.

**Proof:**

This result is implicit in expression (3.2) of Trawinski & Bargman (1964). Note, however, that the negative sign is incorrectly missing from their result.

Q.E.D.

**Theorem 3.2.1:**

Under the model and conditions specified in section 3.1 the ML estimates observe the following first-order equations:

1. $\delta^2 = \sum (Y_i - \hat{Y}_i)^' G_i^{-1} (Y_i - \hat{Y}_i) / T$
2. $[\sum Z_i^' G_i^{-1} Z_i] = \sum Z_i^' G_i^{-1} Y_i$
3. $\text{tr}[V(\delta G/\delta \psi_k)] = (1/\delta^2) \sum U_i^' G_i^{-1} K_i^' [\partial G/\partial \psi_k] K_i U_i$

where "tr" signifies the trace of the matrix, $V(\psi) = \sum K_i^' G_i^{-1} (Y_i - \hat{Y}_i)^' K_i$, $U_i = Y_i - \hat{Y}_i$, and a "hat" over any of the matrices signifies that it is evaluated at the ML estimates of the parameters.

**Proof:**

Taking the derivative of (3.2.2) w.r.t. $\delta^2$ yields

$$(\partial A/\partial \delta^2) = (-T^*/2\delta^2) + (1/2\delta^2) \sum (Y_i - \hat{Y}_i)^' G_i^{-1} (Y_i - \hat{Y}_i).$$

\[ (3.2.5) \]
Setting this expression to zero implies that
\[ \hat{\sigma}^2 = \mathbb{E}(\hat{y}_i - \hat{Z}_i \hat{\theta})' \hat{G}_i^{-1}(\hat{y}_i - \hat{Z}_i \hat{\theta}) / \mathbb{T}. \]  
(3.2.6)

Similarly, differentiating (3.2.2) w.r.t. \( \hat{\theta} \) yields the expression,
\[ (\partial \hat{y}_i / \partial \hat{\theta}) = (1/\sigma^2) [(\Sigma \hat{Z}_i' \hat{G}_i^{-1} \hat{y}_i - (\Sigma \hat{Z}_i' \hat{G}_i^{-1} \hat{Z}_i) \hat{\theta}]. \]  
(3.2.7)

Setting this expression to zero, it follows that
\[ (\Sigma \hat{Z}_i' \hat{G}_i^{-1} \hat{Z}_i) \hat{\theta} = (\Sigma \hat{Z}_i' \hat{G}_i^{-1} \hat{y}_i \]  
with \( \hat{G} = \hat{G}(\hat{\theta}) \) and \( \hat{\theta} \) is the MLE of \( \theta \).

Now, to find the ML equation corresponding to each time series parameter, it follows from the above two lemmas that
\[ (\partial \hat{y}_i / \partial \hat{\psi}_k) = (1/2\sigma^2) \Sigma \hat{u}_i' \hat{G}_i^{-1} K_i \partial \hat{G}_i / \partial \hat{\psi}_k \]  
\( (1/2) \tr[\Sigma (\partial \hat{G}_i / \partial \hat{\psi}_k)] \].  
(3.2.9)

where \( \hat{u}_i = \hat{y}_i - \hat{Z}_i \hat{\theta} \), and \( \Sigma = (\Sigma \hat{K}_i' \hat{G}_i^{-1} K_i \). Setting this expression to zero, it follows that,
\[ \tr[(\hat{G}_i' \hat{G}_i)^{-1} \hat{G}_i' \hat{G}_i^{-1} \hat{u}_i] = (1/\sigma^2) \Sigma \hat{u}_i' \hat{G}_i^{-1} K_i \partial \hat{G}_i / \partial \hat{\psi}_k \]  
(3.2.10)

Q.E.D.

A unique solution will exist for \( \hat{\theta} \) provided that
\[ \Sigma \hat{Z}_i' \hat{G}_i^{-1} \hat{Z}_i = (X_i X_i' \otimes K_i' \hat{G}_i^{-1} K_i \]  
is a non-singular matrix. Since \( X \) is assumed to have full column rank, and each \( G_i \) is positive-definite, this imposes certain restraints on the missing-value patterns. For example, if one or more of the time points are completely empty, then the corresponding columns (and rows) of each \( K_i' \hat{G}_i^{-1} K_i \) would consist entirely of zeroes, and the corresponding columns of \( B \) would not be estimable from the data. Thus, provided that the column vectors of \( \Sigma \hat{Z}_i' \hat{G}_i^{-1} \hat{Z}_i \) span the QT-dimensional Euclidean space, one may write
\[ \hat{\theta} = (\Sigma \hat{Z}_i' \hat{G}_i^{-1} \hat{Z}_i)^{-1} (\Sigma \hat{Z}_i' \hat{G}_i^{-1} \hat{y}_i). \]  
(3.2.11)
It should be evident from expression (3.2.10) that, in light of the quantities \( u_i = x_i - \bar{z}_i \), all the observed data elements are being used to estimate all the time series parameters simultaneously. Moreover, since each element of the covariance structure \( G \) is a function of \( \vartheta \), it follows that that all the elements of \( G \) are simultaneously being estimated by all the data elements as well. This stands in contrast to the procedure described in LaVange(1983), Ch. 4]. It prescribed partitioning the sample into two (equal) subsets - one to estimate the time series parameters, \( \vartheta \), and the second, together with this value, to derive the BAN estimator of the linear model parameters, \( \varphi \). Not only is there a certain arbitrariness in this method, but the sample is also not being used to full advantage. Therein lies one of the distinct advantages of ML estimation over the two-stage BAN estimation procedures.

Oberhofer & Kmenta(1974) described one technique for iterating to a solution to these equations. The procedure consists of alternating between expressions for \( \hat{\varphi} \) and \( \hat{\sigma}^2 \) for the current value of \( \vartheta \), and then, conditionally upon these values, solving the ML equations for \( \vartheta \). It was shown that under general regularity conditions (their Assumption 6), this iterative procedure always converges to a solution of the first-order maximizing equations. An equivalent procedure, based on "concentrating" the likelihood function, is discussed in detail in the next chapter, and the ensuing small-sample properties are discussed at that point.

3.3 PROPERTIES OF ML ESTIMATORS
Kendall & Stuart(1967), Ch. 18] described large-sample properties of ML estimators under the classical assumption of simple random sampling. Under very general conditions, the ML estimator is shown to be consistent and asymptotically efficient. Moreover, if \( \hat{\varphi} \) is the MLE of \( \varphi \), and

\[
I(\varphi) = - E[\partial^2 \Lambda / \partial \varphi \partial \varphi']
\]

(3.3.1)
is the Fisher Information Matrix, then \( \hat{a} \sim \mathcal{N}(a, I(a))^{-1} \) is asymptotically distributed as a multivariate normal distribution with mean 0 and dispersion matrix \( I(a) \). In our case, while different observational units are assumed to be mutually independent, they are not, in general, identically distributed. Recall, however, that expressions (3.1.15 - 3.1.17) had placed the problem in the GLUM(\( y, Z \hat{p}, \mathcal{N} = \sigma^2 W \)) framework. We therefore appeal to the large-sample properties of the GLUM model to establish the large-sample properties for this particular model.

Magnus (1978) studied in detail ML estimation for the GLUM(\( y, Z \hat{p}, \mathcal{N} \)), when \( \mathcal{N} \) is a function of a finite and constant number of parameters, in our case, \( y \) and \( \sigma^2 \). In Sec. 7, the general case is presented, where \( \mathcal{N} \) may be expressed as

\[
\mathcal{N} = \mathcal{N}(I \otimes \Gamma)Q'.
\]

The GLUM model (3.1.15 - 3.1.17) is seen to be a special case of this general structure by taking \( I = \sigma^2 I_N \), \( \Gamma = G \), and

\[
Q = \begin{bmatrix}
K_1 \\
K_2 \\
\vdots \\
K_N
\end{bmatrix}
\]

Thus, properties specified for the ML estimators in Magnus' paper can be applied to this model provided that the underlying assumptions can verified. The remainder of the section is devoted to verifying those assumptions and applying the results.

3.3.1 Regularity

**Theorem 3.3.1:** [Magnus (1978), p. 291]

Under Assumptions 3.1 - 3.6, and Lemma 3.1.1, the multivariate normal density of \( \underline{y} \) with parameters \( \underline{a}' = [\hat{\beta}' \hat{\gamma} \sigma^2] \) is regular with respect to its first and second derivatives, i.e.,

\[
E[\partial^2/\partial \underline{a}] = 0
\]
and

\[- E[\partial^2 \Lambda / \partial \theta \partial \theta'] = E[(\partial \Lambda / \partial \theta)(\partial \Lambda / \partial \theta)']\]  

(3.3.4)

3.3.2 Asymptotic Distribution

The large-sample properties of the ML estimators are considered next. Since the linear model parameters are indexed by \( t \), to avoid a situation where the number of parameters increases without bound, \( T \) is held fixed while the number of observational units is allowed to become infinitely large. The number of rows of \( B \) is held fixed at \( Q \) (i.e., the number of explanatory variables). Thus, the total number of observed data elements \( T^* \) approaches infinity by virtue of \( N \) approaching infinity.

Now, the following assumptions are made.

Assumption 3.7

Every element of the quantity \( Q = (Z'\Omega^{-1}Z) / T^* \) converges as \( N \rightarrow \infty \) to a finite function of \( \psi \), uniformly for \( \psi \) in any compact set.

This quantity reduces to

\[ Q = E[Z_1'G_1^{-1}Z_1 / T^*\sigma^2] = E[\xi_1\xi_1' \otimes K_1'G_1^{-1}K_1] / T^*\sigma^2. \]

This assumption therefore speaks directly to the manner in which the number of observations becomes large. In particular, it is assumed that the number of observations at each time point becomes infinitely large as \( N \rightarrow \infty \). Moreover, if the design matrix reflects the existence of different subpopulations, then this must be true for every subpopulation in the experimental design. Thus, those situations where as \( N \rightarrow \infty \) the missing-value patterns tend to exclude certain time points are not appropriate for this analysis.

Assumption 3.8

Every diagonal element of the quantity

\[ Q = Z'(\partial\Omega^{-1}/\partial \psi_k) \otimes (\partial\Omega^{-1}/\partial \psi_k)Z / (T^*)^2 \]

converges as \( N \rightarrow \infty \), uniformly for \( \psi \) in any compact set \( (k=1,2,...,p+q) \).
In light of the block diagonal structure of \( Q \), this quantity reduces to

\[
Q = \sigma^2 \sum \Gamma Z_i' (\partial G_i^{-1}/\partial \psi_k) G_i (\partial G_i^{-1}/\partial \psi_k) Z_i / (T*)^2.
\]

**Assumption 3.9**

Let \( p_{a,N} = n_a / N \) represent the proportion of observational units revealing the \( a \)-th missing-value pattern based upon a sample of size \( N \) \((a = 1, 2, \ldots, 2^T - 1)\). We assume that there exist quantities \( \pi_a \) \((0 \leq \pi_a \leq 1, \sum \pi_a = 1)\) such that as \( N \to \infty \), \( p_{a,N} \to \pi_a \) for all \( a = 1, 2, \ldots, 2^T - 1 \).

The quantities \( \pi_a \) represent the proportion of the population units revealing the \( a \)-th missing-value pattern. The assumption supposes that the observed proportions converge to their fixed population counterparts as \( N \) becomes large. Alternatively, each \( \pi_a \) represents the probability that an observational unit selected at random would reveal the \( a \)-th missing-value pattern, and we assume that these probabilities remain constant as each new observational unit is added. This is an entirely reasonable assumption to make if individuals are selected according to simple random sampling, and is not thought to be an overly restrictive assumption.

Now, consider the following two results.

**Lemma 3.3.1:**

The quantity \( Q = \text{tr}[ (\partial \Omega^{-1}/\partial \psi_k) \Omega (\partial \Omega^{-1}/\partial \psi_k) \Omega ] / T^* \) converges as \( N \to \infty \) to a finite function of \( \psi \), uniformly for \( \psi \) in any compact set \((k = 1, 2, \ldots, p+q)\).

**Proof:**

Using the expression (3.1.18) for \( Q \) in terms of the missing-value patterns, it also follows that \( \partial \Omega^{-1}/\partial \psi_k \) is a block diagonal matrix with typical diagonal block \( I_{n_a} \otimes (\partial G_a^{-1}/\partial \psi_k) \). Multiplying these block diagonal structures, and using the properties of the trace operator, this quantity reduces to

\[
Q = (\sigma^4 / T^*) \text{tr} \left[ \pi_a [I_{n_a} \otimes (\partial G_a^{-1}/\partial \psi_k)] G_a (\partial G_a^{-1}/\partial \psi_k) G_a \right]
\]

- 54 -
\[ Q = \sum \sigma^4 \sum n_a \text{tr}\{(\partial \Sigma_a^{-1}/\partial \psi_k) \Sigma_a (\partial \Sigma_a^{-1}/\partial \psi_k) \Sigma_a\}, \]

with the summation extending over the entire set of \(2T-1\) missing-value patterns (although some of the \(n_a\) may be 0). The total number of observed data elements \(T^* = \sum n_a T_a\), and dividing the numerator and denominator by \(N\), it follows that

\[ Q = \sum \sigma^4 \sum n_a \text{tr}\{(\partial \Sigma_a^{-1}/\partial \psi_k) \Sigma_a (\partial \Sigma_a^{-1}/\partial \psi_k) \Sigma_a\} / \sum n_a T_a. \tag{3.3.5} \]

For each missing value pattern, the trace is a constant with respect to \(N\), and since each \(p_{a_n} \rightarrow \pi_a\) as \(N \rightarrow \infty\), it follows that as \(N \rightarrow \infty\),

\[ Q \rightarrow \sigma^4 \sum \pi_a \text{tr}\{(\partial \Sigma_a^{-1}/\partial \psi_k) \Sigma_a (\partial \Sigma_a^{-1}/\partial \psi_k) \Sigma_a\} / \bar{T}, \tag{3.3.6} \]

where \(\bar{T} = \sum n_a T_a\) is the overall "average" number of repeated measures per individual in the population \((1 \leq \bar{T} \leq T)\). Thus, \(Q\) converges and the result is established.

Q.E.D.

Lemma 3.3.2:

The quantity \(Q = \text{tr}\{(\partial^2 \Omega^{-1}/\partial \psi_m \partial \psi_n) \Omega\}^2 / (T^*)^2\) converges as \(N \rightarrow \infty\) to zero, uniformly for \(\psi\) in any compact set \((m, n = 1, 2, \ldots, p+q)\).

Proof:

Using the same argument as above, it can be shown that

\[ Q = \sum \sigma^4 \sum n_a \text{tr}\{(\partial^2 \Sigma_a^{-1}/\partial \psi_m \partial \psi_n) \Sigma_a\}^2 / (\sum n_a T_a)^2. \]

Dividing numerator and denominator by \(N^2\), it follows that

\[ Q = \sum \sigma^4 \sum p_{a_n} \text{tr}\{(\partial^2 \Sigma_a^{-1}/\partial \psi_m \partial \psi_n) \Sigma_a\}^2 / N(\sum p_{a_n} T_a)^2. \tag{3.3.7} \]

Since each \(p_{a_n} \rightarrow \pi_a\), by virtue of the \(N\) in the denominator, \(Q\) must converge to 0, and the result is established.

Q.E.D.
Theorem 3.3.2: [Magnus(1978), p. 295]

Under the Assumptions 3.1 - 3.9, and Lemmas 3.1.1, 3.3.1, and 3.3.2, the maximum likelihood estimator of $\theta' = [p' \quad \sigma^2]$ is weakly consistent, asymptotically normal, and asymptotically efficient.

3.3.3 **Asymptotic Covariance of the Linear Model Parameters**

Statistical theory shows that the Cramer-Rao lower bound for the covariance matrix of an estimator $\hat{\theta}$ is equal to the inverse of the Fisher information matrix, $I(\theta) = -E[\partial^2 \Lambda(\theta)/\partial \theta \partial \theta']$ where $\theta' = [p' \quad \sigma^2]$ represents the parameter set, and $\Lambda(.)$ is the log likelihood function. This matrix may be partitioned as

$$I(\theta) = \begin{bmatrix}
i(p) & i(p, \sigma) & i(p, \sigma^2) \\
 i(p, \sigma) & i(\sigma) & i(\sigma, \sigma^2) \\
 (\text{sym.}) & i(\sigma^2)
\end{bmatrix} \quad (3.3.8)$$

where, in general, $i(p, \sigma) = -E[\partial^2 \Lambda(\theta)/\partial p \partial \sigma']$. In this section the structure of $I(\theta)$ is explored with a view towards deriving the asymptotic covariance matrix of $\theta$.

**Lemma 3.3.3:**

Under the conditions of Theorem 3.2.1,

$$i(p) = (1/\sigma^2) [\Sigma Z_i'G_i^{-1}Z_i]$$

**Proof:**

From the previous section, we have that

$$(\partial \Lambda/\partial p) = (1/\sigma^2) [\Sigma Z_i'G_i^{-1}Y_i - (\Sigma Z_i'G_i^{-1}Z_i)\theta].$$

Whence it follows that,

$$(\partial^2 \Lambda/\partial p \partial \sigma') = -(1/\sigma^2) \Sigma Z_i'G_i^{-1}Z_i \quad (3.3.9)$$

Taking expectations and the negative of this expression, it immediately follows that

$$i(p) = (1/\sigma^2) [\Sigma Z_i'G_i^{-1}Z_i] \quad (3.3.10)$$

Q.E.D.
Lemma 3.3.4:

Under the conditions of Theorem 3.2.1,

\[ i(\bar{\theta}, \psi) = 0. \]

Proof:

The first derivative of \( \psi \) w.r.t. \( \bar{\theta} \) may be written as

\[ (\partial \psi / \partial \bar{\theta}) = (1/\sigma^2) \Sigma \psi' \Sigma^{-1} (\psi' \Sigma^{-1} \bar{\theta}). \]

Using Lemma 3.2.2, it follows that

\[ (\partial \psi / \partial \bar{\theta}) \Sigma \psi' \Sigma^{-1} \Sigma^{-1} (\psi' \Sigma^{-1} \bar{\theta}) \]

Taking expectations, since \( E(\psi' \Sigma^{-1} \bar{\theta}) = 0 \), it immediately follows that

\[ i(\bar{\theta}, \psi) = -E(\partial \psi / \partial \bar{\theta}) \Sigma = 0, \quad (3.3.11) \]

for \( k = 1, 2, \ldots, p \cdot q \). Thus,

\[ i(\bar{\theta}, \psi) = 0. \quad (3.3.12) \]

Q.E.D.

Lemma 3.3.5:

Under the conditions of Theorem 3.2.1,

\[ i(\bar{\theta}, \sigma^2) = 0. \]

Proof:

Differentiating (3.2.7) w.r.t. \( \sigma^2 \) results in

\[ (\partial \psi / \partial \sigma) = -(1/\sigma^4) \Sigma \psi' \Sigma^{-1} (\psi' \Sigma^{-1} \bar{\theta}). \quad (3.3.13) \]

Taking expectations, since \( E(\psi_i) = \Sigma_i \bar{\theta} \), it follows that

\[ i(\bar{\theta}, \sigma^2) = (1/\sigma^4) \Sigma \psi' \Sigma^{-1} (\Sigma_i \bar{\theta}) = 0. \quad (3.3.14) \]

Q.E.D.

