1972

Estimation for linear models with unknown diagonal covariance matrix

John David Jobson
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation
Jobson, John David, "Estimation for linear models with unknown diagonal covariance matrix" (1972). Retrospective Theses and Dissertations. 4743.
https://lib.dr.iastate.edu/rtd/4743
INFORMATION TO USERS

This dissertation was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
A Xerox Education Company
JOBSON, John David, 1941-
ESTIMATION FOR LINEAR MODELS WITH UNKNOWN
DIAGONAL COVARIANCE MATRIX.

Iowa State University, Ph.D., 1972
Statistics

University Microfilms, A XEROX Company, Ann Arbor, Michigan
Estimation for linear models with unknown diagonal covariance matrix

by

John David Jobson

A Dissertation Submitted to the Graduate Faculty in Partial Fulfillment of The Requirements for the Degree of DOCTOR OF PHILOSOPHY

Major Subject: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa
1972
PLEASE NOTE:

Some pages may have indistinct print.
Filmed as received.

University Microfilms, A Xerox Education Company
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. A REVIEW OF ESTIMATION IN THE CLASS $\xi$</td>
<td>3</td>
</tr>
<tr>
<td>A. The Least Squares Procedure</td>
<td>4</td>
</tr>
<tr>
<td>B. The Maximum Likelihood Procedure</td>
<td>26</td>
</tr>
<tr>
<td>III. EXISTENCE AND PROPERTIES OF ESTIMATORS FOR A SUBCLASS OF $\xi$</td>
<td>34</td>
</tr>
<tr>
<td>A. The Principal Assumptions</td>
<td>34</td>
</tr>
<tr>
<td>B. The Likelihood Function and Properties of the Maximum Likelihood Estimator</td>
<td>40</td>
</tr>
<tr>
<td>C. Existence and Properties of Simple Least Squares Estimators</td>
<td>89</td>
</tr>
<tr>
<td>D. A Joint Least Squares Procedure and Its Properties</td>
<td>118</td>
</tr>
<tr>
<td>E. Convergence of the Joint Least Squares Estimator to the Maximum Likelihood Estimator</td>
<td>147</td>
</tr>
<tr>
<td>IV. APPLICATION OF THE DERIVED PROPERTIES OF THE ESTIMATORS TO SOME EXAMPLES</td>
<td>174</td>
</tr>
<tr>
<td>V. A COMPARISON OF THE PERFORMANCES OF THE SIMPLE AND JOINT LEAST SQUARES ESTIMATORS FOR A RANDOM COEFFICIENT MODEL</td>
<td>179</td>
</tr>
<tr>
<td>A. The Procedure</td>
<td>179</td>
</tr>
<tr>
<td>B. Summary of Results</td>
<td>182</td>
</tr>
<tr>
<td>VI. SUMMARY</td>
<td>188</td>
</tr>
<tr>
<td>VII. LITERATURE CITED</td>
<td>190</td>
</tr>
<tr>
<td>VIII. ACKNOWLEDGEMENTS</td>
<td>193</td>
</tr>
</tbody>
</table>
I. INTRODUCTION

We consider the problem of estimation in the class $\mathcal{L}$ of linear models defined as follows:

$$Y = X\beta + e ;$$

where $Y$ is an $n \times 1$ vector of observed real random variables, $X$ is an $n \times p$ matrix of rank $p$ of observed real constants, $\beta$ is a $p \times 1$ vector of unknown real parameters, $e$ is an $n \times 1$ vector of unobserved random variables, $E[e] = 0$; $E[ee'] = G$.

$G$ is a diagonal matrix with nonzero diagonal elements whose structure is known up to an $s \times 1$ vector $\theta$ of unknown real parameters. $\theta$ may or may not contain elements of $\beta$. The vectors $\beta$ and $\theta$ are to be estimated.

A review of estimation for some examples in the class $\mathcal{L}$ is given in Section II. Two common procedures, a maximum likelihood procedure, and a least squares procedure are discussed. The least squares procedure uses residuals from a simple least squares fit of $\beta$ to estimate $\theta$ and hence $G$. A weighted least squares procedure is then used to obtain a new estimate of $\beta$. The weights are obtained from the estimates of the diagonal elements of $G$. The maximum likelihood procedure is employed under the additional
assumption that the vector \( e \) is normally distributed.

In Section III a subclass of the class \( \mathcal{L} \) is defined by adding assumptions about the properties of the diagonal elements of \( G \) and the density of \( e \). For this subclass we first discuss the existence and properties of the maximum likelihood estimators of \( \beta \) and \( \theta \). We then discuss the existence and properties of simple least squares estimators of \( \beta \) and \( \theta \). A joint least squares procedure is then developed which uses preliminary estimates of \( \beta \) and \( \theta \) to obtain new estimates. It is shown that this joint least squares estimator has the same asymptotic distribution as the maximum likelihood estimator. Finally we demonstrate that a sequence of estimators can be obtained that converge to a local maximum of the likelihood function. This sequence is obtained by repeatedly applying the joint least squares procedure.