Lemma 3.3.6:

Under the conditions of Theorem 3.2.1,

\[ i(\psi_m, \psi_n) = (1/2) \text{tr} \left[ \Sigma K_i' \Sigma^{-1} K_i (\partial \Sigma / \partial \psi_m) K_i' \Sigma^{-1} K_i (\partial \Sigma / \partial \psi_n) \right]. \]

for \( m, n = 1, 2, \ldots, p \cdot q \).
Proof:

From Theorem 3.2.1, we have that
\[
(\partial \Lambda / \partial \psi_m) = (1/2\sigma^2) \sum_i(g_i^{-1} K_i (\partial g / \partial \psi_m) K_i g_i^{-1} u_i - (1/2) \text{tr}[\nabla (\partial g / \partial \psi_m)].
\]
Considering the second term first, it can be shown that for any two conformable matrix functions A and B of \( \psi_n \),
\[
(\partial / \partial \psi_n) \text{tr}[AB] = \text{tr}[(\partial A / \partial \psi_n) B] + \text{tr}[A (\partial B / \partial \psi_n)].
\]
(3.3.15)
Now, since \( \nabla = \sum_i K_i g_i^{-1} K_i \), it follows that
\[
(\partial / \partial \psi_n) \left[ (-1/2) \text{tr}[\nabla (\partial g / \partial \psi_m)] \right]
= (-1/2) \left[ \text{tr}[\sum_i K_i g_i^{-1} K_i (\partial g / \partial \psi_n) K_i g_i^{-1} K_i (\partial g / \partial \psi_m) K_i g_i^{-1} K_i (\partial g / \partial \psi_m)] + \text{tr}[\nabla (\partial^2 g / \partial \psi_m \partial \psi_n)] \right].
\]
(3.3.16)
For the first term, the usual rules of matrix differentiation are used.
\[
(\partial / \partial \psi_n) \left[ (1/2\sigma^2) \sum_i(g_i^{-1} K_i (\partial g / \partial \psi_m) K_i g_i^{-1} u_i) \right]
= (1/2\sigma^2) \left[ - \sum_i(g_i^{-1} K_i (\partial g / \partial \psi_n) K_i g_i^{-1} K_i (\partial g / \partial \psi_m) K_i g_i^{-1} K_i (\partial g / \partial \psi_m)] + \sum_i(g_i^{-1} K_i (\partial^2 g / \partial \psi_m \partial \psi_n) K_i g_i^{-1} K_i (\partial g / \partial \psi_m)] \right]
- \sum_i(g_i^{-1} K_i (\partial g / \partial \psi_m) K_i g_i^{-1} K_i (\partial g / \partial \psi_n) K_i g_i^{-1} K_i (\partial g / \partial \psi_n)]
\]
The third term is the transpose of the first; and since each term is a scalar, equal to its trace, one may used the properties of the trace operator to arrive at
\[
= (1/2\sigma^2) \text{tr}[K_i g_i^{-1} u_i u_i' g_i^{-1} K_i (\partial g / \partial \psi_n)]
- (1/\sigma^2) \text{tr}[K_i g_i^{-1} u_i u_i' g_i^{-1} K_i (\partial g / \partial \psi_m) K_i (\partial g / \partial \psi_n)]
\]
Taking expectations, since \( u_i = y_i - z_i \), \( \text{E}(u_i u_i') = \text{Var}(u_i) = \sigma^2 g_i \). Thus, after some algebraic manipulations, it follows that
\[
\text{E}(\cdot ) = (1/2) \text{tr}[\left( \sum_i g_i^{-1} K_i (\partial^2 g / \partial \psi_m) K_i g_i^{-1} K_i (\partial g / \partial \psi_n) \right) + \text{tr}[\nabla (\partial^2 g / \partial \psi_m \partial \psi_n)]
\]
(3.3.17)
Thus, combining these two results, we have
\[
-\text{E}(\partial \Lambda / \partial \psi_m \partial \psi_n)
= - (1/2) \text{tr}[\nabla (\partial^2 g / \partial \psi_m \partial \psi_n)]
+ \text{tr}[\sum_i g_i^{-1} K_i (\partial g / \partial \psi_m) K_i g_i^{-1} K_i (\partial g / \partial \psi_n)]
- (1/2) \text{tr}[\sum_i g_i^{-1} K_i (\partial g / \partial \psi_m) K_i g_i^{-1} K_i (\partial g / \partial \psi_n)]
- 58 -
\[ + \frac{1}{2} \text{tr}[V(\partial^2 \omega/\partial \psi_m \partial \psi_n)] \]

Hence

\[ i(\psi_m, \psi_n) = (1/2) \text{tr}[\Sigma K_i'G_i^{-1}K_i(\partial G/\partial \psi_m) K_i'G_i^{-1}K_i(\partial G/\partial \psi_n)]. \]

(3.3.18)

for \(m, n = 1, 2, \ldots, p+q\).

Q.E.D.

Lemma 3.3.7:

Under the conditions of Theorem 3.2.1

\[ i(\psi_k, \sigma^2) = (1/2\sigma^2) \text{tr}[V(\partial G/\partial \psi_k)] \]

for \(k = 1, 2, \ldots, p+q\), where \(V = \Sigma K_i'G_i^{-1}K_i\).

Proof:

From the ML equations, we have that

\[ (\partial \lambda/\partial \sigma^2) = (-T*/2\sigma^2) + (1/2\sigma^4) \Sigma u_i'u_i'G_i^{-1}u_i. \]

Applying Lemma 3.2.2, it follows that

\[ (\partial^2 \lambda/\partial \sigma^2 \partial \psi_k) = -(1/2\sigma^4) \Sigma u_i'u_i'G_i^{-1}K_i[\partial G/\partial \psi_k]K_i'G_i^{-1}u_i. \]

Each element in the summation is a scalar, equal to its trace, and using the permutation property of the trace operator, this reduces to

\[ (\partial^2 \lambda/\partial \sigma^2 \partial \psi_k) = (-1/2\sigma^4) \text{tr}[[\Sigma K_i'G_i^{-1}u_i'u_i'G_i^{-1}K_i][\partial G/\partial \psi_k]]. \]

Taking expectations, since \(u_i = \psi_i - Z_i \xi\), \(E(u_i'u_i') = \text{Var}(u_i) = \sigma^2 G_i\). Thus

\[ -E[\partial^2 \lambda/\partial \sigma^2 \partial \psi_k] = (1/2\sigma^4) \text{tr}[[\Sigma K_i'G_i^{-1}(\sigma^2 G_i)G_i^{-1}K_i][\partial G/\partial \psi_k]]. \]

Hence

\[ i(\psi_k, \sigma^2) = (1/2\sigma^2) \text{tr}[V(\partial G/\partial \psi_k)] \]

(3.3.19)

for \(k = 1, 2, \ldots, p+q\) where \(V = \Sigma K_i'G_i^{-1}K_i\).

Q.E.D.
Lemma 3.3.8:

Under the conditions of Theorem 3.2.1,
\[ i(\sigma^2) = T^*/2\sigma^4. \]

Proof:

It had been shown that
\[ \frac{\partial^2 \lambda}{\partial \sigma^2} = \left(-T^*/2\sigma^4\right) + \left(1/\sigma^6\right) \Sigma \left(\mathbf{y}_i - \mathbf{z}_i \mathbf{g}\right) \mathbf{g}^{-1} \left(\mathbf{y}_i - \mathbf{z}_i \mathbf{g}\right)^t. \]

Differentiating w.r.t. \( \sigma^2 \),

\[ \left[\frac{\partial^2 \lambda}{\partial \sigma^2}\right] = \left(T^*/2\sigma^4\right) - \left(1/\sigma^6\right) \Sigma \text{tr} \left[\mathbf{g}^{-1} \left(\mathbf{y}_i - \mathbf{z}_i \mathbf{g}\right) \left(\mathbf{y}_i - \mathbf{z}_i \mathbf{g}\right)^t\right]. \]

(3.3.20)

As we had used above, \( \mathbb{E} \left(\mathbf{y}_i - \mathbf{z}_i \mathbf{g}\right) \left(\mathbf{y}_i - \mathbf{z}_i \mathbf{g}\right)^t = \text{Var}(\mathbf{u}_i) = \sigma^2 \mathbf{g}_i \), so that

\[ -\mathbb{E}\left[\frac{\partial^2 \lambda}{\partial \sigma^2}\right] = \left(-T^*/2\sigma^4\right) + \left(1/\sigma^6\right) \Sigma \text{tr} \left[\mathbf{g}_i^{-1} (\sigma^2 \mathbf{g}_i)\right]. \]

Since \( \text{tr}(\mathbf{g}_i^{-1} \mathbf{g}_i) = \text{tr}(\mathbf{I} \mathbf{T}_i) = T_i \), and by definition \( \Sigma \mathbf{T}_i = T^* \), it follows that

\[ i(\sigma^2) = -\mathbb{E}\left[\frac{\partial^2 \lambda}{\partial \sigma^2}\right] = \left(-T^*/2\sigma^4\right) + \left(T^*/\sigma^4\right) = T^*/2\sigma^4. \]

(3.3.21)

Q.E.D.

Theorem 3.3.3:

Under the conditions of Theorem 3.2.1,
\[ \text{Asy.Var.}(\mathbf{g}) = \sigma^2 [\Sigma \mathbf{Z}_i \mathbf{g}_i^{-1} \mathbf{Z}_i]^{-1}. \]

Proof:

The Fisher information matrix had been written as

\[ \mathbf{I}(\sigma) = \begin{bmatrix} i(\mathbf{g}) & i(\mathbf{g}, \sigma^2) & i(\mathbf{g}, \sigma^2) \\ i(\mathbf{g}, \sigma^2) & i(\sigma^2) & i(\sigma^2) \\ (\text{sym.}) & i(\sigma^2) & i(\sigma^2) \end{bmatrix} \]

Now, from the previous lemmas, this structure reduces to

\[ \mathbf{I}(\mathbf{g}, \mathbf{w}, \sigma^2) = \begin{bmatrix} \sigma^{-2} [\Sigma \mathbf{Z}_i \mathbf{g}_i^{-1} \mathbf{Z}_i] & 0 & 0 \\ 0 & W_{zz} & W_{z3} \\ 0 & W_{z3} & T^*/2\sigma^4 \end{bmatrix} \]

(3.3.22)
where the elements of $W_{zz}$ were derived in Lemma 3.3.6 and the elements of $W_{z3}$ were derived in Lemma 3.3.7. This is a block diagonal matrix and the inverse of a block diagonal matrix is also a block diagonal matrix with elements equal to the inverse of the individual blocks. In particular, in the upper left-hand corner, we have $\sigma^2[\Sigma Z_i' G_i^{-1}Z_i]^{-1}$ corresponding to $\hat{\theta}$. Thus,

$$\text{Asy.Var}(\hat{\theta}) = \sigma^2[\Sigma Z_i' G_i^{-1}Z_i]^{-1}. \quad (3.3.23)$$

Q.E.D.

**Corollary:**

Under the conditions of Theorem 3.2.1, the linear model parameter estimator $\hat{\theta}$ is asymptotically independent of the covariance parameter estimators $\hat{\varphi}$ and $\hat{\sigma}^2$.

**Proof:**

This follows from the block diagonal structure of $\mathbb{I}(\theta)$ and the asymptotic normality of the joint parameter set.

Q.E.D.

Note that the matrix $W_{zz}$ will not, in general, be identically $0$, so that the time series parameter estimator is asymptotically correlated with $\hat{\sigma}^2$. 

- 61 -
4.1 INTRODUCTION

It can be shown that the first-order ML equations defined in Theorem 3.2.1 for the time series parameters $\theta_k$ ($k=1,2,...,p+q$) form a set of non-linear simultaneous equations in $\theta$. In general, there is no closed-form expression for the solution $\hat{\theta}$ of these equations, and one must therefore resort to numerical methods techniques of the type discussed in Chapter II to iterate to a solution. Oberhofer & Kmenta (1974) described one technique for iterating to a solution to these equations. The procedure consists of alternating between expressions for $\hat{\theta}$ and $\hat{\theta}^2$ for the current value of $\theta$, and then, conditionally upon these values, solving the ML equations for $\theta$. It was shown that under general regularity conditions this iterative procedure always converges to at least a local maximum in the parameter space.

In this chapter, the resources required to perform non-linear optimization will be discussed in detail. A method equivalent to that described by Oberhofer & Kmenta is based upon the "concentrated" log likelihood function, and an expression for this function as well as the corresponding gradient vector will be derived. Initial estimators for the time series will also be presented. Thus, this chapter describes the mechanics of performing non-linear optimization for this particular model, and the ensuing properties are discussed.
4.2 ITERATION SEQUENCE

In Theorem 3.2.1, it had been shown that the ML estimators for \( \rho \) and \( \sigma^2 \) can both be expressed as functions of the time series parameters \( \psi \). Define the quantities,

\[
S(\psi) = \mathbf{Z}_i' \mathbf{G}_i^{-1}(\psi) \mathbf{Z}_i, \\
\tilde{\mathbf{r}}(\psi) = \mathbf{Z}_i' \mathbf{G}_i^{-1}(\psi) \mathbf{r}_i, \\
\tilde{\mathbf{Q}}(\psi) = \mathbf{Z}_i' \mathbf{G}_i^{-1}(\psi) \mathbf{Z}_i.
\]  

(4.2.1)
(4.2.2)
(4.2.3)

Note that these quantities are dependent upon the parameter set only through the time series parameters \( \psi \). It therefore follows that

\[
\hat{\mathbf{Q}} = \left[ \mathbf{Z}_i' \hat{\mathbf{G}}_i^{-1} \mathbf{Z}_i \right]^{-1} \left[ \mathbf{Z}_i' \hat{\mathbf{G}}_i^{-1} \mathbf{r}_i \right] = \hat{\mathbf{Q}}^{-1} \hat{\mathbf{E}},
\]

(4.2.4)

where \( \hat{\mathbf{Q}} = \mathbf{Q}(\hat{\psi}) = \mathbf{Z}_i' \mathbf{G}_i^{-1}(\hat{\psi}) \mathbf{Z}_i \), with similar definitions for \( \hat{S} \) and \( \hat{\mathbf{E}} \). As well,

\[
\hat{\omega}^2 = \mathbf{r}_i' \hat{\mathbf{G}}_i^{-1} \mathbf{r}_i + \hat{\psi}' \left[ \mathbf{Z}_i' \hat{\mathbf{G}}_i^{-1} \mathbf{Z}_i \right] \hat{\mathbf{E}},
\]

(4.2.5)

and substituting the above expression for \( \hat{\mathbf{Q}} \) into this expression yields

\[
\hat{\omega}^2 = \left[ \hat{\mathbf{S}} - \hat{\mathbf{E}}' \hat{\mathbf{Q}}^{-1} \hat{\mathbf{E}} \right] / T^*.
\]

(4.2.6)

Richards (1961) specifically described a method of performing ML estimation when the ML estimators for one set of parameters (in our case, \( \rho \) and \( \sigma^2 \)) can be expressed as a function of a second set of parameters (in our case, \( \psi \)). He suggested substituting \( \hat{\mathbf{Q}}(\psi) \) and \( \hat{\omega}^2(\psi) \) directly into the (log) likelihood function, and to then "... proceed as if performing an ordinary maximum likelihood estimation ..." on the resulting expression. The ML estimators for \( \rho \) and \( \sigma^2 \) are ultimately determined by substituting the ML estimate for \( \psi \) into the expressions for \( \hat{\mathbf{Q}} \) and \( \hat{\omega}^2 \). Clearly, this procedure is equivalent to the iterative scheme described in Oberhofer & Kmenta. Rather than solving for \( \rho \) and \( \sigma^2 \) in terms of \( \psi \) at each iteration, these quantities are incorpo-
rated implicitly in the log likelihood function. Moreover, finding that value of \( \psi \) which maximizes this function is equivalent to solving the corresponding first-order ML equations. Thus, properties observed under the Oberhofer & Kmenta scheme will also apply using this technique.

Now substituting the above expressions for \( \hat{\theta} \) and \( \hat{\sigma} \) into (3.2.2) yields the "concentrated log likelihood function" (i.e., "concentrated" with respect to the parameter, \( \psi \)), and after some algebraic manipulations, this may be expressed as:

\[
\Lambda^*(\psi) = c - (T^*/2)\ln(S - \mathbf{Q}^{-1}\mathbf{F}) - (1/2)I\ln|\mathbf{G}_i|, \quad (4.2.7)
\]

where the constant \( c = (-T^*/2)[1 + \ln(2\pi) - \ln(T^*)] \). It is this function for which we seek a maximizing value, \( \hat{\psi} \). The ML estimation procedure may be summarized in two steps:

1. Find \( \hat{\psi} \) which maximizes (4.2.7), and then
2. Substitute this value into the expressions for \( \hat{\theta} \) and \( \hat{\sigma} \) to derive their ML estimates.

While some non-linear optimization methods do not require derivatives, it is generally the case that gradient methods are more efficient, iterating more quickly to the optimal solution. The derivative of the concentrated log likelihood function with respect to each \( \psi_k \) is established in the following theorem, but first, the following useful result is provided.

**Lemma 4.2.1:**

Under the model and conditions considered thusfar, if \( \mathbf{r} = \mathbf{F}_i\mathbf{G}_i^{-1}\mathbf{v}_i \), then

\[
\mathbf{r}^{(k)} = (\partial \mathbf{r} / \partial \psi_k) = - \mathbf{F}_i \mathbf{G}_i^{-1} \mathbf{K}_i \left[ \partial \mathbf{G}_i / \partial \psi_k \right] \mathbf{K}_i \mathbf{G}_i^{-1} \mathbf{v}_i \quad (4.2.8)
\]

for \( k = 1, 2, \ldots, p+q \). 

---

- 64 -
Proof:

This result follows in a straightforward manner from Lemma 3.2.2.

Q.E.D.

Corollary:

Under the model and conditions considered thusfar,

\[ S^{(k)} = (\partial S/\partial \psi_k) = - \sum_i G_i^{-1} K_i [G_i/\partial \psi_k] K_i' G_i^{-1} y_i \] (4.2.9)

and

\[ Q^{(k)} = (\partial Q/\partial \psi_k) = - \sum_i G_i^{-1} K_i [G_i/\partial \psi_k] K_i' G_i^{-1} z_i \] (4.2.10)

Proof:

These results follow naturally from the definitions of \( S \) and \( Q \) and the logic of the lemma.

Q.E.D.

Note that since each \( z_i = x_i' \otimes K_i \), it follows that \( Q^{(k)} \) may be written as:

\[ Q^{(k)} = - \sum_i [x_i x_i' \otimes K_i' G_i^{-1} K_i (G_i/\partial \psi_k) K_i' G_i^{-1} K_i] \]

Theorem 4.2.1:

Under the model and conditions considered thusfar, the derivative of the concentrated likelihood function w.r.t. \( \psi_k \) is given by:

\[
q_k = (\partial \Lambda^*/\partial \psi_k) = -[T/2(S - \bar{x}' Q^{-1} \bar{x})] \\
[Q^{(k)} - \bar{x}' Q^{-1} \bar{x} + \bar{x}' Q^{-1} Q^{(k)} Q^{-1} \bar{x} - \bar{x}' Q^{-1} \bar{x}^{(k)}] \\
- (1/2)tr[\nabla(\partial G/\partial \psi_k)],
\]

for \( k = 1, 2, \ldots, p+q \), and where \( \nabla = I K_i' G_i^{-1} K_i \), and \( S^{(k)} \), \( \bar{x}^{(k)} \), and \( Q^{(k)} \) had been defined in the previous lemma.
Proof:

From (4.2.7), the concentrated likelihood function is given by:

\[ \Lambda^*(k) = -(T^*/2) \ln(S - r'Q^{-1}r) - (1/2) \sum \ln |G_i| \].

Considering the second term first, from Lemma 3.2.1 we have

\[ \frac{\partial}{\partial \psi_k} \ln |G_i| = \text{tr} \left[ V \frac{\partial G}{\partial \psi_k} \right] \].

(4.2.11)

Next, for the first term, we use two well known rules of matrix differentiation, viz.,

\[ \frac{\partial}{\partial \psi_k} [XYZ] = (\frac{\partial X}{\partial \psi_k})YZ + X(\frac{\partial Y}{\partial \psi_k})Z + XZ(\frac{\partial Y}{\partial \psi_k}), \]

for any conformable matrices \( X, Y, \) and \( Z, \) and

\[ \frac{\partial}{\partial \psi_k} [Q^{-1}] = -Q^{-1} (\frac{\partial Q}{\partial \psi_k}) Q^{-1}. \]

Using these results, it follows that

\[ \frac{\partial}{\partial \psi_k} [S - r'Q^{-1}r] = S(k) - r(k)'Q^{-1}r + r'Q^{-1}Q(k)Q^{-1}r - r'Q^{-1}r(k). \]

(4.2.12)

So that

\[ \frac{\partial}{\partial \psi_k} [\ln(S - r'Q^{-1}r)] = [S - r'Q^{-1}r]^{-1} \]

\[ [S(k) - r(k)'Q^{-1}r + r'Q^{-1}Q(k)Q^{-1}r - r'Q^{-1}r(k)]. \]

(4.2.13)

Thus combining these results, we have:

\[ q_k = \frac{\partial \Lambda^*}{\partial \psi_k} = -\frac{T^*}{2(S - r'Q^{-1}r)} \]

\[ [S(k) - r(k)'Q^{-1}r + r'Q^{-1}Q(k)Q^{-1}r - r'Q^{-1}r(k)] \]

\[ - (1/2) \text{tr} \left[ V \frac{\partial G}{\partial \psi_k} \right] \].

(4.2.14)

where \( S(k), r(k), \) and \( Q(k) \) had been defined in the lemma.

Q.E.D.

The derivative, \( q_k \), of Theorem 4.2.1 is called the "gradient", and the resulting \((p+q)\)-dimensional vector, \( q \), is called the gradient vector.

As discussed in Chapter II, non-linear optimization is an iterative technique, where the current estimate of the pa-
rameter vector is updated to produce the next estimate. Let $\hat{y}^{(u)}$ represent the estimate of $\hat{y}$ on the $u$-th iteration ($u=0,1,2,...$). Then, in general, we write

$$ \hat{y}^{(u+1)} = \hat{y}^{(u)} - tP^{-1}(\hat{y}^{(u)})g(\hat{y}^{(u)}). $$

(4.2.15)

Belsley (1980) strongly recommended the Newton-Raphson approach, that is where $t=1$, and $P$ is the matrix of mixed second-order partial derivatives,

$$ P = \frac{\partial^2 A^*(\hat{y})}{\partial \hat{y} \partial \hat{y}^*}, $$

evaluated at $\hat{y}=\hat{y}^{(u)}$. An analytic expression for this quantity can be derived by performing an analysis similar to that which arrived at Theorem 4.2.1.

However, this expression is very cumbersome, and there are several alternative procedures for the applied worker. Rather than evaluating this equation analytically, one may approximate it numerically by successively perturbing each element of the parameter vector, and manipulating the resulting change in the gradient vector. Box & Jenkins (1970) suggested that the Gauss-Newton approach provides a useful approximation to the Newton-Raphson technique, and several other authors including Ansley & Newbold (1980), Kang (1975), and Spitzer (1979) have successfully applied this procedure. In general, all the gradient methods described in Chapter II can be expected to perform reasonably well in this setting.

Thus, the iteration sequence may be summarized as follows:

1. For $u=0,1,2,...$, let $\hat{y}^{(u)}$ be the estimate of $\hat{y}$ at the $u$-th iteration.
2. Using Theorem 4.2.1, evaluate the vector of partial derivatives,

$$ g = \frac{\partial A^*(\hat{y})}{\partial \hat{y}} $$

at $\hat{y}^{(u)}$.
3. Evaluate the matrix of second-order partial derivatives,

$$ P = \frac{\partial^2 A^*(\hat{y})}{\partial \hat{y} \partial \hat{y}^*}. $$

- 67 -
at \( \hat{\omega}^{(u)} \), either analytically or numerically as appropriate for the gradient procedure being used.

4. The estimate of \( \hat{\omega} \) on the next iteration is taken as

\[
\hat{\omega}^{(u+1)} = \hat{\omega}^{(u)} - P^{-1}g.
\]

5. Continue until convergence is achieved.

We now consider the small-sample properties of the resulting estimator for \( \hat{\omega} \). First, make the following regularity assumption.

**Assumption 4.1** [Regularity Conditions]

There exist fixed numbers \( \lambda > 0 \) and \( M > 0 \) such that when

\[
U = \{ \hat{\omega} \mid \lambda \leq |\hat{\omega}| \leq M \}
\]

the following is true: with probability 1, a number, say \( s \), may be found such that for any \( \hat{\omega}^* \in \mathbb{R}^T \), and \( \hat{\omega}^* \in U \), we have that \( \Lambda(\hat{\omega}^*, \hat{\omega}^*, \sigma^2) \geq s \) implies that \( \hat{\omega}^*(\hat{\omega}^*) \in U \), where \( \hat{\omega}^*(\hat{\omega}^*) \) denotes the value of \( \hat{\omega} \) that maximizes \( \Lambda(\hat{\omega}^*, \hat{\omega}^*, \sigma^2) \).

**Theorem 4.2.2** [Don & Magnus (1980), p. 525]

Under the Assumptions 3.1 - 3.6, and 4.1, and Lemma 3.1.1, the iteration sequence described above converges to a solution of the ML equations for \( \hat{\omega} \) (existence), and it is an unbiased estimator (unbiasedness).

As Don & Magnus indicated, it a very difficult (if not impossible) proposition to verify Assumption 4.1 for any particular model. For the AR(1) case, however, it is suggested that this is satisfied by the existence of a small quantity, say \( \epsilon > 0 \), such that \( |\hat{\omega}^{(u)}| < 1 - \epsilon \) on every iteration. This is a condition which is very likely to be observed in practice.
4.3 INITIAL ESTIMATES OF THE PARAMETERS

4.3.1 Sample Autocorrelations

All the non-linear optimization procedures described in Chapter II require that an estimate of the parameter vector be made available to initiate the calculations. Some care must be taken for the choice can influence how efficiently, and indeed how successfully, the procedure performs. In this section, these issues are considered, and a technique is proposed for deriving initial estimates of the time series parameters.

As a general statement, there is some feeling that iterations should begin from a consistent estimator of the parameters. Consistency is an important concept, of course, in BAN estimation since a BAN estimator exists for the linear model parameter matrix, $B$, only insofar as a consistent estimator exists for the covariance structure. Both LeVan ge (1983) and Kleinbaum (1970) devoted a considerable amount of attention to this property.

Moreover, Jennrich (1969) considered the convergence properties of the Gauss-Newton algorithm for least squares and ML estimation. He found that provided the initial estimate $\hat{\theta}(0)$ is sufficiently "close" to $\theta$, and iterations are continued long enough, the algorithm will eventually converge to $\theta$. Jennrich quoted from Hartley & Booker (1965) who suggested that the sequence should begin from a consistent estimator, and a number of workers, e.g., Nicholls, Pagan & Terrel (1975), echoed this belief. While there may be some theoretical comfort in these results, there is little to appeal to the applied worker. Even if the iteration sequence is begun from a consistent estimator, there is no guarantee that convergence will be realized within a practical number of iterations.

There is no suggestion, however, that consistency is a necessary condition for convergence. Bard [(1974), Sec. 5-20], in fact, cautioned "... not to exaggerate the importance of finding a good initial guess ...," and suggested
that non-linear optimization procedures are reasonably robust with respect to starting values. Other authors agree, and Belsley (1980) for example, observed that even OLS estimators yield satisfactory results for the complete-data univariate regression problem. The operating criterion is "reasonableness" within the context of the problem at hand.

Recall from the previous chapter that we had defined the random variables

\[ a_i(t) = \begin{cases} 1 & \text{if } Y_i(t) \text{ is recorded} \\ 0 & \text{if } Y_i(t) \text{ is missing,} \end{cases} \]

signifying the presence or absence of the corresponding data elements. It had been assumed that these random variables are mutually independent of the \( \{Y_i(t)\} \) and hence the \( \{u_i(t)\} \). Parzen (1963) defined the "amplitude modulated" time series as

\[ v_i(t) = a_i(t)u_i(t) = \begin{cases} u_i(t) & \text{if } Y_i(t) \text{ observed} \\ 0 & \text{if } Y_i(t) \text{ missing.} \end{cases} \] (4.3.1)

[Parzen's original formulation holds, in fact, for more general stochastic processes, for example, a sequence of Bernoulli trials.]

Let \( \bar{a}_i = \frac{\sum a_i(t)}{T} \) be the proportion of data elements observed for the i-th individual, and define

\[ c_i(j,a) = \frac{\sum a_i(t)a_i(t+j)}{T} \] (4.3.2)

to represent the number of pairs of observations from the i-th individual, separated by j time points, for which both data elements are observed (weighted by \( T^{-1} \)). Similarly, one may define

\[ c_i(j,v) = \frac{\sum v_i(t)v_i(t+j)}{T} \] (4.3.3)

with an analogous interpretation. Parzen proposed the quantity

\[ \bar{c}_i(j) = \frac{c_i(j,v)}{c_i(j,a)} \] (4.3.4)
as an estimator of the autocovariance for the i-th individual at lag j. This is just the arithmetic mean of those products $u_i(t)u_i(t+j)$ for which both data elements are observed. The quantity

$$r_i(j) = c_i(j) / c_i(0)$$  \hspace{1cm} (4.3.5)

is proposed as the estimator of the autocorrelation for the i-th individual at lag j. To combine information across individuals, define the quantities

$$c_j = \Sigma c_i(j,v) / \Sigma c_i(j,a)$$
$$= \Sigma \Sigma v_i(t)v_i(t+j) / \Sigma \Sigma a_i(t)a_i(t+j)$$  \hspace{1cm} (4.3.6)

and

$$r_j = c_j / c_0$$  \hspace{1cm} (4.3.7)

for $j=0,1,2,...$.

To demonstrate that these are "reasonable" estimators, consider the following evidence. Dunsmuir & Robinson(1981) discuss the large-sample properties of estimators (4.3.4) and (4.3.5), but as $T\to\infty$ not as $N\to\infty$. Provided that $\bar{a}_i$ and $c_i(j,a)$ converge to fixed values for $j=0,1,2,...$ and for some stronger regularity conditions, both $c_i(j)$ and $r_i(j)$ are consistent estimators of their population counterparts. Moreover, LaVange(1983) considered the estimator

$$r_{i*} = \Sigma \Sigma v_i(t)v_i(t+1) / \Sigma [v_i(t)]^2$$

for the first-order autocorrelation. She showed that $r_{i*}$ is a consistent estimator of $\rho_1$ as $N\to\infty$. No statements are made, however, concerning higher-order autocorrelations. These results recommend the above estimators for the sample autocovariances and autocorrelations.

4.3.2 Initial Estimates for the Time Series Parameters

Box & Jenkins[(1970), Sec. 6.3] provided a detailed discussion on forming initial estimates of the time series parameters $\phi$ based upon the sample autocorrelations $r_j$. For the AR(1) process, for example, $\phi^{(0)} = r_1$, and for the AR(2),
\[
\phi_1(0) = \frac{r_1(1-r_2)}{(1-r_1^2)} \\
\phi_2(0) = \frac{(r_2-r_1^2)}{(1-r_1^2}). \tag{4.3.8}
\]

In general, for the AR(p) process, the Yule-Walker equations may be written as

\[
P\varphi = r \tag{4.3.9}
\]

where \(P\) is a \(p\) by \(p\) matrix, with typical element \(p_{r,s} = r_{1r-s1}\), and \(r\) is the vector of sample autocorrelations, \(r' = [r_1 \ r_2 \ \ldots \ r_p]\). An initial estimate of the AR parameter vector is a solution to this equation.

For the MA(1) case,

\[
\varphi(0) = \frac{[-1+(1-4r_1^2)^{1/2}]}{2r_1}. \tag{4.3.10}
\]

One solution is the reciprocal of the other, so that one of these values can be chosen to satisfy the invertibility condition that \(|\varphi|<1\).

A method for computing initial estimates of the time series parameters as a function of the sample autocorrelations for the general ARMA model is described at length in Box & Jenkins[1970, Appendix A6.2]. It is based upon a method due to Wilson(1969). Provided that all the sample autocorrelations \(r_j\) defined in (4.3.7) exist, this technique may also be used in this setting to provide an initial estimate for \(\varphi\).

Thus, the technique for computing \(\varphi(0)\) may be summarized as follows.

1. Following LaVange(1983) and using the LMI representation, \(y=Z\varphi+u\), compute the OLS estimator,

\[
b = (Z'Z)^{-1}Z'Y.
\]

This (QT by 1) vector can be transformed into the (Q by T) parameter matrix \(B(0)\) in an obvious way.

2. Form the residuals corresponding to the observed data elements as

\[
\hat{u}_i(t) = y_i(t) - x_i'b(t),
\]

where \(b(t)\) represents the \(t\)-th column of \(B(0)\).
3. Form the "amplitude modulated" residuals,
\[ \hat{v}_i(t) = a_i(t) \hat{u}_i(t). \]

4. Substitute \( \hat{v}_i(t) \) into (4.3.6) and (4.3.7) to derive the sample autocorrelations.

5. Finally, use the method described in Box & Jenkins[(1970), Ch. 6] to form an initial estimate of the time series parameter vector \( \theta \) from the sample autocorrelations.

4.4 COMPUTATIONAL CONSIDERATIONS

Analytic procedures for evaluating the (log) likelihood function had been discussed in detail in Chapter II. The main computational obstacle, of course, is the covariance structure \( G \) which is a square matrix of order \( T \) (the number of repeated measures). In order to avoid calculating \( G^{-1} \) directly, research has centered on deriving a factorization matrix \( P \) such that \( G^{-1} = P'P \), and various techniques had been presented there.

In the incomplete data case, these difficulties take on an added dimension of computational complexity. Savin & White(1978) considered the AR(1) error structure, specifically when there is a gap of \( m \) consecutive missing values, and a factorization matrix was presented. Beyond this report, no general method of deriving a similar transformation matrix has been published for various missing value patterns of the ARMA covariance structure. Thus, factorization methods appear to have limited utility in this setting.

The Kalman filter methodology had also been presented as a powerful analytic tool to sequentially update parameter estimates of the ARMA time series parameters and to evaluate the overall likelihood function. Harvey & Phillips(1979) had applied the Kalman filter to the univariate regression model in the complete-data case under an ARMA covariance pattern. Jones(1980,1984a) and Ansley & Kohn(1983) discussed its application for the incomplete-data case in the univariate and multivariate cases respectively. The Kalman
filter is therefore a viable computational tool in this setting.

It is not possible, however, to compute the gradient vector, discussed in Theorem 4.2.1, simultaneously with the likelihood function using this procedure. The Kalman filter would likely be imbedded as a subroutine in a non-linear optimization procedure which evaluated derivatives numerically. The large number of function evaluations would have an obvious impact on the amount of computer resources required.

Alternatively, if the number of repeated measures is small (e.g., T<20), then evaluating $G^{-1}$, and hence $G_a^{-1}$ for any missing value pattern, may prove to be computationally feasible. The remainder of this section is devoted to describing one methodology for performing these calculations.

Recall that there are A distinct missing value patterns, indexed by the subscript "a", and let $G_a=K_aGK_a'$ represent the covariance structure associated with the a-th missing value pattern. Now the observed data elements in this pattern induce a partition in the covariance structure $G$, and without loss of generality, we assume that the first $T_a$ repeated measures are those which are observed, while the final $(T-T_a)$ repeated measures are missing. Thus, we write

$$G = \begin{bmatrix}
G_{11}(a) & G_{12}(a) \\
G_{21}(a) & G_{22}(a)
\end{bmatrix} \quad (4.4.1)$$

where $G_{11}(a)$ is the covariance structure among the observed data elements in the a-th missing value pattern. The superscript "a" is dropped for convenience where it is understood that these results hold for each missing value pattern.

Goodnight(1979) discussed at length the application of the SWEEP operator in statistical computing. The SWEEP operator is a highly structured sequence of row operations on a symmetric positive semi-definite matrix. The operations single out a particular row (say the j-th row) to drive the calculations, and a central figure is the diagonal element $g_{jj}$. This element is called the "pivot", and the entire operation is called "pivoting on the j-th diagonal element".
In our application, if we sequentially pivot on the diagonal elements of $G_{11}$ (i.e., corresponding to the observed data elements), then $G$ is transformed to become

$$G^* = \begin{bmatrix}
G_{11}^{-1} & G_{11}^{-1}G_{12} \\
-G_{21}G_{11}^{-1} & G_{22} - G_{21}G_{11}^{-1}G_{12}
\end{bmatrix}$$

(4.4.2)

Thus, the desired matrix $G_{11}^{-1}$ may be read in those positions corresponding to observed data elements in this missing value pattern.

The power of the SWEEP operator lies in its ability to advance from one missing value pattern to another. Goodnight showed that the SWEEP operator is cumulative and reversible, and this has the following implications for our purposes. Let $a_2$ represent another missing value pattern. Then to derive $G_{a2}^{-1}$, one pivots $G^*$ on those diagonal elements corresponding to data elements observed in $a_1$ but no longer in $a_2$ (i.e., sweeps them "out"), and then pivots on those diagonal elements corresponding to data elements now observed in $a_2$ but originally missing (i.e., sweeps them "in"). One only pivots on those diagonal elements corresponding to repeated measures which change status from one missing value pattern to the next. The desired $G_{a}^{-1}$ is always read in those positions corresponding to the observed repeated measures in the pattern. By a judicious choice of the sequence in which patterns are processed, the number of sweep operations can be minimized with a corresponding impact on the amount of computer resources required. A by-product of these computations is the determinant $|G_{a}|$, and its value can be similarly updated during the sweep operations from one pattern to the next.

These computations would be feasible when the number of observed data elements is relatively small since relatively few sweep operations would be required. If there were a large number of observed repeated measures, then one could approach the calculations from the other end.
Since the covariance structure $G$ is positive-definite, let $H = G^{-1}$. Then $H$ may be partitioned in a manner corresponding to $G$ above. Ignoring the superscript "a", we may write

$$
\begin{bmatrix}
H_{11} & H_{12} \\
H_{21} & H_{22}
\end{bmatrix}^{-1} = H^{-1} = G =
\begin{bmatrix}
G_{11} & G_{12} \\
G_{21} & G_{22}
\end{bmatrix}
$$

(4.4.3)

From the theory of partitioned matrices, and focussing on the upper left-hand element (corresponding to the observed data elements) it follows that

$$
G_{11} = [H_{11} - H_{12}H_{22}^{-1} H_{21}]^{-1}
$$
i.e.,

$$
G_{11}^{-1} = H_{11} - H_{12}H_{22}^{-1} H_{21}
$$

(4.4.4)

However, sweeping the matrix $H$ on the diagonal elements of $H_{22}$ (corresponding to the missing data elements), results in

$$
H^* =
\begin{bmatrix}
H_{11} - H_{12}H_{22}^{-1} H_{21} & -H_{12}H_{22}^{-1} \\
H_{22}^{-1} H_{21} & H_{22}^{-1}
\end{bmatrix}
$$

(4.4.5)

Thus, the desired quantity is again found in those positions corresponding to the observed repeated measures in this pattern. The same results hold for updating $G_a^{-1}$ and $|G_a|$ from one missing value pattern to the next.

Some research has appeared in the literature for computing the matrix $H$. Judge et al(1980) provided an explicit expression for $H$ and $|H|$ under an AR(1) and an AR(2) covariance structure, while Tiao & Ali(1971) provided similar expressions for the ARMA(1,1) model. Galbraith & Galbraith(1974) presented a method of deriving $H$ under the general ARMA(p,q) model, and explicit expressions were provided for the AR(p), MA(1), and ARMA(1,1) models.
Chapter V
HYPOTHESIS TESTING

5.1 INTRODUCTION

It is generally acknowledged that there are two important steps in any inferential strategy. The first, of course, is to specify a model and to provide expressions for the estimators of the associated parameters, and this has largely been the topic of the last two chapters. Expressions had been presented for the ML estimators of the parameters of the incomplete longitudinal design, and a technique had been described to perform these calculations. A second and equally important component of the inferential process is to specify hypotheses of interest and to derive a procedure for evaluating these assertions. This is the topic of the current chapter.

The distinction between these two components is not as sharp as this discussion suggests. Model building is very frequently an iterative process, with inference alternating between proposing a model for consideration, estimating the associated parameters, and testing for its compatibility with the data, and on the basis of these results, submitting a revised model for scrutiny. Hypothesis testing therefore forms an integral component of the model building function as well. Eventually, a "final" model which provides a parsimonious description of the underlying stochastic process is accepted.

There are two broad categories of hypotheses which will be studied in depth in this chapter. The first examines the covariance structure among the repeated measures. Generally, it is very difficult to attach a physical interpretation to any particular ARMA model; consequently, this largely
serves the model building and identification process. It is significant only insofar as the final model provides an effective vehicle for pursuing more substantive hypotheses. These are the hypotheses among the linear model parameters for assessing differences across subpopulations, for patterns over time, and for the influence of concomitant variables. These issues are also considered in this chapter.