The application of the joint least squares procedure to several examples given in Section II is discussed in Section IV. The properties of the simple and joint least squares procedures are used to obtain properties for estimators in the specific examples.

In Section V we compare the performance of the simple and joint least squares estimators for a random coefficient model. A measure of the adequacy of the large sample results for a sample of size 40 is obtained from a Monte Carlo study.
II. A REVIEW OF ESTIMATION IN THE CLASS £

It is well known that for the class £ of linear models where \( \beta \) is not an unknown parameter of \( G \) the best linear unbiased estimator of \( \beta \) when \( G \) is known is given by the generalized least squares estimator

\[
\hat{\beta}^* = (X'G^{-1}X)^{-1}X'G^{-1}Y. \tag{2.0.1}
\]

If \( e \) is also normally distributed, \( \hat{\beta}^* \) is the maximum likelihood estimator of \( \beta \). The simple least squares estimator which is defined by

\[
\hat{\beta} = (X'X)^{-1}X'Y \tag{2.0.2}
\]

is also unbiased but in general less efficient than \( \hat{\beta}^* \). Conditions under which \( \hat{\beta}^* \) and \( \hat{\beta} \) have the same covariance matrix are given by Zyskind (1967) and Williams (1967).

Since the covariance matrix of \( \hat{\beta} \) is a function of \( G \) an estimate of \( \Theta \) is required before an estimate of the precision of \( \hat{\beta} \) can be constructed from the sample. If \( G \) is unknown and an estimator \( \hat{\Theta} \) of \( \Theta \) is available a reasonable estimator of \( \beta \) is given by

\[
\bar{\beta} = (X'G^{-1}X)^{-1}X'G^{-1}Y; \tag{2.0.3}
\]

where \( \hat{\Theta} \) is obtained from \( G \) by replacing \( \Theta \) by \( \hat{\Theta} \). The estimator \( \bar{\beta} \) may or may not be more efficient than \( \hat{\beta} \). The case wherein the estimator \( \hat{\beta} \)
of \( \Theta \) is obtained from an independent sample has been discussed by C.R. Rao (1965b, 1965c). We shall consider the case where \( \hat{\Theta} \) is obtained from the simple least squares residuals. We shall call this approach to the estimation of \( \beta \) and \( \Theta \) the least squares procedure.

If \( e \) is normally distributed the method of maximum likelihood can be used to estimate \( \beta \) and \( \Theta \). The values of \( \beta \) and \( \Theta \), \( \tilde{\beta} \) and \( \tilde{\Theta} \), that maximize the likelihood function are called the maximum likelihood estimators.

A. The Least Squares Procedure

The least squares procedure begins with the residuals, \( \hat{e} = Y - X\hat{\beta} \), where \( \hat{\beta} \) is given by (2.0.2). These residuals are used to estimate the covariance matrix \( G \) by estimating the unknown parameter \( \Theta \). A new estimator of \( \beta \) is then obtained from (2.0.3). A study of this procedure requires an investigation of the techniques of covariance matrix estimation using least square residuals.

Covariance matrix estimation has been studied recently by C.R. Rao (1970) and Chew (1970). The techniques discussed by these authors have also been used by Hildreth and Houck (1968) and Hartley, Rao and Kiefer (1969). The least squares residuals are defined by

\[
\hat{e} = Y - X\hat{\beta},
\]

\[
= Y - X(X'X)^{-1}X'Y,
\]

\[
= e - X(X'X)^{-1}X'e,
\]

\[
= M e; \tag{2.1.1}
\]
where $M = I - X(X'X)^{-1}X'$.

It follows that

$$E[ee'] = MGM.$$ 

Let $g$ be the vector of diagonal elements of $G$, let $M$ be the matrix whose elements are the squares of the elements of $M$, and let $e$ be the vector whose elements are the squares of the elements of $e$. Since $G$ is diagonal we may write

$$E[ee'] = Mg. \quad (2.1.2)$$ 

Depending on additional assumptions about the structure of $G$ the elements of $g$, $\{g_i\}_{i=1}^n$, may not be all different. If the elements of $g$ are all different then (2.1.2) is a system of $n$ equations in the $n$ unknowns $\{g_i\}_{i=1}^n$.