5.1.1 Likelihood Ratio Test Procedures

The technique of ML estimation has been discussed under this particular model. A method of test construction which is closely allied to ML estimation is the Likelihood Ratio (LR) method, originally proposed in 1928 by Neyman and Pearson. It is adopted exclusively for the purpose of this study.

Let \( \Omega \) represent the parameter space under the terms and conditions of the model. An hypothesis imposes restrictions on the parameters, constraining them to a particular region of the parameter space, denoted by \( \omega \). The likelihood ratio test statistic is defined as

\[
\Delta^* = \sup_{a \in \omega} L(a) / \sup_{a \in \Omega} L(a),
\]

(5.1.1)

where \( \theta' = [\theta' \ y' \ \sigma^2] \) represents the parameter set, and \( L(\cdot) \) represents the likelihood function. If the subscript "0" signifies the null hypothesis and the subscript "\( a \)" is used for the alternative or unrestricted case, then ignoring pathological cases, the above equation may be expressed as

\[
\Delta^* = L(\hat{\theta}_0) / L(\hat{\theta}_a),
\]

(5.1.2)

that is, the ratio of the maximum likelihoods attainable under the respective hypotheses.

Frequently, the hypothesis of interest may be confined to a certain subset of the model parameters. Following Kendall & Stuart[(1967), Ch. 24], let \( \theta' = [\theta' \ \eta'] \), where there are \( r \) elements in \( \theta \) and \( s \) elements in \( \eta \) (\( r, s \neq 0 \)). Consider the null hypothesis,
\[ H_0: \mu = \mu_0 \quad (5.1.3) \]

versus the alternative hypothesis,
\[ H: \mu \neq \mu_0. \quad (5.1.4) \]

Then (5.1.2) may be expressed as
\[ \Delta^* = \frac{L(\mu_0, \hat{\mu}_0)}{L(\mu_a, \hat{\mu}_a)}, \quad (5.1.5) \]

where \( \hat{\mu}_0 \) represents the ML estimate of \( \mu \) under the restrictions imposed by \( H_0 \).

The LR test procedure prescribes finding some critical value, say \( c_0 \), such that \( \Delta^* < c_0 \) suggests that \( H_0 \) is not consistent with these data under the model, and should be rejected. The critical value \( c_0 \) is determined from the distribution of \( \Delta^* \). However, under this particular model, this distribution cannot be specified in any useable form, and one must therefore appeal to large-sample properties. A thorough discussion of the manner in which the number of observations becomes large had been presented in Section 3.2.2, and applies in full force for the current discussion.

Let \( \Lambda(\cdot) \) represent the log likelihood function, and consider the following result.

**Theorem 5.1.1:**

Under Assumptions 3.1 - 3.9 and Lemma 3.1.1, and under the null hypothesis (5.1.3), as \( N \to \infty \), the quantity
\[ \Delta = -2 \ln \Delta^* = (-2)[\Lambda(\mu_0, \hat{\mu}_0) - \Lambda(\mu_a, \hat{\mu}_a)] \]
converges in distribution to a random variate having the central chi-square distribution with \( r \) degrees of freedom.

**Proof:**

In Theorem 3.3.1, it had been demonstrated that the density function of \( \mu \) with parameters \( \mu' = [\mu' \ \sigma^2] \) is regular with respect to its first and second derivatives. Wilks[(1962), Th. 13.8.1] showed that under the null hypothesis, if the underlying density is regular with respect to its second derivatives, the quantity
\[ \Delta = (-2)[\Lambda(\hat{\mu}_0, \hat{\sigma}_0) - \Lambda(\hat{\mu}_a, \hat{\sigma}_a)] \]  

(5.1.6)

converges as \( N \to \infty \) to a chi-square with degrees of freedom equal to the number of independent restrictions imposed by the null hypothesis, in this case, \( r \). Thus, the result is established.

Q.E.D.

Of central interest, therefore, is the ability to evaluate the (log) likelihood function at its maximizing value under the null and alternative hypotheses. Deriving expressions for ML estimators under these hypotheses is of secondary interest, and is critical only insofar as it leads to maximum (log) likelihood functions. Thus, estimation techniques described in the previous two chapters will be expanded to accommodate a wider variety of restrictions in the parameter space. These quantities are primarily used in the evaluation of \( \Delta \) in (5.1.6) for the appropriate test statistic.

5.1.2 Overview of Specific Hypotheses

It had been mentioned that there are two broad categories of hypotheses to be discussed in this chapter. The first supports the model building and identification function by examining the covariance structure among the repeated measures. There are some instances where a particular ARMA pattern can be specified \textit{a priori}, either because the data are known to satisfy these conditions, or as frequently occurs in econometric applications, it is an integral part of the overall model. For example, an AR(4) model is sometimes postulated for quarterly data. It is generally the case, however, that the form of the covariance matrix is not known \textit{a priori} but must be determined from an examination of the data. Indeed, one cannot be sure whether the Toeplitz structure in general is consistent with the data, much less any specific ARMA model.
Let the symbol \( \Sigma \) represent the covariance structure among the repeated measures, i.e.,

\[
\text{Var}(u_{i*}) = \Sigma
\]

(5.1.7)

where \( \Sigma \) is an \((T \times T)\) matrix, whose \((r,s)\)-th element is

\[
\sigma_{r,s} = \text{Cov}(u_i^{(r)}, u_i^{(s)}), \text{ for } r,s=1,2,\ldots,T \text{ (} i=1,2,\ldots,N \text{)}.
\]

Define the following three hypotheses which impose progressively more structure on \( \Sigma \), but where no restraints are imposed on the linear model parameter matrix \( \mathbf{B} \).

\( H_1 \): No structure is imposed on \( \Sigma \),

(5.1.8)

\( H_2 \): \( \Sigma \) observes the general Toeplitz structure,

\[
\Sigma = \mathcal{G}(\mathbf{r}),
\]

(5.1.9)

where \( \mathbf{r}' = [r_0 \ r_1 \ \ldots \ r_{T-1}] \). Thus, \( \mathcal{G} \) is a \((T \times T)\) Toeplitz matrix whose \((r,s)\)-th element is \( \gamma_{r-s} \), but where no ARMA structure is superimposed on \( \mathcal{G} \).

\( H_3 \): \( \Sigma \) follows a particular ARMA model,

\[
\Sigma = \sigma^2 \mathcal{G}(\mathbf{\phi}),
\]

(5.1.10)

where \( \mathbf{\phi}' = [\phi' \ \theta' ] \). In this case, \( \mathcal{G} \) is a \((T \times T)\) Toeplitz matrix, whose \((r,s)\)-th element is \( g_{r-s} \). This is the familiar case studied in depth in the previous two chapters.

In the model building and identification process, there are clearly two tests of interest. First, does the general Toeplitz paradigm provide an effective means of summarizing the covariance structure? A negative response would argue against the broad class of ARMA covariance patterns. Otherwise, one would want to determine which particular ARMA pattern most closely describes the covariance structure, and several competing ARMA patterns would be assessed at this time.

Given that a specific covariance model \( \Sigma = \sigma^2 \mathcal{G}(\mathbf{\phi}) \) has been identified, linear combinations of the elements of the parameter matrix \( \mathbf{B} \) taking the form:

\[
\mathbf{H}_4: \mathbf{CBD} = \mathbf{0}
\]

(5.1.11)
would be considered, where \( C \) is a known \( (c \times Q) \) matrix, of rank \( c \leq Q \), and \( D \) is a known \( (T \times d) \) matrix of rank \( d \leq T \). By manipulating the \( C \) and \( D \) matrices, various hypotheses among the explanatory variables over the time domain can be specified.

In the following sections, hypothesis testing procedures will be considered in detail. In all cases, the strategy consists of three steps.

1. Specify a method to derive the ML estimates under the null (restricted) hypothesis as a means to evaluating the log likelihood function, say \( \Lambda(\hat{\theta}_0) \), at this point.

2. Specify a method to derive the ML estimates under the alternative (unrestricted) hypothesis as a means to evaluating the log likelihood function, say \( \Lambda(\hat{\theta}_a) \), at this point.

3. Compute the likelihood ratio test statistic,

\[
\Delta = (-2)[\Lambda(\hat{\theta}_0) - \Lambda(\hat{\theta}_a)],
\]

and compare it against the percentile of the appropriate \( \chi^2 \) distribution.

Finally, a subscript will frequently be added to a parameter or function to emphasize to which hypothesis it explicitly refers.

5.2 \textbf{HYPOTHESES CONCERNING THE COVARIANCE STRUCTURE}

5.2.1 \textbf{ML Estimation Under} \( H_1 \)

Consider the hypothesis

\[ H_1: \text{No structure is imposed on} \, \Sigma, \]

Since there is no particular pattern imposed upon the covariance matrix \( \Sigma \), this is just the usual multivariate linear model, i.e., the GLM-\( \text{FR}(Y^*, XB, \Sigma) \) with normality. The intrusion of missing data recalls the incomplete longitudinal design, and a general methodology for performing ML estimation
in the presence of missing data, called the EM algorithm [Dempster, Laird & Rubin(1977)], had been described in Chapter II. It is recommended under the present circumstances.

The algorithm is an iterative procedure, with each iteration consisting of two steps. In the estimation (E) step, the conditional expectations of the (theoretical) complete-data sufficient statistics are computed, conditional upon the observed data elements and the current estimates of the parameters. The maximization (M) step then updates the ML estimators based upon the complete sufficient statistics. In the GLMM, if \( Y^* = [X^* Y^*] \) represents the augmented data array, then the complete-data sufficient statistics consist of the sums of squares and cross products matrix \( Y^* Y^* \). The E-step maximum likelihood estimator is the usual OLS expression \( \hat{\beta} = (X'X)^{-1}X'Y^* \).

The process is continued until successive estimates of the parameters converge. Dempster, Laird & Rubin(1977) showed that provided the underlying density function is a member of the exponential family of distributions, this sequence of estimators will correspond to a monotone non-decreasing sequence in the likelihood function, representing at least a local maximum for \( \hat{\beta} \) and \( \hat{\Sigma} \) in the likelihood space. Moreover, in a small Monte Carlo study, Hosking(1980) demonstrated the viability of this procedure under a variety of conditions.

In the current context, it is assumed that every element of \( \Sigma \) can be estimated from these data. This usually amounts to requiring that for all \( r, s = 1, 2, \ldots, T \), at least one pair of data elements \( (Y_i^{(r)}, Y_i^{(s)}) \) be observed in the data array. Let \( \hat{\beta}_1 \) and \( \hat{\Sigma}_1 \) represent the ML estimates derived using this technique. (The subscript "1" is used to emphasize the relationship to \( H_1 \)). Using the LMI representation, it follows that the log likelihood function, evaluated at these quantities, is computed as

\[
\Lambda_1(\hat{\beta}_1, \hat{\Sigma}_1) = c_1 - (1/2) \sum (Y_i - Z_i \hat{\beta})' \hat{\Sigma}_1^{-1} (Y_i - Z_i \hat{\beta}) - (1/2) \ln |\hat{\Sigma}_1| \quad (5.2.2)
\]

- 83 -
where $\hat{\Sigma}_i = K_i \hat{\Sigma}_i K_i'$ is the estimated covariance among the repeated measures captured for the $i$-th individual, and the constant $c_i = (-T*/2)\ln(2\pi)$.

5.2.2 ML Estimation Under $H_2$

$H_2$ had been defined as

$$H_2: \Sigma = G(\gamma), \quad (5.2.3)$$

where $\gamma' = [\gamma_0 \gamma_1 \cdots \gamma_{T-1}]$. $G$ is the general Toeplitz matrix, that is

$$G(\gamma) = \begin{bmatrix}
\gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{T-2} & \gamma_{T-1} \\
\gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{T-3} & \gamma_{T-2} \\
\vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
\gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \cdots & \gamma_0 & \gamma_1
\end{bmatrix} \quad (5.2.4)$$

It is assumed that every element of $\gamma$, and hence of $G(\gamma)$, can be estimated from the data, and this usually amounts to requiring that for all $j = 1, 2, \ldots, T-1$, at least one pair of data elements $(Y_i(t), Y_i(t+j))$ be observed in the data array. This is a considerably weaker restriction than imposed under $H_1$.

ML estimation under $H_2$ follows in a straightforward manner from the procedures described in the previous two chapters. The main results are established in the following two theorems.

Theorem 5.2.1:

Under the model and conditions considered thusfar, and under the restrictions imposed by $H_2$, the ML estimator of $\theta$ is given by

$$\hat{\theta}_2 = \left[ \Sigma_{i} \left( \hat{\Sigma}_i^{-1} \Sigma_i \right)^{-1} \right]^{-1} \left[ \Sigma_{i} \left( \hat{\Sigma}_i^{-1} \hat{\Sigma}_i \right)^{-1} \right]$$

(5.2.5)

where $\Sigma_i = K_i \hat{\Sigma}_i K_i'$, and $G = G(\gamma)$.

- 84 -
Proof:

Using the LMI representation, the log likelihood function (ignoring constant terms) is given by

$$
\Lambda_{z}(\theta, \Sigma) = -(1/2)\sum \ln |G_{1}(\Sigma)| \\
- (1/2)\sum (\Sigma_{1}-\Sigma_{1}\theta)'G_{1}^{-1}(\Sigma)(\Sigma_{1}-\Sigma_{1}\theta)
$$

(5.2.6)

The result follows in a straightforward manner, by taking the partial derivative w.r.t. $\theta$, setting this expression to zero, and solving for the parameter vector.

Q.E.D.

One could also specify the ML conditions for the Toeplitz parameter vector $\tau$. In general, they resemble expression (3.2.10), but as in the ARMA case, there is no closed-form expression for the solution to these equations, and one must therefore resort to non-linear optimization procedures to iterate to its ML estimate under $H_{2}$.

The techniques described in Chapter IV for "concentrating" the log likelihood function are followed in this case. Recalling definitions (4.2.1) - (4.2.3), the ML estimator for $\theta$ may be expressed as

$$
\hat{\theta}_{z}(\tau) = Q_{z}^{-1} \hat{\Sigma}
$$

(5.2.7)

where the subscript "2" emphasizes the reference to $H_{2}$, and all functions are functions of $\tau$ rather than $\omega$. Substituting the above expression into (5.2.6) yields the $H_{2}$-restricted concentrated log likelihood function,

$$
\Lambda_{z}^{*}(\tau) = c_{z} - (1/2)[S_{z}-E_{z}'Q_{z}^{-1}E_{z}] - (1/2)\sum \ln |G_{1}|,
$$

(5.2.8)

where the constant $c_{z}=-(T/2)\ln(2\pi)$. It is this function for which a maximizing value $\hat{\tau}_{z}$ is desired.

As in Chapter IV, it is desirable to provide an expression for the gradient vector. The main result is presented in the next theorem, but first some additional notation is required. Define the matrices,
\[ E_0 = \frac{\partial S}{\partial \gamma} = \begin{bmatrix} 1 & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & 1 \end{bmatrix} = I_T \]

\[ E_1 = \frac{\partial S}{\partial \gamma_1} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 1 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \]

\( E_1 \) is a matrix where all the elements of those diagonals one position away from the main diagonal are 1 and the remaining elements are 0. Similarly, one may define \( E_2, \ldots, E_{T-1} \), where, in general, \( E_j \) is a matrix where all the elements on the pair of diagonals, \( j \) units away from the main diagonal are 1 and the remaining elements are 0.

**Lemma 5.2.1:**

Under the model and conditions considered thusfar, and under hypothesis \( H_2 \),

\[ \xi(j) = \frac{\partial \xi}{\partial \gamma_j} = - \xi \mathbf{Z}_i' \mathbf{G}_i^{-1} \mathbf{K}_i \mathbf{E}_j \mathbf{K}_i' \mathbf{G}_i^{-1} \xi \quad (5.2.9) \]

for \( j=0,1,\ldots,T-1 \).

**Proof:**

This result follows in a straightforward manner from Lemma 3.2.2.

Q.E.D.

**Corollary:**

Under the conditions of this lemma,

\[ S(j) = \frac{\partial S}{\partial \gamma_j} = - \xi \mathbf{Z}_i' \mathbf{G}_i^{-1} \mathbf{K}_i \mathbf{E}_j \mathbf{K}_i' \mathbf{G}_i^{-1} \xi \quad (5.2.10) \]

and

\[ \Omega(j) = \frac{\partial \Omega}{\partial \gamma_j} = - \xi \mathbf{Z}_i' \mathbf{G}_i^{-1} \mathbf{K}_i \mathbf{E}_j \mathbf{K}_i' \mathbf{G}_i^{-1} \mathbf{Z}_i \quad (5.2.11) \]
Proof:

These results follow naturally from the definitions of $S$ and $Q$.

Q.E.D.

**Theorem 5.2.2:**

Under the model and conditions considered thusfar and under the hypothesis $H_2$, the derivative of the concentrated likelihood function w.r.t. $\gamma_j$ is given by:

$$q_j = \left( \frac{\partial \Lambda_2}{\partial \gamma_j} \right) =$$

$$- \frac{1}{2} \left[ S(j) - \epsilon(j)'Q^{-1}\epsilon + \epsilon(j)'Q^{-1}(j)Q^{-1}\epsilon - \epsilon(j)'Q^{-1}\epsilon(j) \right]$$

$$- \frac{1}{2} \text{tr}[VE_j],$$

(5.2.12)

for $j=0,1,\ldots,T-1$, where $V(z)=\Sigma_k G_t^{-1}(z)K_t$, and $S(j)$, $\epsilon(j)$, and $Q(j)$ had been defined in the previous lemma.

Proof:

This result follows in a straightforward manner from Theorem 4.2.1 and the previous lemma.

Q.E.D.

Considerations for performing non-linear optimization had been discussed at length in Chapter IV, and apply in full force in this setting. Starting from an initial estimate $\gamma^{(0)}$ of the Toeplitz parameter vector, the iteration sequence is pursued until the parameter vector converges to its ML estimate $\gamma$ under $H_2$. In practice, this procedure is particularly sensitive to patterns in the covariance matrix. If the data do conform to the Toeplitz model, then our experience suggests that the procedure performs well, converging within 4-5 iterations. However, it frequently fails to converge for even modest departures from the Toeplitz structure.

This situation can be ameliorated to a large extent by a judicious choice for the initial estimate of the Toeplitz
parameter vector. Since, under $H_2$, $\gamma_j$ is the autocovariance at lag $j$, an obvious choice is simply the corresponding sample autocovariance

$$c_j = \Sigma v_i(t)v_i(t+j) / \Sigma a_i(t)a_i(t+j), \quad (5.2.13)$$

for $j=0,1,...,T-1$. An initial estimator is

$$\hat{\gamma}^{(0)} = \xi, \quad (5.2.14)$$

the vector of sample autocovariances. However, in our experience, this estimator is frequently too coarse to allow the iteration sequence to converge to a solution. A better choice may be to fit a higher-order ARMA model, and to use the ensuing covariance sequence at convergence as an initial estimator for the Toeplitz vector.

Finally, the log likelihood function evaluated at the ML estimates under this hypothesis is simply the $H_2$-restricted concentrated log likelihood function (5.2.8) evaluated at $\hat{\xi}_2$. Thus,

$$\Lambda_2^*(\hat{\xi}_2) = c_2 - (1/2)\{\hat{\xi}_2 - \hat{\xi}_2 \hat{\xi}_2^{-1}\hat{\xi}_2\}$$

$$- (1/2)\Sigma \ln|G_1(\hat{\xi}_2)|, \quad (5.2.15)$$

where the constant $c_2 = (-T^2/2) \ln(2\pi)$.