The method of moments suggests the estimation equations

$$\hat{e} = \hat{M} g. \quad (2.1.3)$$

If $M^{-1}$ exists we obtain

$$\hat{g} = M^{-1} \hat{e}. \quad (2.1.4)$$
Sufficient conditions for the rank of \( \hat{M} \) to be full are given by Hartley, Rao, and Kiefer (1969) and by C.R. Rao (1970). Hartley, Rao and Kiefer give the sufficient condition

\[
\max_{i} \sum_{j=1}^{n} |a_{ij}| < 1 \quad \text{for } i = 1, 2, \ldots, n;
\]

where

\[
a_{ii} = 1 - \hat{m}_{ii} \quad i = 1, 2, \ldots, n;
\]

\[
a_{ij} = -\hat{m}_{ij} \quad i \neq j \quad i, j = 1, 2, \ldots, n;
\]

\[
\hat{M} = \{\hat{m}_{ij}\}_{i,j=1}^{n}.
\]

This condition reduces to

\[
\min_{i} \hat{m}_{ii}^{\frac{1}{2}} > \frac{1}{2}
\]

since

\[
\sum_{j=1}^{n} \hat{m}_{ij} \hat{m}_{ij} = \hat{m}_{ii}^{\frac{1}{2}} \Rightarrow 0 < \hat{m}_{ii} < 1.
\]
Sufficient conditions due to C.R. Rao (1970) are summarized in the following two lemmas.

**Lemma 2.1.1.** Given that the following conditions are satisfied:

1. \( X_{nxp} \) is of rank \( p \) where \( n \geq 2p+1 \)
2. \( X' \) can be partitioned after a rearrangement of columns as \( (X'_1 : X'_2 : X'_3) \) where
   a. \( X_1 \) is \( pxm \) of rank \( p \),
   b. \( X_2 \) is \( pxm \) of rank \( p \) and
   c. \( X_3 \) is such that not all its rows depend on a proper subset of rows of \( X_1 \) and such that not all its rows depend on a proper subset of rows \( X_3 \),

then \( \hat{M} \) is of full rank.

**Lemma 2.1.2.** Given that every column vector belonging to the linear manifold \( M(X) \) contains more than \( (p+1) \) nonzero elements then \( \hat{M} \) is of full rank. The linear manifold \( M(X) \) is the subspace generated by all possible linear combinations of the columns of \( X \).

C.R. Rao (1970) defines an estimator \( Y'AY \) of a linear function of the diagonal elements \( \sum_{i=1}^{n} c_i g_i \) to be MINQUE if the Euclidean norm of \( A, \|A\| \) (the square root of the trace of \( A \)) is minimized subject to the conditions:

1. \( AX = 0 \)
2. \( \sum_{i=1}^{n} a_{ij} g_i = \sum_{i=1}^{n} c_i g_i \)
He then proves that the MINQUE method of estimation has the following desirable properties:

1. The MINQUE is invariant for orthogonal transformations of $Y$.

2. If $S_1$ is the MINQUE of $\sum_{i=1}^{n} c_i g_i$ and $S_2$ is the MINQUE of $\sum_{i=1}^{n} d_i g_i$ then $S_1 + S_2$ is the MINQUE of $\sum_{i=1}^{n} (c_i + d_i) g_i$.

3. If all $g_i$, $i = 1, 2, \ldots, n$ are different the MINQUE of $g$ has the minimum average variance in the class of quadratic unbiased estimators for any symmetric a priori distribution of $g_1, \ldots, g_n$ over which the average is taken.

C.R. Rao (1970) proves that if the $g_i$, $i = 1, 2, \ldots, n$ are all different then the MINQUE of $g$ is given by $\hat{g}$ in (2.1.4). If only $k < n$ of the $g_i$, $i = 1, 2, \ldots, n$ are different then the equations for the MINQUE of $g$ are obtained from (2.1.4) as follows. Without loss of generality let the first $n_1$ elements of $g$ be equal to $\gamma_1$, the next $n_2$ elements of $g$ be equal to $\gamma_2$, ..., the last $n_k$ elements of $g$ be equal to $\gamma_k$, $\sum_{j=1}^{k} n_j = n$. Add up the first $n_1$ equations to obtain the first equation in $\gamma_1, \gamma_2, \ldots, \gamma_k$; then add up the next $n_2$ equations to obtain the second equation in $\gamma_1, \gamma_2, \ldots, \gamma_k$ and so on; to obtain a total of $k$ equations. The equations are given by

$$\hat{W} = \hat{H} \gamma; \quad (2.1.5)$$
where

$$h_{ij} = \sum_{a_{i-1}+1}^{a_i} \sum_{a_{j-1}+1}^{a_j} m_{rs} \quad i,j = 1, 2, \ldots, k;$$

$$a_i = \sum_{t=1}^{i} n_t \quad i = 1, 2, \ldots, k; \quad \mathbf{H} = \{ h_{ij} \}_{i,j=1}^{k};$$

$$\hat{W}_i = \sum_{a_{i-1}+1}^{a_i} e_s \quad i = 1, 2, \ldots, k; \quad \hat{W} = \{ \hat{W}_i \}_{i=1}^{k}.$$  