5.2.3 LR Test Procedures

With respect to the covariance structure among the repeated measures, there are two hypothesis test procedures of interest. First, does the general Toeplitz structure provide an effective means of summarizing the covariance pattern? That is, a test of

$$H_2: \xi = G(\gamma)$$

versus

$$H_1: \xi \neq G(\gamma). \quad (5.2.16)$$

No restrictions are imposed upon $B$ in either case.
Following the general method, the LR test statistic is computed as

$$\Delta_{21} = (-2) [ \Lambda_2^{*} (\hat{z}_2) - \Lambda_1 (\hat{z}_1, \hat{z}_1) ]$$  \hspace{1cm} (5.2.17)$$

where $\Lambda_2^{*} (\hat{z}_2)$ had been defined in (5.2.15) and $\Lambda_1 (\hat{z}_1, \hat{z}_1)$ had been defined in (5.2.2). There are $T(T+1)/2$ parameters under $H_1$, and $T$ independent parameters under $H_2$. The number of independent restrictions imposed under $H_2$, therefore, is the difference

$$d_{21} = T(T-1)/2.$$ \hspace{1cm} (5.2.18)

An appropriate test of $H_2$ versus $H_1$ is to compare the quantity $\Delta_{21}$ against the appropriate percentile of the $\chi^2$ distribution with $T(T-1)/2$ degrees of freedom. Larger values provide greater evidence against $H_2$.

If there is no evidence forthcoming against $H_2$, then the Toeplitz paradigm may be subsumed into the overall design. In light of the sequential nature of the model building process, it is now $H_2$ which sets the context for any further specification of the covariance structure, and it becomes the appropriate alternative hypothesis. Thus, the question of whether a particular ARMA($p,q$) model effectively summarizes the covariance pattern may be considered as a test of

$$H_3: \Gamma = \sigma^2 \tilde{G}(\theta)$$

versus

$$H_2: \Gamma = \tilde{G}(\tau).$$ \hspace{1cm} (5.2.19)

where, again, no restrictions are imposed upon $\mathbb{B}$ in either case. $H_3$ is the familiar case studied in depth in the previous two chapters. Its ($H_3$-restricted) concentrated log likelihood function, $\Lambda_3^{*}(\theta)$, had been defined in (4.2.7). Thus, the LR test statistic is computed as

$$\Delta_{32} = (-2) [ \Lambda_3^{*} (\hat{\theta}_3) - \Lambda_2^{*} (\hat{\theta}_2) ]$$ \hspace{1cm} (5.2.20)
There are $T$ Toeplitz parameters under $H_2$, and $p+q+1$ parameters (including $\sigma^2$) under $H_3$, so that the number of independent restrictions imposed under $H_3$ is the difference

$$d = T - (p+q+1).$$  \hfill (5.2.21)

Thus, a test of $H_3$ versus $H_2$ is to compare the quantity $\Delta d$ against the appropriate percentile of the $\chi^2$ distribution with $T - (p+q+1)$ degrees of freedom. Larger values signify greater evidence against $H_3$.

5.3 HYPOTHESES CONCERNING THE LINEAR MODEL PARAMETERS

5.3.1 The Null Hypothesis

In this section, the general linear hypothesis of the form

$$H_0: \bar{\Xi} = CBD = 0$$  \hfill (5.3.1)

is considered, under some specific covariance model, $\Xi = \sigma^2 \C(\psi)$. $\C$ is a $(c \times q)$ matrix with full row rank, and is used, for example, to compare subpopulations of observational units or to assess the influence of concomitant variables. $D$ is a $(T \times d)$ matrix with full column rank, and is used to draw inference over the time points.

Let $\xi$ be the $(cd \times 1)$ vector formed by rolling out $\bar{\Xi}$ by rows. It follows from a result in Neudecker (1969) that if $\rho$ similarly represents $B$ rolled out by rows then

$$\xi = (\C \otimes D')\rho = L\rho$$  \hfill (5.3.2)

where $L = (\C \otimes D')$. The hypothesis may therefore be expressed as

$$H_0: \xi = L\rho = 0$$  \hfill (5.3.3)

The appropriate alternative hypothesis, $H_3$, specifies the same ARMA covariance model, but with no restrictions imposed upon the linear model parameter matrix $B$.

It is important to observe that under $H_0$, only the form of the covariance matrix is required to adhere to the spe-
cific ARMA\( (p, q) \) covariance structure. The ARMA parameter vector \( \psi \) itself is not restricted to any particular value, and is therefore free to assume whatever value maximizes the likelihood function. In large samples, this is of little consequence, since the linear model parameter estimator \( \hat{\beta} \) had been shown to be asymptotically independent of the covariance parameter estimators \( \hat{\psi} \) and \( \hat{\sigma}^2 \). In small samples, this is not necessarily the case, and in practice, it is necessary to re-estimate \( \psi \) under \( H_4 \).

The technique used to evaluate this hypothesis follows the general pattern established at the beginning of the chapter. Of immediate interest is a method to estimate the parameters \( \sigma' = [\beta' \ \gamma' \ \sigma^2] \) under the restrictions imposed by \( H_4 \).

### 5.3.2 ML Estimation Under \( H_4 \)

The objective is to maximize the (log) likelihood function under the the null hypothesis. The method of Lagrange multipliers is readily available for this purpose. Let \( \lambda \) be a \((cd \ by \ 1)\) vector of Lagrange multipliers. Then, ignoring the constant terms, the criterion function under \( H_4 \) may be expressed as

\[
\Lambda(\sigma) = -(T* / 2) \ln(\sigma^2) - (1/2) \sum \ln |G_i| \nonumber \]
\[
- (1/2\sigma^2) \sum (Y_i - Z_i \hat{\beta})' \hat{G}_i^{-1} (Y_i - Z_i \hat{\beta}) - \lambda' L \beta. \quad (5.3.4)
\]

The ML estimators are found by taking partial derivatives w.r.t. each parameter, setting these expressions to zero, and solving the resultant equations. The main result is established in the following theorem.

**Theorem 5.3.1:**

Under the model and conditions presented thusfar, and under the restrictions imposed by \( H_4 \), the ML estimators of \( \sigma^2 \) and \( \beta \) are given by

\[
\hat{\sigma}^2 = \frac{1}{T} \sum (Y_i - Z_i \hat{\beta})' \hat{G}_i^{-1} (Y_i - Z_i \hat{\beta})
\]

\[
= \frac{\hat{\psi} - \hat{\lambda}' \hat{\beta} \psi'}{T*}
\]

and

\[
- 91 -
\]
\[ \hat{\xi}_4 = [\hat{\xi}^2 - \hat{\xi}^{-1} \xi (\hat{\lambda}^2)^{-1} \hat{\xi} - \hat{\xi}^{-1}] \hat{\xi} = \hat{M} \hat{\xi}, \]

where \( S(\xi), \xi(\xi) \), and \( Q(\xi) \) had been defined in (4.2.1) - (4.2.3), and

\[ M = Q^{-1} - Q^{-1} \xi (\hat{\lambda}^2)^{-1} \hat{\xi} \] 

A hat over any of these matrices signifies that it is evaluated at the corresponding ML estimate.

**Proof:**

Taking the derivative of (5.3.4) w.r.t. \( \sigma^2 \) gives

\[ (\partial \lambda / \partial \sigma^2) = -(T^2 / 2 \sigma^2) + (1/2 \sigma^4) \Sigma (\lambda_1 - \lambda_1 \xi)^T \lambda_1^{-1} (\lambda_1 - \lambda_1 \xi) \]

(5.3.5)

and setting this expression to zero,

\[ \hat{\sigma}^2 = \Sigma (\lambda_1 - \lambda_1 \xi)^T \lambda_1^{-1} (\lambda_1 - \lambda_1 \xi) / T^2 \]

(5.3.6)

or,

\[ \hat{\sigma}^2 = \left[ S - 2 \xi \hat{\xi} + \xi \hat{\xi} \xi \right] / T^2 \]

(5.3.7)

where \( S, \xi \) and \( Q \) have been defined in (4.2.1) - (4.2.3). As well, taking the partial derivative w.r.t. \( \lambda \) and setting this expression to zero yields

\[ \Lambda^T \hat{\lambda} = 0. \]

(5.3.8)

Now,

\[ (\partial \lambda / \partial \xi) = (1 / \sigma^2) \left[ \Sigma \xi \xi^T \lambda_1^{-1} \lambda_1 - \Sigma \xi \xi^T \lambda_1^{-1} \lambda_1 \xi \right] - \lambda \lambda^T \]

so that,

\[ \hat{\xi} - \hat{\xi} \xi + \sigma^2 \lambda \lambda^T \lambda = 0. \]

(5.3.9)

Pre-multiplying each element by \( \hat{\lambda}^2 \), it follows that

\[ \hat{\lambda}^2 \hat{\lambda} - \hat{\lambda} \hat{\lambda} - \sigma^2 \hat{\lambda} \hat{\lambda} = 0. \]

Using (5.3.8), and solving for \( \lambda \) results in

\[ \hat{\lambda} = ((\hat{\lambda}^2 \hat{\lambda})^{-1} \hat{\lambda} \hat{\lambda}^{-1} \hat{\lambda}) / \sigma^2. \]

(5.3.10)

Substituting this expression into (5.3.9) and solving for \( \hat{\xi} \) gives the ML estimator for \( \xi \), i.e.,

\[ \hat{\xi} = \hat{\xi}^2 - \hat{\xi}^{-1} \xi (\hat{\lambda}^2)^{-1} \hat{\xi} - \hat{\xi}^{-1} \xi (\hat{\lambda}^2)^{-1} \hat{\xi} - \hat{\xi}^{-1} \xi (\hat{\lambda}^2)^{-1} \hat{\xi} - \hat{\xi}^{-1} \xi (\hat{\lambda}^2)^{-1} \hat{\xi}. \]

(5.3.11)
i.e.,
\[ \hat{\theta} = \hat{M} \hat{F} \]
(5.3.12)

where
\[ M = Q^{-1} - Q^{-1}L'(LQ^{-1}L')^{-1}LQ^{-1}. \]
(5.3.13)

Note that since \( M \) is a function of \( Q(\varphi) \), \( M(\varphi) \) is also a function of \( \varphi \). Moreover, since \( Q \) is a symmetric matrix, it follows that \( M \) is a symmetric matrix as well. If one writes
\[ M = Q^{-1} - P, \]
where
\[ P = Q^{-1}L'(LQ^{-1}L')^{-1}LQ^{-1}, \]
(5.3.14)

then notice that
\[ PQP = Q^{-1}L'(LQ^{-1}L')^{-1}LQ^{-1} \]
\[ Q^{-1}L'(LQ^{-1}L')^{-1}LQ^{-1} = P. \]

Thus,
\[ MOM = (Q^{-1} - P) Q (Q^{-1} - P) = Q^{-1} - P = M. \]
(5.3.15)

Substituting (5.3.12) into (5.3.7) and using (5.3.15) finally results in
\[ \hat{\sigma}^2 = \left[ \hat{S} - 2 \hat{E} \hat{M} \hat{F} + \hat{F}^T \hat{M} \hat{F} \right] / T^*, \]
i.e.,
\[ \hat{\sigma}^2 = \left[ \hat{S} - \hat{F}^T \hat{M} \hat{F} \right] / T^*. \]
(5.3.16)

Q.E.D.

Note that in light of expression (5.3.11), it is tempting to write
\[ \hat{\theta}_3 = \tilde{\theta} - \hat{Q}^{-1}L'(L\hat{Q}^{-1}L')^{-1}L\tilde{\theta} \]
where \( \tilde{\theta} = \hat{Q}^{-1} \hat{F} \) suggests the "unrestricted" (\( H_3 \)) ML estimator for \( \theta \). However, this is not correct, since in the development of (5.3.11),
\[ \tilde{\theta} = Q^{-1}(\hat{\theta}_4) \hat{F}(\hat{\theta}_4) \]
is evaluated as the ML estimator under \( H_4 \), while the unrestricted estimator,
\[ \hat{\theta}_3 = Q^{-1}(\hat{\theta}_3) \hat{F}(\hat{\theta}_3) \]
is evaluated at the ML estimator under H₃. Thus, these two quantities are based upon different estimators of \( \hat{\psi} \).

One could also specify the ML conditions for the time series parameters under H₄, and in general they resemble expression (3.2.10). As in the unrestricted case, however, there are no closed-form expressions for the solutions to these equations, and one must therefore resort to non-linear optimization procedures to iterate to the ML estimator \( \hat{\theta} \) under the null hypothesis.

The techniques described in Chapter IV for "concentrating" the log likelihood function are again followed in this situation. Substituting the ML expressions (5.3.12) and (5.3.16) into the log likelihood function yields the H₄-restricted concentrated log likelihood function, and after some algebraic manipulations, it may be expressed as

\[
\Lambda_{\hat{\theta}}(\hat{\psi}) = c_4 \left( -\frac{T^*}{2} \ln[S - \hat{\psi}' M \hat{\psi}] - \frac{1}{2} T^* \ln |G_i| \right),
\]

where \( c_4 = (-T^*/2)[1+\ln(2\pi)-\ln(T^*)] \). As in Chapter IV it is desirable to provide an expression for the gradient vector to expedite the calculations. The main result in established in the next theorem, but first, the following useful result is proven.

**Lemma 5.3.1:**

Under the model and conditions considered thusfar, and under hypothesis H₄, if \( M = Q^{-1} - Q^{-1} L'(Q^{-1} L')^{-1} Q^{-1} \), where L is functionally independent of \( \psi \), then

\[
M' = (\partial M/\partial \psi_k) = -MQ(k)M,
\]

where \( Q(k) = (\partial Q/\partial \psi_k) \) is defined in Lemma 4.2.1.

**Proof:**

\[
M' = (\partial M/\partial \psi_k) = (\partial/\partial \psi_k)[Q^{-1} - Q^{-1} L'(Q^{-1} L')^{-1} Q^{-1}]
\]

Using the usual rules of matrix differentiation,

\[
M' = -Q^{-1}Q(k)Q^{-1} + Q^{-1}Q(k)Q^{-1} L'(Q^{-1} L')^{-1} Q^{-1} - Q^{-1} L'(Q^{-1} L')^{-1} Q^{-1} Q(k) Q^{-1} L'(Q^{-1} L')^{-1} Q^{-1}
\]

- 94 -
+ \mathbf{Q}^{-1}\mathbf{L}'(\mathbf{LQ}^{-1}\mathbf{L}')^{-1}\mathbf{LQ}^{-1}\mathbf{Q}(\mathbf{k})\mathbf{Q}^{-1}.

Using the matrix \( \mathbf{P} \) defined in (5.3.14),
\[
\mathbf{M}(\mathbf{k}) = -\mathbf{Q}^{-1}\mathbf{Q}(\mathbf{k})\mathbf{Q}^{-1} + \mathbf{Q}^{-1}\mathbf{Q}(\mathbf{k})\mathbf{P} \\
- \mathbf{P}\mathbf{Q}(\mathbf{k})\mathbf{P} + \mathbf{P}\mathbf{Q}(\mathbf{k})\mathbf{Q}^{-1} \\
= - (\mathbf{Q}^{-1}\mathbf{P})\mathbf{Q}(\mathbf{k})/(\mathbf{Q}^{-1}-\mathbf{P}).
\]

Thus,
\[
\mathbf{M}(\mathbf{k}) = -\mathbf{MQ}(\mathbf{k})\mathbf{M}.
\]
(5.3.18)

Q.E.D.

**Theorem 5.3.2:**

Consider the model and conditions presented thusfar. Under
\[ \mathbf{H}_4: \mathbf{\hat{r}} = \mathbf{L}\mathbf{\hat{a}} = \mathbf{0}, \]
the derivative of the \( \mathbf{H}_4 \)-restricted concentrated log likelihood function w.r.t. \( \mathbf{\hat{r}}_k \) is given by:
\[
\mathbf{q}_k = (\partial \mathbf{A}_4^*/\partial \mathbf{r}_k) = -[T*/2(S-\mathbf{r}'\mathbf{Mr})] \\
[S(\mathbf{k}) - \mathbf{r}(\mathbf{k})'\mathbf{Mr} + \mathbf{r}'\mathbf{MQ}(\mathbf{k})\mathbf{Mr} - \mathbf{r}'\mathbf{Mr}(\mathbf{k})] \\
- (1/2)\text{tr} [\mathbf{V}(\partial \mathbf{G})/\partial \mathbf{r}_k],
\]
(5.3.19)

for \( k=1,2,\ldots,p+q \), and where \( \mathbf{V}=\mathbf{K}_1'\mathbf{G}_1^{-1}\mathbf{K}_1 \), and \( S(\mathbf{k}) \), \( \mathbf{r}(\mathbf{k}) \) and \( \mathbf{Q}(\mathbf{k}) \) had been defined in Lemma 4.2.1.

**Proof:**

The proof follows in a straightforward manner from Theorem 4.2.1 and the previous lemma.

Q.E.D.

Procedures for performing non-linear optimization had been discussed in some detail in Chapter IV, and apply in full force in this setting. The iteration sequence is pursued until the time series parameters converge to the \( \mathbf{H}_4 \)-restricted ML estimator \( \mathbf{\hat{r}}_4 \). A reasonable choice for the initial estimator \( \mathbf{\hat{r}}_4(0) \) is simply the ML estimate from the unrestricted case, \( \mathbf{\hat{r}}_3 \).

In practice, provided that the hypothesis is not "profoundly" inconsistent with the data (e.g., \( p \geq 0.005 \)), the
procedure appears to perform well, generally converging within 3-4 iterations. However, our experience suggests that for $p \leq 0.005$, the procedure may converge to inadmissible values, or fail to converge entirely, and this, in itself, is sufficient warning that the hypothesis may not be consistent with the data.

Finally, the log likelihood function, evaluated at the ML estimators under this hypothesis, is simply the $H_4$-restricted concentrated log likelihood function (5.3.17) evaluated at $\hat{\theta}_4$,

$$\Lambda^*_4(\hat{\theta}_4) = c_4 - (T^*/2)\ln[S_4 - \hat{\theta}_4^* \hat{\Lambda}_4^*] - (1/2)\sum \ln|G_1(\hat{\theta}_4)|,$$  (5.3.20)

where $c_4 = -T^*/2[1 + \ln(2\pi) - \ln(T^*)]$.

5.3.3 LR Test Procedure

Consider a test of the hypothesis

$$H_4: \text{CBD} = 0,$$

versus

$$H_3: \text{CBD} \neq 0.$$  (5.3.21)

assuming a particular ARMA covariance pattern $\Sigma = \sigma^2 G(\phi)$ in either case. Following the general procedure, the LR test statistic is computed as

$$\Delta_{43} = (-2)[\Lambda^*_4(\hat{\theta}_4) - \Lambda^*_3(\hat{\theta}_3)]$$  (5.3.22)

where $\Lambda^*_4(\hat{\theta}_4)$ had been defined in (5.3.20), while $\Lambda^*_3(\hat{\theta}_3)$ had been defined in (4.2.7). There are

$$d_{43} = cd$$  (5.3.23)

independent restrictions imposed on the linear model parameters under $H_4$. Thus, a test of $H_4$ versus $H_3$ is to compare the quantity $\Delta_{43}$ against the appropriate percentile of the $\chi^2$ distribution with $cd$ degrees of freedom. Larger values suggest greater evidence against $H_4$. 

- 96 -
Chapter VI
EXAMPLES

6.1 INTRODUCTION

In the previous chapters, a methodology had been presented for drawing inferences from the incomplete longitudinal design under an ARMA covariance model. Expressions for the ML estimators of the parameters had been derived, and a discussion was included concerning the mechanics of performing these calculations using non-linear optimization techniques. In addition, procedures for testing various hypotheses in support of the model building and identification process, as well as for the more substantive hypotheses among the linear model parameters, were described. This methodology therefore presents a unified inference strategy.

In this chapter, examples will be presented to illustrate this methodology. Two datasets will be studied in detail. The first consists of data generated at random from a known stochastic mechanism; the second is derived from a dataset published in the statistical literature. If the underlying structures of these datasets can be verified, then this would represent a vote of confidence in these techniques.

Moreover, this chapter also serves a useful pedagogical purpose. Model building is very frequently a sequential process, with a decision to add more structure to the model contingent upon a satisfactory fit by the current model. The sequential nature of this process will be highlighted throughout these examples.
6.1.1 Consistency of the Stochastic Mechanism

It has been implicitly assumed that the underlying stochastic mechanism is consistently observed across all the time points under consideration. This is manifested primarily by the Toeplitz structure in the covariance matrix $\Sigma$ among the repeated measures. In particular, the elements along any diagonal of $\Sigma$ must all be the same, and a formal test of significance for this assertion had been presented in the previous chapter.

It would nevertheless be useful to have a diagnostic procedure readily available to address this issue. An obvious approach is to examine the autocovariance structure among the repeated measures, and specifically the contribution made by each time point to this overall measure. One would expect that if the stochastic mechanism was consistently observed, then these time-specific autocovariances would be approximately the same across all time points. Time-specific autocorrelations, however, are a better choice since they are bounded by 1.

Accordingly, define the autocovariance at time $t$ and lag $j$ as

$$\gamma_j(t) = E(v_i(t)v_i(t+j)). \quad (6.1.1)$$

Using the notation established in Section 4.3, this quantity can estimated by

$$c_j(t) = \frac{\varepsilon v_i(t) v_i(t+j)}{\varepsilon a_i(t) a_i(t+j)} \quad (6.1.2)$$

for $j=0,1,2,\ldots,T-1$ and $t=1,2,\ldots,T-j$. The overall sample autocovariances (4.3.6) are seen to be weighted averages of these time-specific quantities. The corresponding time-specific sample autocorrelations are defined as

$$r_j(t) = c_j(t) / c_0(t). \quad (6.1.3)$$

An expression for the approximate variance of any sample autocorrelation was derived in Bartlett (1946). Box & Jenkins (1970), Sec. 2.1] recommended that, in support of the model identification process, the quantity
\[ \text{Var}(r_j) = \frac{[1 + 2\sigma_{r_k}^2]}{n} \]

be used, with the summation extending from \( k=1 \) to \( j-1 \). The
denominator \( n \) represents the number of data elements which
constitute the relevant time series. For the overall sample
autocorrelations (4.3.7), this is simply the total number of
observed data elements \( T = \sum a_i(t) \), so that

\[ \text{Var}(r_j) = \frac{[1 + 2\sigma_{r_k}^2]}{\sum a_i(t)}. \quad (6.1.4) \]

For the time-specific autocorrelations, since only one pair
of residuals \( v_i(t)v_i(t+j) \) is contributed to this measure by
each observational unit, the denominator is the number of
observed data elements specific to that time point, \( i_a(t) \).
Hence,

\[ \text{Var}[r_j(t)] = \frac{[1 + 2\sigma_{r_k}^2(t)]}{i_a(t)}. \quad (6.1.5) \]

for \( j=1,2,...,T-1 \) and \( t=1,2,...,T-j \).

Thus, a useful diagnostic device for assessing the con-
sistency of the underlying stochastic mechanism is to pro-
cede as follows. For each lag \( j=1,2,...,T-1 \), one could plot
the time-specific autocorrelations for \( t=1,2,...,T-j \), plus
and minus two (say) standard errors. If for any two time
points, the corresponding intervals overlap, then this would
suggest that the \( j \)-th order autocorrelation is consistent at
those time points. All such pairwise comparisons could be
made. Alternatively, one could generate the corresponding
interval for the overall sample autocorrelation at lag \( j \).
If each of the time-specific intervals intersects this over-
all interval, then this would provide evidence in support of
the consistency of the stochastic mechanism. One might want
to make allowances for the number of comparisons being made.
An example of these plots will be provided in the next sec-
tion.
6.2 Example 1
6.2.1 Description of the Dataset

The first dataset consisted of data generated at random from a known stochastic structure. A data matrix, \( Y^* \), of size \( N=80 \) and \( T=7 \) was formed, and \( Y^* \) may be considered as consisting of two components, namely a deterministic component \( XB \), and a stochastic component \( U^* \), such that

\[
Y^* = XB + U^*.
\]

The deterministic component was generated using

\[
X = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\]

(6.2.1)

i.e., a cell-mean model, where 1 represents a (40 by 1) column of 1's, and 0 is a (40 by 1) column of 0's. The linear model parameter matrix \( B \) was defined to be

\[
B = \begin{bmatrix}
10 & 15 & 20 & 25 & 30 & 35 & 40 \\
10 & 20 & 30 & 40 & 50 & 60 & 70
\end{bmatrix}
\]

(6.2.2)

Thus, the deterministic component reflects the existence of two subpopulations, each with 40 observational units. A linear trend in the cell means over the 7 time points is evident for both subpopulations; but while the cell means are equal at Time 1, they diverge thereafter.

For each observational unit, disturbance terms were generated using an ARMA(1,1) covariance model, with parameters,

\[
\phi = 0.75,
\]

(6.2.3)

\[
\theta = -0.50,
\]

(6.2.4)

and, \( \sigma^2 = 1 \).

(6.2.5)

The covariance sequence \( \gamma_0, \gamma_1, \ldots, \gamma_{T-1} \) corresponding to this ARMA model is given by

\[
\gamma' = [4.57 \ 3.93 \ 2.95 \ 2.21 \ 1.66 \ 1.24 \ 0.93],
\]

(6.2.6)

while the covariance matrix \( \Sigma \) among the repeated measures is that (7 by 7) Toeplitz matrix function of this vector, \( G(\gamma) \).
Since \( G \) is a positive-definite matrix, from the Cholesky decomposition, there exists a lower-triangular matrix \( F \) such that

\[
G = FF'.
\] (6.2.7)

Independent and identically distributed \( N(0,1) \) random variates \( \epsilon_i^{(t)} \) were generated using the SAS random number generating function RANNOR [SAS Institute(1982a)] for \( t=1,2,\ldots,T \) and \( i=1,2,\ldots,N \). RANNOR uses a prime modulus multiplicative generator as described in Fishman & Moore(1982). The random vector \( \mathbf{\xi}_i = [\epsilon_i^{(1)} \epsilon_i^{(2)} \ldots \epsilon_i^{(T)}] \) therefore follows the MVN(0,\( \sigma^2 I \)), and

\[
\mathbf{u}_i = F\mathbf{\xi}_i
\] (6.2.8)

is distributed as a MVN(0,\( \sigma^2 G \)) for \( i=1,2,\ldots,N \). Thus, taking \( \mathbf{U}^* \) to be that matrix whose \( i \)-th row is \( \mathbf{u}_i \), one arrives at the complete data array,

\[
\mathbf{y}^* = \mathbf{Xb} + \mathbf{U}^*.
\] (6.2.9)

as required.

Finally, elements of \( \mathbf{y}^* \) were deleted at random using the SAS random number generating function RANUNI [SAS Institute(1982a)]. A target of one-third missing data elements was set, although only 176 of the 560 data elements (31.43%) were actually deleted. The frequency of missing values by Group and Time is presented in Table 6.2.1. A complete listing of the dataset is reported in Appendix A.
<table>
<thead>
<tr>
<th>Time</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1</td>
<td>15</td>
<td>18</td>
<td>13</td>
<td>9</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>86</td>
</tr>
<tr>
<td>Group2</td>
<td>17</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>9</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>TOTAL</td>
<td>32</td>
<td>29</td>
<td>25</td>
<td>24</td>
<td>28</td>
<td>18</td>
<td>20</td>
<td>176</td>
</tr>
</tbody>
</table>

**TABLE 6.2.1**
Frequency of kissing values by Group and Time

6.2.2 *Model Building and Identification*

In the first instance, the diagnostic procedures discussed in the previous section were applied to the data in an effort to examine the underlying stochastic structures. Table 6.2.2 details the time-specific, and overall, sample autocorrelations, together with their approximate standard errors. No substantial differences were observed between the time-specific values and their overall sample counterparts. The value of \( \hat{r}_3(1) = 0.0567 \) (s.e. = 0.2451) for the lag 3 autocorrelation at Time 1 appears to be rather small compared to the overall value of \( \hat{r}_3 = 0.4101 \) (s.e. = 0.0886); however, in light of its large standard error, it cannot be considered as "significantly" different.

A plot of the time-specific first-order sample autocorrelations (plus and minus two standard errors) is presented in Figure 6.2.1. A reference line is drawn at the overall first-order value, \( \hat{r}_1 = 0.8178 \). Clearly, none of the time-specific values deviated substantially from this quantity. Thus, these observations suggest that the underlying sto-
Figure 6.2.1: Plot of the Time-Specific First-Order Sample Autocorrelations, Plus and Minus Two Standard Errors


<table>
<thead>
<tr>
<th>TIME1</th>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
<th>LAG 4</th>
<th>LAG 5</th>
<th>LAG 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7110</td>
<td>0.6604</td>
<td>0.0567</td>
<td>0.3487</td>
<td>0.2782</td>
<td>0.1607</td>
</tr>
<tr>
<td></td>
<td>(0.1443)</td>
<td>(0.2047)</td>
<td>(0.2451)</td>
<td>(0.2453)</td>
<td>(0.2550)</td>
<td>(0.2617)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TIME2</th>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
<th>LAG 4</th>
<th>LAG 5</th>
<th>LAG 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9727</td>
<td>0.5938</td>
<td>0.4345</td>
<td>0.4139</td>
<td>0.1829</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1400)</td>
<td>(0.2381)</td>
<td>(0.2656)</td>
<td>(0.2792)</td>
<td>(0.2909)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TIME3</th>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
<th>LAG 4</th>
<th>LAG 5</th>
<th>LAG 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7504</td>
<td>0.6158</td>
<td>0.4939</td>
<td>0.4042</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1348)</td>
<td>(0.1966)</td>
<td>(0.2290)</td>
<td>(0.2476)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TIME4</th>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
<th>LAG 4</th>
<th>LAG 5</th>
<th>LAG 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8329</td>
<td>0.5790</td>
<td>0.5205</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1336)</td>
<td>(0.2064)</td>
<td>(0.2336)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TIME5</th>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
<th>LAG 4</th>
<th>LAG 5</th>
<th>LAG 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7762</td>
<td>0.6683</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1387)</td>
<td>(0.2059)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TIME6</th>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
<th>LAG 4</th>
<th>LAG 5</th>
<th>LAG 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8537</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1270)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OVERALL</th>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
<th>LAG 4</th>
<th>LAG 5</th>
<th>LAG 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8176</td>
<td>0.5817</td>
<td>0.4101</td>
<td>0.3553</td>
<td>0.1628</td>
<td>0.0882</td>
</tr>
<tr>
<td></td>
<td>(0.0511)</td>
<td>(0.0780)</td>
<td>(0.0886)</td>
<td>(0.0934)</td>
<td>(0.0968)</td>
<td>(0.0976)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POPULATION</th>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
<th>LAG 4</th>
<th>LAG 5</th>
<th>LAG 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.8594</td>
<td>0.6445</td>
<td>0.4834</td>
<td>0.3626</td>
<td>0.2719</td>
<td>0.2038</td>
</tr>
</tbody>
</table>

TABLE 6.2.2

Time-Specific and Overall Sample Autocorrelations, with Approximate Standard Errors in Parentheses, Compared Against Theoretical Population Values

The stochastic mechanism is consistent (as expected) across the time points.

In the most general case ($H_1$), no particular structure was imposed upon the covariance matrix $\Gamma$ among the repeated measures, and ML estimation was performed using the EM algorithm. Orchard & Woodbury (1972) had observed that this methodology may converge rather slowly, and this was found to be the case in this example. Convergence was achieved
only after 20 iterations, the value of the log likelihood function at convergence was computed to be -551.90. The ML estimate for \( \Sigma \) is presented in Table 6.2.3, and while there appears to be some variability along the diagonals, casual inspection suggests that a Toeplitz structure is a plausible covariance model for these data.

\[
\begin{array}{cccccccc}
\text{TIME1} & \text{TIME2} & \text{TIME3} & \text{TIME4} & \text{TIME5} & \text{TIME6} & \text{TIME7} \\
3.1707 & 3.3009 & 2.6352 & 2.0592 & 1.4538 & 0.9093 & 0.5756 \\
(4.57) & (3.93) & (2.95) & (2.21) & (1.66) & (1.24) & (0.93) \\
3.9414 & 3.4990 & 2.9468 & 2.2038 & 1.5430 & 1.1817 & 1.1817 \\
(4.57) & (3.93) & (2.95) & (2.21) & (1.66) & (1.24) & (1.24) \\
3.9730 & 3.7504 & 2.9587 & 2.3075 & 1.9640 & 1.9640 & 1.9640 \\
(4.57) & (3.93) & (2.95) & (2.21) & (1.66) & (1.24) & (1.24) \\
4.4477 & 3.9469 & 3.1181 & 2.4348 & 2.4348 & 2.4348 & 2.4348 \\
(4.57) & (3.93) & (2.95) & (2.21) & (1.66) & (1.24) & (1.24) \\
(4.57) & (3.93) & (2.95) & (2.21) & (1.66) & (1.24) & (1.24) \\
(\text{symmetric}) & & & & & & \\
(4.57) & (3.93) & (2.95) & (2.21) & (1.66) & (1.24) & (1.24) \\
3.5225 & & & & & & \\
(4.57) & & & & & & \\
\end{array}
\]

**TABLE 6.2.3**

ML Estimate of \( \Sigma \) Under the General Covariance Model, with the Corresponding Population Values in Parentheses

Techniques described in Section 5.2 were used to perform ML estimation under the general Toeplitz paradigm (\( H_2 \)).
maximum likelihood estimate of the Toeplitz parameter vector at convergence was computed as

$$\hat{\theta}' = [3.69 \ 3.31 \ 2.64 \ 2.04 \ 1.55 \ 1.05 \ 0.44] . \ (6.2.10)$$

Compared to the theoretical structure (6.2.6), the ML estimates were consistently 10-20% lower, although the same basic decreasing structure was maintained.

The value of the log likelihood function at convergence was calculated to be -561.64. Thus, a large-sample test of significance for the appropriateness of the Toeplitz model was computed as

$$X^2 = (-2)[-561.64 - (-551.90)] = 19.48,$$

with 7(6)/2=21 degrees of freedom, $p=0.55$. Thus, the methodology does not reject the hypothesis of a Toeplitz pattern in the covariance matrix.

Given this result, efforts were next directed towards determining which particular ARMA model provided a parsimonious description of the covariance structure ($H_3$). Several competing models were presented for consideration, and for each model, the ML estimates of $\psi$ and $\sigma^2$, together with the corresponding covariance sequence, are reported in Table 6.2.4. Additionally, a test of significance for the adequacy of each model was performed against the alternative of the general Toeplitz pattern, and the results of these procedures are reported in Table 6.2.5.

Clearly, a time series model with at least two parameters is required for these data. Both the AR(2) model ($p=0.26$) and the ARMA(1,1) model ($p=0.31$) provided an adequate fit, with the latter being marginally preferred. The three-parameter models all provided a considerably better fit ($0.50 \leq p \leq 0.55$), but at the expense of an extra parameter. Thus, the methodology confirms that the ARMA(1,1) offers a plausible vehicle to model the covariance structure, and if parsimony were the overriding criterion, one would be justified in adopting either the AR(2) or the ARMA(1,1) time series model.
<table>
<thead>
<tr>
<th>ML ESTIMATES</th>
<th>CORRESPONDING COVARIANCE SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lag 0</td>
</tr>
<tr>
<td>AR(1):</td>
<td></td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>0.8847</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>0.7987</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td></td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td></td>
</tr>
<tr>
<td>AR(2):</td>
<td></td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>1.2839</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>0.4309</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td></td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td></td>
</tr>
<tr>
<td>ARMA(1,1):</td>
<td></td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>0.8137</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>-0.5813</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td></td>
</tr>
<tr>
<td>ARMA(1,2):</td>
<td></td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>0.7753</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>-0.6062</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td></td>
</tr>
<tr>
<td>ARMA(2,1):</td>
<td></td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>1.0540</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>-0.2293</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td></td>
</tr>
<tr>
<td>POPULATION:</td>
<td></td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>0.75</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>-0.50</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 6.2.4
ML Estimates of $\phi$ and $\sigma^2$ for Various ARMA Models, with the Corresponding Covariance Sequence, Compared Against the Theoretical Population Values
TABLE 6.2.5
Results of Hypothesis Testing for Various ARMA Models

<table>
<thead>
<tr>
<th>MODEL</th>
<th>$\Lambda_3(\hat{\cdot})$</th>
<th>$\chi^2$</th>
<th>D.F.</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>-586.22</td>
<td>49.17</td>
<td>5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>AR(2)</td>
<td>-564.29</td>
<td>5.30</td>
<td>4</td>
<td>0.26</td>
</tr>
<tr>
<td>ARMA(1,1)</td>
<td>-564.03</td>
<td>4.78</td>
<td>4</td>
<td>0.31</td>
</tr>
<tr>
<td>ARMA(1,2)</td>
<td>-562.70</td>
<td>2.12</td>
<td>3</td>
<td>0.55</td>
</tr>
<tr>
<td>ARMA(2,1)</td>
<td>-562.77</td>
<td>2.26</td>
<td>3</td>
<td>0.52</td>
</tr>
</tbody>
</table>

6.2.3 ML Estimation and Hypothesis Testing

Maximum likelihood estimation under the ARMA(1,1) covariance model was performed using the techniques described in Chapters III and IV. Convergence was achieved in 6 iterations. The ML estimates for the covariance parameters were computed as

$$\hat{\phi} = 0.8137,$$  \hfill (6.2.11)

$$\hat{\theta} = -0.5613,$$  \hfill (6.2.12)

and, $$\hat{\sigma}^2 = 0.5602.$$  \hfill (6.2.13)

The ARMA parameter estimates compare favourably with their theoretical counterparts, $\phi = 0.75$ and $\theta = -0.50$, while $\hat{\sigma}^2$ appears to underestimate its theoretical value of 1. The ML estimate of $B$, with the corresponding asymptotic standard errors, is presented in Table 6.2.6.

For the most part, the ML estimates compare favourably with their theoretical counterparts. With the exception of the Group 2 cell mean at Time 1, the theoretical values of (6.2.2) all lie within two standard errors of their ML estimates. A plot of the estimated cell means over time for the two subpopulations is presented in Figure 6.2.2.
TABLE 6.2.6
ML Estimate of the Linear Model Parameter Matrix B, with the Asymptotic Standard Errors in Parentheses

| GROUP1 | 10.03 (0.33) | 15.07 (0.32) | 19.80 (0.31) | 24.85 (0.31) | 29.93 (0.31) | 34.99 (0.31) | 39.82 (0.32) |
| GROUP2 | 9.23 (0.33)  | 19.45 (0.31) | 29.56 (0.31) | 39.74 (0.31) | 49.72 (0.31) | 59.75 (0.31) | 69.63 (0.32) |

TABLE 6.2.7
Results of Hypothesis Testing Among the Linear Model Parameters

<table>
<thead>
<tr>
<th>LINEAR HYPOTHESIS</th>
<th>( \Lambda_x(\hat{\theta}) )</th>
<th>( \chi^2 )</th>
<th>D.F.</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Main Effect</td>
<td>-1383.38</td>
<td>1638.70</td>
<td>12</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Linear Component</td>
<td>-1366.16</td>
<td>1604.27</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Quadratic Component</td>
<td>-564.83</td>
<td>1.61</td>
<td>2</td>
<td>0.45</td>
</tr>
<tr>
<td>Equal Linear Components</td>
<td>-956.08</td>
<td>784.09</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Equality at Time 1</td>
<td>-565.48</td>
<td>2.90</td>
<td>1</td>
<td>0.09</td>
</tr>
<tr>
<td>Equality at Time 2</td>
<td>-600.37</td>
<td>72.66</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Finally, hypothesis testing \( (H_0) \) was performed to verify those structures incorporated in the linear model parameter matrix. In all cases, the ARMA(1,1) model was used, and hypothesis testing was performed against the alternative of this covariance pattern, but with no restrictions in the linear model parameter space. These procedures are summarized in Table 6.2.7.

The methodology confirms that significant differences were observed across the time points \( (p<0.0001) \), and suggests that the trend may be summarized by a linear \( (p<0.0001) \) but not a quadratic \( (p=0.45) \) function over time. A significant difference was recorded at the second time point \( (p<0.0001) \), although no difference was observed at Time 1 \( (p=0.09) \) as expected.

The test for equality at Time 1 provided more evidence against the null hypothesis than was anticipated. The estimated cell mean for Group 2 at Time 1 was computed as 9.23, more than two standard errors below the theoretical value of 10. This may well reflect normal statistical variability in the random number generator, or the influence of the large number of data elements deleted at the first time point. Nevertheless, no significant difference was observed between the two cell means at Time 1 as mandated by the underlying structure.
Figure 6.2.2: Plot of Estimated Cell Means Over Time for the Two Subpopulations
6.2.4 Comparison Against Univariate Estimation Procedures

An alternative inference strategy is to perform a univariate analysis on those data elements observed at each time point, serially across the T repeated measures. Computationally, this would entail a considerably less demanding procedure by several orders of magnitude, and statistically, one would not be bound by an ARMA Toeplitz pattern in the covariance matrix among the repeated measures. However, this approach ignores any intercorrelation structure among the repeated measures, and there may be some sacrifice in statistical power. In this section, a comparison is made between the ML estimation procedure, using the ARMA(1,1) covariance structure as reported above, and this piecemeal univariate approach.