The MINQUE of $\gamma$ is then given by

$$\gamma = H^{-1} \hat{W}. \quad (2.1.6)$$

A common alternative to (2.1.6) has been to replace $\mathbf{H}$ by the diagonal matrix

$$N = \begin{bmatrix} n_1 - 1 \\ & n_2 - 1 \\ & & \ddots \\ & & & n_k - 1 \end{bmatrix}.$$
Thus the estimator of $\gamma_j$ is given by

$$
\bar{\gamma}_j = \frac{\sum_{i=1}^{n_j} e_{ij}}{n_j - 1}.
$$

(2.1.7)

Since $H = N + O(\frac{1}{n})$ the procedure is justified in large samples. Rao and Subrahmaniam (1971) have shown that (2.1.6) provides a more efficient estimator for $\gamma$ than (2.1.7) for small samples in some special cases.

A problem with the MINQUE approach is the possibility of negative estimates of the $\gamma_j$, $j = 1, 2, \ldots, k$. Clearly this is not possible if (2.1.7) is used. Rao and Subrahmaniam (1971) employed the following adjustment to the MINQUE procedure to eliminate negative estimates.

$$
\gamma_j^* = \begin{cases} 
\gamma_j & \text{if } \gamma_j > \epsilon \\
\bar{\gamma}_j & \text{if } \gamma_j < \epsilon 
\end{cases}
$$

where $\epsilon$ is an arbitrary small number.

For the case in which only $k < n$ of the $g_i$, $i = 1, 2, \ldots, n$ are distinct, say $\gamma_1, \gamma_2, \ldots, \gamma_k$, Chew (1970) suggests the following alternative procedure. If we assume without loss of generality that the first $n_1$ elements of $g$ are equal to $\gamma_1$, the next $n_2$ elements of $g$ are equal to $\gamma_2$, $\ldots$ and the last $n_k$ elements of $g$ are equal to $\gamma_k$.
we may rewrite (2.1.3) as

$$
\hat{e} = \hat{F} \gamma \quad ;
$$

(2.1.8)

where

$$
\hat{F}_{nxk} = \{\hat{f}_{ij}\} ;
$$

$$
\hat{f}_{ij} = \sum_{r=a_{j-1}+1}^{a_j} \frac{m_{ir}}{r=a_{j-1}+1} \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, k;
$$

$$
\hat{a}_j = \sum_{s=1}^{j-1} n_s \quad j = 1, 2, \ldots, k .
$$

If we let \( W = \hat{e} - E[\hat{e}] \) we may write

$$
\hat{e} = \hat{F} \gamma + W ;
$$

(2.1.9)

Chew (1970) uses the least squares estimator

$$
\gamma^* = (\hat{F}'\hat{F})^{-1}\hat{F}'\hat{e} .
$$

(2.1.10)

The least squares estimator may also yield negative estimates.

An example of the application of these techniques is provided by the following nonparametric unequal variance model.
\[ y_{ij} = \sum_{k=1}^{p} \beta_{ki} x_{ijk} + e_{ij} \quad i = 1, 2, \ldots, n_j \; ; \]
\[ j = 1, 2, \ldots, m \; ; \]

\[ E[e_{ij}] = 0 \quad i = 1, 2, \ldots, n_j \; ; \]
\[ j = 1, 2, \ldots, m \; ; \]

\[ E[e_{ij} e_{rs}] = \sigma_j^2 \quad i, r = 1, 2, \ldots, n_j \; ; \]
\[ j = s, j, s = 1, 2, \ldots, m \; ; \]

= 0 otherwise.

In matrix notation we write

\[ Y = X\beta + e \]

where

\[ E[ee'] = \begin{bmatrix}
\sigma_1^2 I_{n_1} & 0 \\
\vdots & \ddots \\
0 & \sigma_m^2 I_{n_m}
\end{bmatrix} \]

;
Let
\[ n_T = \sum_{j=1}^{k} n_j . \]

Then the MINQUE of \( \sigma_j^2 \), \( j = 1, 2, \ldots, m \) is given by (2.1.5) and the least squares estimator by (2.1.10).

If \( n_j = 1 \), \( j = 1, 2, \ldots, m \) then the MINQUE of \( \sigma_j^2 \), \( j = 1, 2, \ldots, m \) is given by (2.1.4).

For the case \( n_j = n \), \( j = 1, 2, \ldots, m \); and \( X_{ijk} = X_{rjk} \), \( i, r = 1, 2, \ldots, n \); an alternative procedure is given by Mandel (1964).