Ordinary least squares (OLS) estimates of the linear model parameters from the piecemeal univariate analyses are reported in Table 6.2.8, together with the computed standard errors. Compared to Table 6.2.6, there does not appear to be a great deal of difference in the estimates arising from the two approaches. However, with two exceptions, the univariate estimates are biased further away from the known population values than under the ARMA multivariate approach. Moreover, the Group 2 cell means both at Time 1 and Time 2 are now more than two standard errors away from their population values. The estimated standard errors corresponding to these cell means are generally smaller using the ARMA multivariate procedures. An exception occurs for both cell means at Time 1.

The OLS estimate of the disturbance variance$^1$ for each of the 7 time points was computed to be:

\[
\begin{array}{cccccccc}
\text{TIME:} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hat{\sigma}^2: & 1.99 & 3.39 & 3.96 & 3.96 & 3.55 & 4.09 & 3.95 \\
\end{array}
\]

\[(6.2.14)\]

$^1$ The OLS estimator, i.e., weighted by \((N-2)^{-1}\), is reported rather than the ML estimator since it is the former which would be used to draw inference in a univariate analysis.
<table>
<thead>
<tr>
<th>TIME1</th>
<th>TIME2</th>
<th>TIME3</th>
<th>TIME4</th>
<th>TIME5</th>
<th>TIME6</th>
<th>TIME7</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.65</td>
<td>14.60</td>
<td>19.33</td>
<td>24.40</td>
<td>29.60</td>
<td>34.88</td>
<td>39.82</td>
</tr>
<tr>
<td>(0.28)</td>
<td>(0.39)</td>
<td>(0.38)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
<td>(0.36)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GROUP2</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8.96</td>
<td>19.06</td>
<td>29.62</td>
<td>39.26</td>
<td>49.35</td>
<td>59.46</td>
<td>69.94</td>
</tr>
<tr>
<td>(0.29)</td>
<td>(0.34)</td>
<td>(0.38)</td>
<td>(0.40)</td>
<td>(0.38)</td>
<td>(0.36)</td>
<td>(0.36)</td>
</tr>
</tbody>
</table>

**TABLE 6.2.8**

OLS Estimates of the Linear Model Parameters From the Piecemeal Univariate Analyses, with the Corresponding Standard Errors in Parentheses

Under the Toeplitz paradigm, the disturbance variance is held to be constant across all 7 time points, and from Table 6.2.4, the comparative value is $\hat{\gamma}_0=3.69$ under the ARMA(1,1) model. Four of the univariate variances are higher, while three are smaller.

On the basis of this very limited comparison, the ARMA methodology gives a good accounting of its statistical properties. Estimates of the linear model parameters appear to be somewhat less biased and more precise using the ARMA estimation techniques, and the ARMA estimate of the disturbance variances is no worse than those derived from the univariate procedures. This agrees in broad terms with larger simulation studies by Hosking (1980) and Kmenta & Gilbert (1970), for example.

However, a significant by-product of the ARMA multivariate approach is an estimate of the intercorrelation structures among the repeated measures. It provides for a fuller understanding of the temporal relationships among the dependent variables, and therefore forms an integral component of any inference strategy. This, it is felt, redresses to a large extent the heavy computational burden.
6.3 EXAMPLE 2

6.3.1 Description of the Dataset

The second dataset was drawn from a paper published in the statistical literature. Zerbe (1979) described a methodology for comparing growth curves across several subpopulations. The parametric assumptions underpinning standard procedures [e.g., Grizzle & Allen (1969)] were criticized as being untenable in many situations likely to arise in practice, and a non-parametric alternative was proposed. Best linear unbiased estimators for the mean growth curves were presented. Moreover, a randomization test, together with an approximating F-distribution, was described for testing overall equality, or equality restricted to a certain time point or a certain interval of time.

Data were presented [Zerbe's Table 2] to illustrate these procedures, and the association between hyperglycemia and relative hyperinsulinemia was investigated. Standard glucose tolerance tests were administered to 13 Control and 20 Obese individuals. The response measure of interest was the plasma inorganic phosphate level, and these were determined from blood samples withdrawn at 0, 0.5, 1.0, 1.5, 2, 3, 4, and 5 hours after an oral glucose challenge.

The role of the observation at "Hour 0" is ambiguous from a reading of the paper. It appears, however, to be more an expression of inherent differences existing among the individuals than a measure of the response to the test procedures. Thus, for the purpose of the current analysis, the Hour-0 reading will be considered as a concomitant variable, to account for individual differences and to control for biases between the two subpopulations. This marks a point of departure from the analysis reported in Zerbe's paper, but as will be seen, the overall inferences remain unaffected.

Adjusted cell means (adjusted to the Hour-0 mean value) across the 7 time points and for both treatment groups are presented in Figure 6.3.1. Mean plasma phosphate levels de-
Figure 6.3.1: Plot of the Adjusted Cell Means Over Time for the Control and Obese Subpopulations
cline rapidly during the first two hours, but then appear to reverse and to increase thereafter. The mean response is lowest after 1.5 hours for the Control group, and at the 2-hour point for Obese individuals. This suggests a duality in the biological response to a glucose challenge, with the 2-hour mark representing the point of departure.

<table>
<thead>
<tr>
<th>HOUR</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1450</td>
<td>0.0100</td>
<td>-0.0143</td>
<td>0.0115</td>
<td>0.0008</td>
<td>0.0576</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>0.1561</td>
<td>0.0742</td>
<td></td>
<td>0.0779</td>
<td>0.0168</td>
<td>0.0348</td>
<td>0.0314</td>
</tr>
<tr>
<td></td>
<td>0.1650</td>
<td></td>
<td></td>
<td>0.1396</td>
<td>0.1106</td>
<td>0.0568</td>
<td>0.0355</td>
</tr>
</tbody>
</table>

(symmetric)

|      | 0.3526 | 0.1996 | 0.1360 | 0.0940 |
|      |       | 0.3011 | 0.2111 | 0.1541 |
|      |       |       | 0.2659 | 0.1929 |
|      |       |       |       | 0.2137 |

TABLE 6.3.1
ML Estimate of Σ under the General Covariance Model

This conjecture is reinforced by an examination of the covariance structures among the repeated measures. The ML estimate of Σ (i.e., weighted by N⁻¹) under the general covariance model (H₁), adjusted for Hour-0 differences and pooled across the two subpopulations, is reproduced in Table 6.3.1. Considering elements along the main diagonal, the estimated variances of readings taken at 0.5, 1.0, and 1.5 hours after the procedure appear to cluster around 0.15.
The corresponding values at 2, 3, 4, and 5 hours appear to cluster around the 0.30 level — nearly twice as large.

Table 6.3.2a presents the sample autocorrelations, together with their approximate standard errors, specific to the equally-spaced observations at times 0.5, 1.0 and 1.5 hours after the procedure. The first-order sample autocorrelations are observed to vary from a value of \( \hat{r}_1(1) = 0.0687 \) (s.e. = 0.1741) to a value of \( \hat{r}_1(3) = 0.8456 \) (s.e. = 0.1741) — a difference which may be considered as significantly different. The second-order sample autocorrelations also vary widely, from a value of \( \hat{r}_2(1) = -0.0985 \) (s.e. = 0.1749) to a value of \( \hat{r}_2(2) = 0.4987 \) (s.e. = 0.2098), although by virtue of the large standard errors, this difference was not considered significant. The corresponding sample autocorrelations specific to the observations at times 2, 3, and 4 hours after the procedure are detailed in Table 6.3.2b. [Note that the entries in Table 6.3.2a are not comparable with those described in Table 6.3.2b, since the former describe the correlation structure among observations separated by half-hour intervals, while the latter are specific to observations separated by one-hour intervals.] By and large, the autocorrelation structure appears to have stabilized by the second hour. These first-order sample autocorrelations vary from a value of \( \hat{r}_1(1) = 0.5660 \) (s.e. = 0.1741) to a value of \( \hat{r}_1(3) = 0.7253 \) (s.e. = 0.1741).

Thus, the weight of evidence forthcoming from an examination of the sample first- and second-order moments strongly points to a duality in the biological response to an oral glucose challenge with the Hour-2 reading representing a watershed mark. The initial decline in the mean response is reversed for both treatment groups by the second hour, and the autocorrelations achieve some measure of stability only at that point. Thus, for the purpose of the current illustration, only those four repeated measures observed at the equally-spaced observation times 2, 3, 4, and 5 hours after the test procedure will be considered.
<table>
<thead>
<tr>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOUR 0.5</td>
<td>0.0687</td>
<td>-0.0985</td>
</tr>
<tr>
<td></td>
<td>(0.1741)</td>
<td>(0.1749)</td>
</tr>
<tr>
<td>HOUR 1.0</td>
<td>0.4755</td>
<td>0.4987</td>
</tr>
<tr>
<td></td>
<td>(0.1741)</td>
<td>(0.2098)</td>
</tr>
<tr>
<td>HOUR 1.5</td>
<td>0.8456</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1741)</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 6.3.2A**

Sample Autocorrelations Specific to Observations Taken at Hours 0.5, 1.0, 1.5, with Approximate Standard Errors in Parentheses.

<table>
<thead>
<tr>
<th>LAG 1</th>
<th>LAG 2</th>
<th>LAG 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOUR 2</td>
<td>0.3660</td>
<td>0.3857</td>
</tr>
<tr>
<td></td>
<td>(0.1741)</td>
<td>(0.2230)</td>
</tr>
<tr>
<td>HOUR 3</td>
<td>0.7011</td>
<td>0.5119</td>
</tr>
<tr>
<td></td>
<td>(0.1741)</td>
<td>(0.2451)</td>
</tr>
<tr>
<td>HOUR 4</td>
<td>0.7253</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.1741)</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 6.3.2B**

Sample Autocorrelations Specific to Observations Taken at Hours 2, 3, 4, with Approximate Standard Errors in Parentheses.
6.3.2 Data Analysis

The dataset for this analysis includes \( N = 33 \) subjects and \( T = 4 \) time points; there are two subpopulations and one concomitant variable. There are no missing values evident in the data array; however, with one exception, this makes no difference insofar as the estimation procedures are concerned. Estimation remains an iterative process under \( H_2, H_3, \) and \( H_4 \). The only difference is that the incidence matrices \( K_i = I_T \) for all \( i = 1, 2, \ldots, N \). ML estimation under \( H_4 \), of course, is a considerably easier proposition with complete data. The usual multivariate analysis procedures [e.g., Timm(1975)] are substituted for the EM algorithm. The maximum likelihood estimator of \( \Sigma \) (i.e., weighted by \( N^{-1} \)) is computed rather than the usual unbiased estimator (i.e., weighted by \( (N-Q)^{-1} \)) in order to be consistent with estimation under the other hypotheses.

In the model building stage, an attempt was made to identify which covariance model provided the most parsimonious description of this structure. The general Toeplitz model was first scrutinized for consistency with the data, and contingent upon an affirmative response, various competing ARMA models were then presented for consideration. The results of these procedures are reported in Table 6.3.3. Moreover, ML estimates of \( \psi \) and \( \sigma^2 \), together with the resulting covariance sequence, for each ARMA model under consideration are reported in Table 6.3.4.

The analysis suggests that the Toeplitz covariance structure is only marginally acceptable for these data (\( p = 0.12 \)). Ordinarily, one would prefer a significance level exceeding 0.25 in a model fitting exercise. An examination of the ML estimate of the general covariance matrix \( E \) among the repeated measures (Table 6.3.1) suggests a reason for the relatively poor fit. While the elements along each of the minor diagonals appear to be relatively constant as required, the variances observed along the main diagonal vary from 0.21 to 0.35. This is evidently sufficient in this example
COVARIANCE MODEL

<table>
<thead>
<tr>
<th></th>
<th>$x^2$</th>
<th>D.F.</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toeplitz vs.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Matrix</td>
<td>10.2210</td>
<td>6</td>
<td>0.12</td>
</tr>
<tr>
<td>AR(1) vs. Toeplitz</td>
<td>0.0521</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td>AR(2) vs. Toeplitz</td>
<td>0.0468</td>
<td>1</td>
<td>0.83</td>
</tr>
<tr>
<td>ARMA(1,1) vs. Toeplitz</td>
<td>0.0471</td>
<td>1</td>
<td>0.83</td>
</tr>
</tbody>
</table>

**TABLE 6.3.3**

Results of Hypothesis Testing on the Covariance Structure

to weaken the Toeplitz assumption. Nevertheless, there is no "significant" evidence against the Toeplitz assumption for the last four repeated measures, and one may proceed to the next analysis stage.

Of the competing ARMA patterns, the AR(1) model appears to provide a satisfactory and parsimonious fit to these data. The ML estimates for the covariance parameters were computed as

$$\hat{\varphi} = 0.7098, \quad (6.3.1)$$

and,

$$\hat{\sigma}^2 = 0.1405. \quad (6.3.2)$$

The ML estimate for the linear model parameter matrix $B$, and the corresponding asymptotic standard errors are presented in Table 6.3.5.

Substantive hypotheses among the linear model parameters were considered next, and the results of these test procedures are summarized in Table 6.3.6. The significance of the covariate at Hour 0 ($p=0.0014$) suggests that much of the
<table>
<thead>
<tr>
<th>ESTIMATES</th>
<th>CORRESPONDING COVARIANCE SEQUENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lag 0</td>
</tr>
<tr>
<td>AR(1):</td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.7098</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.1405</td>
</tr>
<tr>
<td>AR(2):</td>
<td></td>
</tr>
<tr>
<td>$\phi_1$</td>
<td>0.7164</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>-0.0092</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.1405</td>
</tr>
<tr>
<td>ARMA(1,1):</td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.7038</td>
</tr>
<tr>
<td>$\theta$</td>
<td>-0.0122</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.1405</td>
</tr>
</tbody>
</table>

**TABLE 6.3.4**

ML Estimates of $\phi$ and $\sigma^2$ for Various ARMA Models, with the Corresponding Covariance Sequence

Variation in these repeated measures could be accounted for by inherent differences among the observational units. However, no systematic difference was observed between the Normal and Obese subpopulations across the four time points, either collectively ($p=0.56$) or individually ($0.24 \leq p \leq 0.84$).

Moreover, while the analysis originally reported in Zerbe's paper did not adjust for Hour-0 differences among the subjects in this study population, it is nevertheless instructive to compare the results from the two procedures. Zerbe tested for equality between the two subpopulations at Hours 3, 4, and 5 collectively, and presented evidence at $0.55 \leq p \leq 0.60$ against the null hypothesis (depending upon which of his techniques was actually used). This compares reasonably well with $p=0.43$ reported in Table 6.3.6 from the present methodology. While it is not clear how much of the
<table>
<thead>
<tr>
<th></th>
<th>HOUR2</th>
<th>HOUR3</th>
<th>HOUR4</th>
<th>HOUR5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTROL</td>
<td>3.1230</td>
<td>3.5054</td>
<td>3.6519</td>
<td>4.0682</td>
</tr>
<tr>
<td></td>
<td>(0.1522)</td>
<td>(0.1522)</td>
<td>(0.1522)</td>
<td>(0.1522)</td>
</tr>
<tr>
<td>OBESE</td>
<td>3.0995</td>
<td>3.2715</td>
<td>3.6112</td>
<td>3.9307</td>
</tr>
<tr>
<td></td>
<td>(0.1214)</td>
<td>(0.1214)</td>
<td>(0.1214)</td>
<td>(0.1214)</td>
</tr>
<tr>
<td>HOUR 0</td>
<td>0.5537</td>
<td>0.6005</td>
<td>0.5148</td>
<td>0.4890</td>
</tr>
<tr>
<td></td>
<td>(0.1394)</td>
<td>(0.1394)</td>
<td>(0.1394)</td>
<td>(0.1394)</td>
</tr>
</tbody>
</table>

**TABLE 6.3.5**
ML Estimate of the Linear Model Parameter Matrix B, with the Asymptotic Standard Errors in Parentheses

Discrepancy is simply due to differences in the analytic approaches, the conclusions drawn are basically the same.

As in the previous example, one could perform a piecemeal univariate analysis on these repeated measures, testing for differences between the Control and Obese subpopulations. F-statistics from this approach are reported in Table 6.3.7; all tests were performed with 1 and 30 degrees of freedom. Overall, the ARMA multivariate test procedures compare favourably with these univariate results. For the most part, they show less evidence against the null hypothesis, and are therefore marginally less conservative than the piecemeal univariate technique. However, both approaches appear to be similarly powerful for the purposes of hypothesis testing.
<table>
<thead>
<tr>
<th>LINEAR HYPOTHESIS</th>
<th>$X^2$</th>
<th>D.F.</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariate, Hour 0</td>
<td>17.7087</td>
<td>4</td>
<td>0.0014</td>
</tr>
</tbody>
</table>

Control vs. Obese:

<table>
<thead>
<tr>
<th></th>
<th>$X^2$</th>
<th>D.F.</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours 2, 3, 4, 5</td>
<td>3.0050</td>
<td>4</td>
<td>0.56</td>
</tr>
<tr>
<td>Hours 3, 4, 5</td>
<td>2.7478</td>
<td>3</td>
<td>0.43</td>
</tr>
<tr>
<td>Equality at Hour 2</td>
<td>0.2199</td>
<td>1</td>
<td>0.64</td>
</tr>
<tr>
<td>Equality at Hour 3</td>
<td>1.3654</td>
<td>1</td>
<td>0.24</td>
</tr>
<tr>
<td>Equality at Hour 4</td>
<td>0.0418</td>
<td>1</td>
<td>0.84</td>
</tr>
<tr>
<td>Equality at Hour 5</td>
<td>0.4746</td>
<td>1</td>
<td>0.49</td>
</tr>
</tbody>
</table>

**TABLE 6.3.6**

Results of Hypothesis Testing Among the Linear Model Parameters
<table>
<thead>
<tr>
<th>F-VALUE</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hour2</td>
<td>0.1609</td>
</tr>
<tr>
<td>Hour3</td>
<td>1.1801</td>
</tr>
<tr>
<td>Hour4</td>
<td>0.0405</td>
</tr>
<tr>
<td>Hour5</td>
<td>0.5740</td>
</tr>
</tbody>
</table>

**TABLE 6.3.7**

Hypothesis Testing for Control vs. Obese Subpopulations from Piécelleal Univariate Analyses

6.4 **DISCUSSION**

On the whole, the methodology acquits itself well under this scrutiny. The techniques reveal themselves to be applicable under a variety of conditions, from small datasets to intermediate-sized, for complete or incomplete data, in the presence or absence of concomitant variables, and so on. It therefore represents a wide-ranging analytic device.

Stochastic structures known to exist in the linear parameter model parameter matrix B were confirmed by the procedures. Some concern was raised with respect to the Group 2 cell mean at Time 1 for the generated dataset. However, in view of the fact that 14 cell means were estimated, it is perhaps not surprising that one should lie beyond two standard errors of its theoretical value. This may well reflect normal statistical variability in the random number generator. Moreover, no differences were observed between the Control and Obese subpopulations in the second dataset, replicating other published results.

The ARMA time series models provided a persuasive accounting of their properties throughout this chapter. It is
generally difficult to discriminate among several competing ARMA models unless there are intervening theoretical considerations. For example, any of the two- or three-parameter models in Example 1 would have provided an effective means of summarizing the covariance structure. The methodology did confirm the suitability of the ARMA(1,1) model in this case, and this represents a vote of confidence in these techniques.

An alternative analysis strategy which was explored in this chapter is to perform a series of univariate analyses, in a piecemeal fashion, across the T repeated measures. Computationally, this represents a considerably less demanding analytic approach, but at the expense of information concerning the covariance structure among the repeated measures. Evidence was presented to suggest that the ARMA multivariate techniques perform no worse than these univariate procedures, and in many cases, considerably better. Furthermore, the intercorrelation structures really form an integral component of an overall inference strategy, providing insight into the temporal relationships among the repeated measures, and were thought to justify the heavy computational burden.