We write
\[ Y_i = X\beta_i + e_i \]

where
\[ Y_i \text{x} \times 1; \hspace{1em} X \text{mxp}; \hspace{1em} \beta_i \text{pxl}; \hspace{1em} e_i \text{mxl} . \]

Define
\[ \hat{\beta}_i = (X'X)^{-1} X'Y_i . \]
From (2.1.3) we have

\[ e = M g. \]

Summing over \( i = 1, 2, \ldots, n \) we have

\[ \frac{1}{n} \sum_{i=1}^{n} e_i = M g. \]

thus

\[ \frac{1}{n} \sum_{i=1}^{n} e_i^2 = \sum_{r=1}^{m} m_{jr} e_r. \]

Since \( M \) is idempotent

\[ m_{jj} = \sum_{r=1}^{m} m_{jr}^2. \]

and hence

\[ m_{jj}^2 = m_{jj} - \sum_{s=1}^{k} m_{js}^2. \]

Therefore

\[ \frac{1}{n} \sum_{i=1}^{n} e_i^2 = m_{jj} e_j^2 + \sum_{s=1}^{k} (g_s - 1) m_{js}^2. \]
Mandel (1964) suggests the estimator

\[ \hat{g}_{j}^* = \frac{1}{n-1} \sum_{i=1}^{n} \frac{\hat{e}_{ij}^2}{m_{jj}} \]  

(2.1.11)

ignoring the terms in \( m_{js}^2 \), \( j \neq s \). We show in Section III that
\( \hat{M} = I + O(\frac{1}{n}) \) and hence \( \hat{g}_{j}^* \) is similar to MINQUE for large \( n \). We note that \( \hat{g}_{j}^* \) cannot be negative.

In some heteroscedastic linear models the diagonal elements of the covariance matrix \( G \), \( g_{ii} \), \( i = 1, 2, ..., n \) are known functions of the explanatory variables \( x_{i} \), \( j = 1, 2, ..., p \) and an unknown \( s \times 1 \) vector parameter \( \Theta \). We write

\[ g_{i} = f(x_{i1}, x_{i2}, ..., x_{ip}; \Theta) \quad i = 1, 2, ..., n; \]

or

\[ g_{i} = f_{i}(\Theta). \]  

(2.1.12)

Then (2.1.3) becomes

\[ \hat{e} = \hat{M}f(\Theta) \]  

(2.1.13)

where \( f(\Theta) \) and \( \hat{e} \) are the \( n \times 1 \) vectors of elements \( f_{i}(\Theta) \) and \( \hat{e}_{i}^2 \), \( i = 1, 2, ..., n \). In some cases it may be possible to write (2.1.13) in
the form

\[ \hat{\mathbf{e}} = \mathbf{M} \hat{\mathbf{F}} \hat{\mathbf{\Theta}} ; \]  

(2.1.14)

or

\[ \hat{\mathbf{e}} = \mathbf{M} \hat{\mathbf{F}} \hat{\mathbf{\Theta}} + \mathbf{W} ; \]  

(2.1.15)

where \( \mathbf{F} \) is an \( n \times s \) matrix whose elements are functions of \( \mathbf{X} \) and
\( \mathbf{W} = \hat{\mathbf{e}} - E[\hat{\mathbf{e}}] \). From (2.1.15) the simple least squares estimator is given by

\[ \hat{\mathbf{\Theta}}^* = (\hat{\mathbf{MF}})' (\hat{\mathbf{MF}})^{-1} (\hat{\mathbf{MF}})' \hat{\mathbf{e}} . \]  

(2.1.16)

As shown by Griffiths (1971) the MINQUE of \( \hat{\mathbf{\Theta}} \) is given by

\[ \overline{\Theta} = (\mathbf{F}' \hat{\mathbf{M}} \mathbf{F})^{-1} \mathbf{F}' \hat{\mathbf{e}} . \]  

(2.1.17)

A form of the random coefficient model which was mentioned by Rubin (1950) and later discussed by Hildreth and Houck (1968) results in a heteroscedastic linear model. The diagonal elements of the covariance matrix have the form given by (2.1.12) and the equation (2.1.3) can be written in the form of (2.1.14) and (2.1.15).

The model is given as follows.

\[ y_i = \sum_{k=1}^{p} \gamma_{ik} x_{ik} \quad i = 1, 2, \ldots, n ; \]
where
\[ E[Y_{ik}] = \beta_k \quad i = 1, 2, \ldots, n; \]
\[ k = 1, 2, \ldots, p; \]

\[ Y_{ik} = \beta_k + \delta_{ik}, \quad E[\delta_{ik}] = 0; \]

\[ E[\delta_{ik}^2] = \alpha_k \quad i = 1, 2, \ldots, n \]
\[ k = 1, 2, \ldots, p \]

and for any \( i, m \) and \( j \neq k \)

\[ E[\delta_{ij}\delta_{mk}] = 0. \]

We rewrite the model as
\[ y_i = \sum_{k=1}^{p} x_{ik} \beta_k + e_i \quad i = 1, 2, \ldots, n; \]

where
\[ e_i = \sum_{k=1}^{p} x_{ik} \delta_{ik}, \quad E[e_i] = 0, \]

\[ E[e_i^2] = \sum_{k=1}^{p} \alpha_k x_{ik}^2, \quad (2.1.18) \]
In matrix notation we write

\[ Y = X\beta + e \]

and

\[ E[ee'] = G, \]

where \( G \) is a diagonal matrix with \( i^{th} \) diagonal elements \( f_i(\alpha) = \sum_{k=1}^{p} x_{ik}^2 \alpha_k \). We let \( \hat{X} \) represent the matrix whose elements are the squares of the elements of \( X \). Equations (2.1.14) and (2.1.15) then have the form

\[ \hat{e} = \hat{M}\hat{X}\hat{\alpha} \]  \hspace{1cm} (2.1.19)

and

\[ \hat{e} = \hat{M}\hat{X}\hat{\alpha} + \hat{W} \]  \hspace{1cm} (2.1.20)

Hildreth and Houck give the simple least squares estimator \( \hat{\alpha} \) obtained from (2.1.17)

\[ \hat{\alpha} = \left[ (\hat{M}\hat{X})'(\hat{M}\hat{X}) \right]^{-1} (\hat{M}\hat{X})' \hat{e}. \]  \hspace{1cm} (2.1.21)

They also demonstrate that the estimator
\[ \beta^* = (X' G^{-1} X) ^{-1} X' G^{-1} Y \]  

(2.1.22)

is consistent. The matrix \( \hat{G} \) is obtained from \( G \) by replacing \( \alpha \) by \( \hat{\alpha} \).

Theil (1971, p. 623) suggests that a more efficient estimator might be obtained using

\[ \hat{\alpha}^* = [ (\hat{M} \hat{X})' \Omega^{-1} (\hat{M} \hat{X}) ]^{-1} (\hat{M} \hat{X})' \Omega^{-1} \hat{e} \]  

(2.1.23)

where \( \hat{\Omega} \) is an estimator of \( \Omega \) the covariance matrix of \( \hat{e} \). If the \( e_i, \ i = 1, 2, \ldots, n \) are normally distributed then \( \text{Var} \ [e_i^2] = 2 [ E[e_i^2] ]^2 \). Under the normality assumption Theil suggests that since to \( O(\frac{1}{n}) \) the covariance matrix \( \Omega \) is diagonal these diagonal elements can be estimated by \( 2 [ \sum_{k=1}^{p} \hat{\alpha}_k x_{ik}^2 ]^2 \). Thus \( \hat{\Omega} \) is a diagonal matrix whose elements are easily obtained from the diagonal elements of \( \hat{G} \).

An alternative estimator of \( \alpha \) given by Hildreth and Houck is given by

\[ \tilde{\alpha} = (X' MX)^{-1} X' \hat{e} \]  

(2.1.24)

Froehlich (1971) and Griffiths (1971) have shown that \( \tilde{\alpha} \) is the MINQUE for \( \alpha \). This estimator of \( \alpha \) may also be used to construct a consistent estimator of \( \beta \) as in (2.1.21).

The estimators of \( \alpha \) given by (2.1.21), (2.1.23) and (2.1.24) may contain negative elements. Hildreth and Houck (1968) and Griffiths (1971)
have suggested alternative procedures which restrict $\alpha$ to contain non-negative elements. Hildreth and Houck demonstrate that the estimator

$$\bar{\alpha}_k = \max \left\{ \alpha_k^*, 0 \right\} \quad k = 1, 2, \ldots, p;$$

is biased but has smaller mean square error than $\alpha_k^*$, $k = 1, 2, \ldots, p$. Hildreth and Houck also give a quadratic programming approach which minimizes

$$(\hat{e} - MX\alpha)' (\hat{e} - MX\alpha)$$

subject to

$$\alpha \geq 0.$$ 

A study of the performance of these estimators for a particular example is given by Griffiths (1971).

The following random coefficient model was discussed by Fisk (1967)

$$y_{ij} = X'_j \beta_{ij} \quad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, k;$$

$$X'_j l \times p; \quad \beta_{ij} p \times 1.$$ 

The $\beta_{ij}$ are independent and identically distributed random vectors with mean $\beta$ and covariance matrix.
We rewrite the model as follows.

\[ y_{ij} = X_j' \beta + X_j' (\beta_{ij} - \beta), \]

\[ = X_j' \beta + e_{ij} \quad i = 1, 2, \ldots, n \]

\[ j = 1, 2, \ldots, k; \]

where

\[ e_{ij} = X_j' (\beta_{ij} - \beta); \]

\[ E[e_{ij}^2] = X_j' VX_j = \sigma_j^2. \]

Let

\[ \bar{y}_i = \frac{1}{n} \sum_{i=1}^{n} y_{ij}, \quad \bar{e}_j = \frac{1}{n} \sum_{i=1}^{n} e_{ij}, \quad j = 1, 2, \ldots, k. \]