While the piecemeal univariate approach ignores information in the off-diagonal elements of the covariance matrix, at the opposite extreme, one could employ the entire roster of $T(T+1)/2$ covariance elements of the general linear multivariate model. The strength of the ARMA time series models lies in the fact that they offer a reasonable compromise between these two extremes. They not only provide a parsimonious description of the entire covariance matrix by using only 3 or 4 parameters, but they can also accommodate a wide variety of covariance models. They therefore provide an attractive model by which to analyze repeated measures experiments.

This may have broader implications in the presence of missing values. For example, vagaries in the missing-value
patterns could very easily result in two time points which are simultaneously observed by a very small number of observational units. Under the general covariance model, this would likely result in an unreliable estimate of the corresponding element of the covariance matrix. The ARMA parameters are estimated by a wider constituency of data elements, and are therefore less susceptible to these peculiarities.

At the same time, it should be clear that the ARMA models offer no panacea for the analysis of repeated measures experiments. The Toeplitz pattern provided only a marginally satisfactory fit to the data reported in Zerbe(1979). Clearly, the techniques cannot be applied blindly, and some effort must be made to ensure that the ARMA model is appropriate for any dataset under consideration. This is the value of the diagnostic procedures introduced in this chapter, and the formal tests of significance derived in the previous chapter.

Moreover, the analyses sharply draw into focus limitations in a visual inspection of the estimated covariance matrix among the repeated measures. Whereas in the generated data, the methodology tolerated a fair amount of variability along the diagonals, seemingly minor variations along the main diagonal in Zerbe's dataset were apparently sufficient to weaken the Toeplitz assumption. A better strategy, however, is to examine the progression in the covariance sequence across each row of the upper-triangle of this matrix. If the same basic decreasing pattern is evident for each row, then there is greater confidence in the Toeplitz paradigm. This was the case in Example 1, but not in Example 2.

This is essentially the rationale behind the diagnostic procedures introduced in this chapter. Examining the time-specific autocorrelation sequences across the various time points is equivalent to examining the rows of the covariance matrix. These plots are nevertheless fairly limited in the information they convey, and the advantage of a formal test of significance for the general Toeplitz pattern is established.

- 126 -
Finally, there was some suggestion that $\sigma^2$ may have been underestimated by this methodology. This points to the general need for a systematic evaluation of the small-sample properties of all these estimators using a Monte Carlo methodology.
Chapter VII
SUMMARY AND CONCLUSIONS

A unified inference strategy has been presented for the analysis of the incomplete longitudinal design under an autoregressive-moving average covariance structure. The issue of primary concern was the joint estimation of the parameters indigenous to this model. A review of the literature had recommended maximum likelihood, and it was adopted as the estimation criterion for this study.

The first-order maximizing equations were derived for each set of parameters. Large-sample properties of the ML estimators of the parameters were considered, and they were shown to be consistent, asymptotically normal, and asymptotically efficient. Moreover, an expression for the asymptotic covariance matrix among the linear model parameters was presented.

This estimator was, however, shown to be a function of the ML estimator of the ARMA time series parameter vector. No closed-form solution was forthcoming for this parameter, and non-linear optimization procedures were prescribed to iterate to a maximizing value. The concentrated log likelihood function, its gradient, and a procedure for calculating the initial values for these parameters were documented in this study. Moreover, the convergence properties of the ensuing iteration sequence were investigated, and under suitable regularity conditions, it was shown to result in an unbiased estimator for the linear model parameter vector.

A second and equally important component of this inference strategy was to specify procedures for testing hypotheses among the indigenous parameters. Two broad categories of hypotheses were identified. The first examined the co-
variance structure among the repeated measures. A likelihood ratio test procedure was described to test the assertion that the general Toeplitz paradigm is consistent with the data; and, contingent upon an affirmative response, a second procedure was prescribed for any particular ARMA covariance model. In addition, procedures for testing substantive hypotheses among the linear model parameters were also presented.

Two datasets were studied to illustrate these procedures. The first consisted of data generated at random from a known stochastic mechanism; the second was derived from a paper published in the statistical literature. The structures inherent in these datasets were largely verified by the methodology, and therefore represents a vote of confidence in these techniques.

The methodology is critically dependent upon a rather large number of parametric assumptions. Obviously, the more closely the data conform to these assumptions, the more relevant the analysis will be. However, the same criticisms which were laid against standard growth curve analysis will also apply in full force to this model. One point of contention certainly is multivariate normality among the disturbance terms from each observational unit. It might be well worthwhile to verify this assumption, particularly in the presence of missing data, before embarking on these procedures.

One of the cornerstones upon which the methodology has been laid is the Toeplitz structure in the covariance matrix among the repeated measures. It is specifically assumed that the elements along any diagonal of this matrix are all the same. Yet as was illustrated in one of the datasets, seemingly minor variations along a single diagonal may be sufficient to weaken this assumption. The general Toeplitz model is seen to be a fairly restrictive assumption, and is likely to invalidate these procedures in many situations arising in practice.
If however the data do conform to this pattern, then the ARMA time series representations offer a vast panorama of covariance models. This stands in contrast to the assumption of "compound symmetry" underlying standard repeated measures analysis. The ARMA models extend well beyond this horizon, accommodating a wide variety of covariance structures, and therefore offer a convenient vehicle in which to analyze repeated measures experiments.

The advantage, of course, is that the covariance matrix is dependent upon a small number of parameters, generally no more than three or four. It therefore offers a reasonable compromise between a piecemeal univariate approach which ignores information in the off-diagonal elements of the covariance matrix, and the general linear multivariate model which requires the entire set of T(T+1)/2 distinct covariance elements. This may also have broader implications under the influence of missing values. Vagaries in the missing-value patterns can easily result in unreliable estimates for one or more of these T(T+1)/2 parameters. The ARMA parameters are estimated by a wider constituency of data elements, more effectively insulating the estimators from these peculiarities.

This also has important implications with respect to multivariate power calculations in preparation for an applied research effort. To determine the appropriate sample size, for example, an estimate of the covariance matrix among the repeated measures is required. It is obviously easier to provide values for (or, to systematically vary) a small number of ARMA parameters than the entire roster of T(T+1)/2 parameters under the general covariance model.

One of the great disadvantages of the current techniques is the expense of performing these calculations. While estimates of computer time and resources are tenuous and invariably installation-dependent, it is suggested that estimation under an ARMA covariance model is 2–3 times more expensive than for the corresponding analysis under the EM
algorithm. Moreover, while the cost appears to increase linearly with the sample size N, there is an exponential increase with the number of time points T. Unless analytic procedures like the Kalman filter can be implemented to expedite the calculations, estimation under these covariance models will likely remain an expensive proposition.

There are two areas where the present methodology can be easily extended. Whereas the current study was restricted to a single dependent measure, there is a corresponding need to consider a multivariable (or, "multivariate") repeated measures analysis. The methodology would want, in the most general case, to accommodate different ARMA models for the various response variables, with perhaps a test for uniformity across the entire set. In addition, growth curve analysis under an ARMA covariance model is an obvious extension which could easily be accommodated by the basic model.

Finally, while the large-sample properties of these estimators have been investigated, there is a need to consider the small-sample properties as well. There was some suggestion that the white noise variance $\sigma^2$ was underestimated by the methodology, and a more systematic analysis is clearly warranted. More generally, a Monte Carlo study could be mounted to investigate the characteristics of all the ML estimators with respect to the shape of their sample distribution, bias, mean square error, and so on.
BIBLIOGRAPHY


<table>
<thead>
<tr>
<th>TIME1</th>
<th>TIME2</th>
<th>TIME3</th>
<th>TIME4</th>
<th>TIME5</th>
<th>TIME6</th>
<th>TIME7</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.68</td>
<td>14.41</td>
<td>18.60</td>
<td>23.83</td>
<td>28.80</td>
<td>34.07</td>
<td>38.83</td>
</tr>
<tr>
<td>7.89</td>
<td>13.64</td>
<td>18.72</td>
<td>22.31</td>
<td>27.82</td>
<td>32.92</td>
<td>38.78</td>
</tr>
<tr>
<td>8.90</td>
<td>14.88</td>
<td>21.16</td>
<td>26.02</td>
<td>30.84</td>
<td>35.71</td>
<td>39.73</td>
</tr>
<tr>
<td>10.51</td>
<td>13.59</td>
<td>18.72</td>
<td>22.31</td>
<td>27.82</td>
<td>32.92</td>
<td>38.78</td>
</tr>
<tr>
<td>9.46</td>
<td>14.78</td>
<td>18.74</td>
<td>23.82</td>
<td>28.74</td>
<td>33.28</td>
<td>36.08</td>
</tr>
<tr>
<td>10.83</td>
<td>12.91</td>
<td>18.59</td>
<td>22.55</td>
<td>27.91</td>
<td>31.72</td>
<td>36.80</td>
</tr>
<tr>
<td>7.92</td>
<td>13.89</td>
<td>18.74</td>
<td>24.37</td>
<td>29.44</td>
<td>34.54</td>
<td>38.48</td>
</tr>
<tr>
<td>7.97</td>
<td>13.89</td>
<td>18.74</td>
<td>24.37</td>
<td>29.44</td>
<td>34.54</td>
<td>38.48</td>
</tr>
<tr>
<td>9.29</td>
<td>13.43</td>
<td>17.47</td>
<td>22.19</td>
<td>34.60</td>
<td>38.71</td>
<td></td>
</tr>
<tr>
<td>10.09</td>
<td>14.97</td>
<td>19.23</td>
<td>23.50</td>
<td>28.11</td>
<td>37.76</td>
<td>41.64</td>
</tr>
<tr>
<td>8.84</td>
<td>14.67</td>
<td>19.59</td>
<td>24.39</td>
<td>36.32</td>
<td>41.36</td>
<td></td>
</tr>
<tr>
<td>10.05</td>
<td>13.71</td>
<td>17.99</td>
<td>22.82</td>
<td>27.78</td>
<td>31.75</td>
<td>36.29</td>
</tr>
<tr>
<td>9.75</td>
<td>13.71</td>
<td>17.99</td>
<td>22.82</td>
<td>27.78</td>
<td>31.75</td>
<td>36.29</td>
</tr>
<tr>
<td>10.96</td>
<td>16.08</td>
<td>21.25</td>
<td>27.10</td>
<td>31.05</td>
<td>35.15</td>
<td>40.11</td>
</tr>
<tr>
<td>11.44</td>
<td>15.84</td>
<td>24.54</td>
<td>29.16</td>
<td>33.36</td>
<td>38.45</td>
<td></td>
</tr>
<tr>
<td>10.74</td>
<td>15.03</td>
<td>20.49</td>
<td>24.28</td>
<td>31.76</td>
<td>36.02</td>
<td>42.51</td>
</tr>
<tr>
<td>10.52</td>
<td>15.27</td>
<td>24.30</td>
<td>31.10</td>
<td>33.43</td>
<td>38.07</td>
<td></td>
</tr>
<tr>
<td>8.33</td>
<td>14.67</td>
<td>18.11</td>
<td>22.18</td>
<td>29.59</td>
<td>32.63</td>
<td>38.88</td>
</tr>
<tr>
<td>7.88</td>
<td>13.67</td>
<td>18.11</td>
<td>22.18</td>
<td>29.59</td>
<td>32.63</td>
<td>38.88</td>
</tr>
<tr>
<td>11.22</td>
<td>13.69</td>
<td>19.33</td>
<td>25.36</td>
<td>32.69</td>
<td>36.51</td>
<td></td>
</tr>
<tr>
<td>8.99</td>
<td>13.69</td>
<td>19.33</td>
<td>25.36</td>
<td>32.69</td>
<td>36.51</td>
<td></td>
</tr>
<tr>
<td>12.65</td>
<td>17.16</td>
<td>21.97</td>
<td>31.87</td>
<td>35.90</td>
<td>38.64</td>
<td></td>
</tr>
<tr>
<td>11.04</td>
<td>18.64</td>
<td>22.48</td>
<td>27.34</td>
<td>35.90</td>
<td>38.64</td>
<td></td>
</tr>
</tbody>
</table>

---

GROUP=1

---
<table>
<thead>
<tr>
<th>TIME1</th>
<th>TIME2</th>
<th>TIME3</th>
<th>TIME4</th>
<th>TIME5</th>
<th>TIME6</th>
<th>TIME7</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:32</td>
<td>8:05</td>
<td>7:39</td>
<td>7:30</td>
<td>8:81</td>
<td>8:97</td>
<td>7:21</td>
</tr>
<tr>
<td>30:92</td>
<td>38:60</td>
<td>29:52</td>
<td>26:66</td>
<td>30:38</td>
<td>29:05</td>
<td>26:05</td>
</tr>
<tr>
<td>49:04</td>
<td>60:12</td>
<td>60:91</td>
<td>59:66</td>
<td>60:12</td>
<td>60:41</td>
<td>60:41</td>
</tr>
<tr>
<td>8:40</td>
<td>8:40</td>
<td>8:40</td>
<td>8:40</td>
<td>8:40</td>
<td>8:40</td>
<td>8:40</td>
</tr>
<tr>
<td>16:10</td>
<td>16:10</td>
<td>16:10</td>
<td>16:10</td>
<td>16:10</td>
<td>16:10</td>
<td>16:10</td>
</tr>
<tr>
<td>26:44</td>
<td>26:44</td>
<td>26:44</td>
<td>26:44</td>
<td>26:44</td>
<td>26:44</td>
<td>26:44</td>
</tr>
<tr>
<td>58:77</td>
<td>58:77</td>
<td>58:77</td>
<td>58:77</td>
<td>58:77</td>
<td>58:77</td>
<td>58:77</td>
</tr>
</tbody>
</table>
INDEX OF REFERENCES CITED

Aigner (1971) ... 13
Anderson (1957) ... 34
Ansky & Kohn (1983) ... 23, 73
Ansky & Newbold (1980) ... 21, 67
Ansky (1979) ... 22-23
Balestra & Nerlove (1966) ... 28
Bart (1974) ... 17, 20, 69
Bartlett (1946) ... 98
Beach & Mackinnon (1978a) ... 14-15
Beach & Mackinnon (1978b) ... 14, 16
Beale & Little (1975) ... 35
Belsley (1980) ... 18, 20, 22, 67, 70
Bjorn (1981) ... 36
Box & Jenkins (1970) ... 6-9, 12, 16, 21, 46-47, 67, 71-73, 98
Cochrane & Orcutt (1949) ... 12, 22, 70
da Silva (1975) ... 27-28, 31, 37
Damsleth (1983) ... 16
Davidson (1959) ... 19
Dempster, Laird & Rubin (1977) ... 33, 83
Dent & Mink (1978) ... 15
Dent (1977) ... 22
Diehl (1983) ... 27
Dixon (1983) ... 21
Don & Magnus (1980) ... 14, 58
Draper & Smith (1981) ... 17, 20
Dunsanmu & Robinson (1981) ... 71
Dwyer (1967) ... 48
Finley (1981) ... 47
Fishman & Moore (1982) ... 101
Fletcher & Powell (1963) ... 19
Fletcher (1969) ... 17
Fletcher (1980) ... 17
Fletcher (1981) ... 17
Galbraith & Galbraith (1974) ... 76
Gardner et al (1980) ... 23
Goldfeld & Quandt (1970) ... 17
Goodnight (1979) ... 74
Griliches & Rao (1969) ... 14
Grizzle & Allen (1969) ... 26, 114
Harris (1975) ... 26
Hartley & Booker (1965) ... 20, 69
Hartley & Hocking (1971) ... 34-35
Harvey & Phillips (1979) ... 23, 73
Hechman & Singer (1982) ... 24
Henley & Richard (1983) ... 27
Hildreth (1969) ... 13
Hocking & Smith (1968) ... 34-35
Hosking (1980) ... 35, 83, 113
Hussain (1969) ... 28
Huynh & Feldt (1970) ... 26
Jennrich (1969) ... 20, 69
Jones (1980) ... 23, 73
Jones (1984a) ... 37, 73
Jones (1984b) ... 37
Judge et al (1980) ... 11-12, 17, 30, 32, 76
Kakwani (1967) ... 29
Kalman & Bucy (1961) ... 22
Kalman (1960) ... 22
Kang (1975) ... 21, 67
Kendall & Stuart (1967) ... 51, 78
Kiefer (1980) ... 24, 26
Kleibbaum (1970) ... 35, 44, 69
Kleibbaum (1973) ... 36-37
Kmenta & Gilbert (1970) ... 31, 113
Kmenta (1971) ... 31-32, 37
LeVan (1983) ... 37, 42-44, 51, 69, 71-72
Leeper & Woolson (1982) ... 37
Magnus (1978) ... 14, 52
Malinvaud (1980) ... 32
Margaudt (1963) ... 19-20
McGillchrist et al (1981) ... 21, 46
Meinhold & Singpurwalla (1983) ... 23
Murray (1972) ... 17
Nelson (1974) ... 15
Neudecker (1969) ... 90
Newbold (1974) ... 22
Newbold (1981) ... 32
Nicholls & Hail (1979) ... 22
Nicholls, Pagan & Terrell (1975) ... 21-22, 69
Oberhofer & Kmenta (1974) ... 51, 62-63
Orchard & Woodbury (1972) ... 33, 35, 104
<table>
<thead>
<tr>
<th>Reference</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Osborn (1976)</td>
<td>15, 21</td>
</tr>
<tr>
<td>Osborn (1982)</td>
<td>16</td>
</tr>
<tr>
<td>Pagan &amp; Nicholls (1976)</td>
<td>14, 22</td>
</tr>
<tr>
<td>Parks (1967)</td>
<td>30-31</td>
</tr>
<tr>
<td>Parzen (1963)</td>
<td>70</td>
</tr>
<tr>
<td>Patel &amp; Khatri (1981)</td>
<td>36</td>
</tr>
<tr>
<td>Pesaran (1973)</td>
<td>22</td>
</tr>
<tr>
<td>Pierce (1971)</td>
<td>13-14</td>
</tr>
<tr>
<td>Powell (1981)</td>
<td>17</td>
</tr>
<tr>
<td>Rao (1973)</td>
<td>19</td>
</tr>
<tr>
<td>Richards (1961)</td>
<td>63</td>
</tr>
<tr>
<td>Rubin &amp; Szatrowski (1982)</td>
<td>33-34</td>
</tr>
<tr>
<td>Sandland &amp; McGilchrist (1979)</td>
<td>21</td>
</tr>
<tr>
<td>SAS Institute (1982a)</td>
<td>101</td>
</tr>
<tr>
<td>SAS Institute (1982b)</td>
<td>21-22</td>
</tr>
<tr>
<td>Savin &amp; White (1978)</td>
<td>22, 73</td>
</tr>
<tr>
<td>Spitzer (1979)</td>
<td>14, 21, 67</td>
</tr>
<tr>
<td>Swamy &amp; Arora (1972)</td>
<td>27</td>
</tr>
<tr>
<td>Theil (1971)</td>
<td>13, 32</td>
</tr>
<tr>
<td>Tiao &amp; Ali (1971)</td>
<td>22, 76</td>
</tr>
<tr>
<td>Timm (1975)</td>
<td>39, 119</td>
</tr>
<tr>
<td>Trawinski &amp; Bargman (1964)</td>
<td>35, 48-49</td>
</tr>
<tr>
<td>Walker (1964)</td>
<td>13</td>
</tr>
<tr>
<td>Wallace &amp; Hussain (1969)</td>
<td>27-28</td>
</tr>
<tr>
<td>Whittle (1951)</td>
<td>13</td>
</tr>
<tr>
<td>Whittle (1962)</td>
<td>13</td>
</tr>
<tr>
<td>Wilks (1962)</td>
<td>79</td>
</tr>
<tr>
<td>Wilson (1969)</td>
<td>21, 72</td>
</tr>
<tr>
<td>Woolson et al (1978)</td>
<td>36</td>
</tr>
<tr>
<td>Zellner (1967)</td>
<td>28-29</td>
</tr>
<tr>
<td>Zerbe (1979)</td>
<td>114, 121, 126</td>
</tr>
</tbody>
</table>