In matrix notation we have

\[ \bar{Y} = X \beta + \bar{e} \]

where

\[ \bar{Y}kx1; \quad Xkxp; \quad \bar{e}kx1. \]
The simple least squares estimator of $\beta$ is given by

$$\hat{\beta} = (X'X)^{-1} X'Y.$$ 

Unbiased estimators of the $\sigma_j^2$, $j = 1, 2, \ldots, k$ are provided by

$$S_j^2 = \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 / (n-1) \quad j = 1, 2, \ldots, k.$$

The $\frac{p(p+1)}{2}$ distinct elements of $V$ are estimated as follows. Let the vector $g$ represent the distinct elements of $V$ written in lexicographical order. Thus

$$g' = (V_{11}, V_{12}, \ldots, V_{1p}, V_{22}, V_{23}, \ldots, V_{2p}, \ldots, V_{pp}).$$

Let

$$X'_j V X'_j = Z'(j) g$$

where

$$Z'(j) = (x_{j1} x_{j1}, x_{j1} x_{j2}, \ldots, x_{j1} x_{jp}, x_{j2} x_{j2}, \ldots, x_{jp} x_{jp}).$$

Let
\[ Z = \begin{bmatrix} Z'(1) \\ \vdots \\ Z'(x) \end{bmatrix}, \quad S = \begin{bmatrix} S_1^2 \\ \vdots \\ S_k^2 \end{bmatrix} \]

Thus
\[ E[S] = Zg. \]

We write
\[ S = Zg + W \]

where
\[ W = S - E[S]. \]

The simple least squares estimator of \( g \) is given by
\[ ^\wedge g = (Z'Z)^{-1} Z'S. \]

If we let
\[ E[WW'] = \Omega \]

then a more precise estimator of \( g \) is given by
\[ ^* g = (Z'\Omega^{-1}Z)^{-1} (Z'\Omega^{-1}S). \]
\[ \Omega \text{ may be estimated from estimates of the sample moments of the } S_j^2, \]
\[ j = 1, 2, \ldots, k. \text{ For instance if the } \beta_{ij} \text{ are normally distributed the diagonal elements of } \Omega \text{ have the form } 2[\sigma_j^2]^2. \text{ In the case of normality } \hat{\beta} \text{ and } \hat{\sigma}^2 \text{ are asymptotically normal.} \]

Another form of heteroscedasticity is that used by Prais and Houthakker (1955) in the study of consumer behavior. The model is given by:

\[ y_i = \sum_{k=1}^{p} \beta_k x_{ik} + e_i \quad i = 1, 2, \ldots, n; \]

(2.1.25)

\[ E[e_i] = 0; \quad E[e_i^2] = \alpha \left( \sum_{k=1}^{p} \beta_k x_{ik} \right)^2; \quad i = 1, 2, \ldots, n \]
\[ \alpha \text{ an unknown constant} \]
\[ E[e_i e_j] = 0 \quad i \neq j; \]

In matrix notation we write:

\[ Y = X\beta + e; \]

\[ Y_{nx1}; \quad X_{nxp}; \quad \beta_{kx1}; \quad e_{nx1}; \]

\[ E[e] = 0; \]

\[ E[ee'] = \alpha G; \]
where $G$ is a diagonal matrix with diagonal elements given by

$$g_i = \left( \sum_{k=1}^{p} \beta_k x_{ik} \right)^2.$$

Theil (1971, p. 25) suggests that simple least squares be used to obtain an estimator $\hat{\beta}$ of $\beta$ and that $\hat{g}_i = \left( \sum_{k=1}^{p} \hat{\beta}_k x_{ik} \right)^2$ be used to estimate $g_i = \left( \sum_{k=1}^{p} \beta_k x_{ik} \right)^2$. A second stage estimator of $\beta$ is given by

$$\beta^* = \left( X' G^{-1} X \right)^{-1} \left( X' G^{-1} Y \right).$$

where $G$ is obtained from $G$ by replacing $g_i$ by $\hat{g}_i$, $i = 1, 2, \ldots, n$. If desired $\alpha$ may be estimated from the second stage residuals using

$$\alpha^2 = \frac{n}{\sum_{i=1}^{n} \left[ y_i - \sum_{k=1}^{p} \hat{\beta}_k x_{ik} \right]^2} \left( \sum_{k=1}^{p} \hat{\beta}_k x_{ik} \right)^2 \frac{1}{n-p}.$$
B. The Maximum Likelihood Procedure

For the class $\mathcal{L}$ of linear models given by (1.1.1) under the assumption of normality the logarithm of the likelihood function is given by

$$
\log L = c - \frac{1}{2} \sum_{i=1}^{n} \log g_i - \frac{1}{2} (y - XB)' G^{-1} (y - XB) 
$$

where $c$ is a real constant and $g_i$, $i = 1, 2, ..., n$ are the diagonal elements of $G$.

Under certain regularity conditions that we shall discuss in Section III equations for the maximum likelihood estimators are obtained by differentiating (2.2.1) with respect to the elements of $\beta$ and unknown parameters of $g$ and setting the derivatives equal to zero. The existence of the maximum likelihood estimator and its properties are discussed in Section III. We assume in this section that the equations for the maximum likelihood estimator exist and have a unique solution which corresponds to the maximum of the likelihood function. We consider some specific examples of heteroscedasticity where the maximum likelihood procedure has been employed.

The following model was discussed by Rutemiller and Bowers (1968).

$$
y_i = \beta_0 + \sum_{k=1}^{p} x_{ik} \beta_k + e_i, \quad i = 1, 2, ..., n;
$$
\[ E[e_i] = 0 ; \quad E[e_i^2] = \gamma_0 + \sum_{k=1}^{p} \gamma_k x_{ik} = \xi_i ; \]

\[ E[e_i e_j] = 0 \quad i \neq j \quad i, j = 1, 2, \ldots, n . \]

Rutemiller and Bowers (1968) used the "method of scoring" as described by C. R. Rao (1965a, p. 302). This method is an iterative technique which begins with a first order Taylor Series expansion of the first derivatives of log \( L \), \( \frac{\partial \log L}{\partial \beta_j} \), \( \frac{\partial \log L}{\partial \gamma_j} \), with respect to the unknown parameters \( \beta_j \), \( \gamma_j \), evaluated at preliminary estimates \( \hat{\beta}_{0j} \), \( \hat{\gamma}_{0j} \), \( j = 0, 1, 2, \ldots, p \). Equations for the changes to be made to the preliminary estimates \( (\beta_j - \hat{\beta}_{0j}) \), \( (\gamma_j - \hat{\gamma}_{0j}) \), \( j = 0, 1, 2, \ldots, p \) are obtained by using the fact that

\[
\left( \frac{\partial \log L}{\partial \beta_j} \right) = 0 \quad \text{and} \quad \left( \frac{\partial \log L}{\partial \gamma_j} \right) = 0
\]

when evaluated at the values of \( \beta_j \) and \( \gamma_j \), \( j = 0, 1, 2, \ldots, p \) that maximize \( L \). The quantities

\[
\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} , \quad \frac{\partial^2 \log L}{\partial \beta_j \partial \gamma_k} , \quad \frac{\partial^2 \log L}{\partial \gamma_j \partial \gamma_k} ,
\]
are replaced by their expectations.

i.e.

\[
\left( \frac{\partial \log L}{\partial \beta_i} \right)_{\beta = \beta_0}^\wedge + \sum_{j=1}^{p} E \left[ \frac{\partial^2 \log L}{\partial \beta_j \partial \beta_i} \right] (\beta_j - \beta_{0j})
\]

\[
\gamma = \gamma_0 + \sum_{j=1}^{p} E \left[ \frac{\partial^2 \log L}{\partial \gamma_j \partial \beta_i} \right] (\gamma_j - \gamma_{0j})
\]

Replacing \( \frac{\partial^2 \log L}{\partial \gamma_j \partial \beta_i} \) and \( \frac{\partial^2 \log L}{\partial \beta_i \partial \beta_j} \) \( i, j = 0, 1, 2, ..., p \) by their expectations and \( \frac{\partial \log L}{\partial \beta_j} \) by zero on the left hand side we obtain

\[
0 = \left( \frac{\partial \log L}{\partial \beta_i} \right)_{\beta = \beta_0}^\wedge + \sum_{j=1}^{p} E \left[ \frac{\partial^2 \log L}{\partial \beta_j \partial \beta_i} \right] (\beta_j - \beta_{0j})
\]

\[
\gamma = \gamma_0 + \sum_{j=1}^{p} E \left[ \frac{\partial^2 \log L}{\partial \gamma_j \partial \beta_i} \right] (\gamma_j - \gamma_{0j})
\]
In matrix notation we write the system

\[ 0 = d + A \begin{bmatrix} \beta - \beta \\ \gamma - \gamma \end{bmatrix} \]

and solving we obtain

\[ \begin{bmatrix} \beta - \beta \\ \gamma - \gamma \end{bmatrix} = -A^{-1} d ; \]

where \( d \) is the \((2p+2)\times1\) vector of elements \( (\tau, \eta) \) and \( A \) is the \(2(p+1) \times 2(p+1)\) matrix of elements

\[
\begin{pmatrix}
\frac{\partial \log L}{\partial \beta_j} & \frac{\partial \log L}{\partial \gamma_j} \\
\frac{\partial \log L}{\partial \beta_k} & \frac{\partial \log L}{\partial \gamma_k}
\end{pmatrix}
\]

\( j = 0, 1, 2, \ldots, p \) and \( A \) is the

\[
E \left[ \frac{\partial^2 \log L}{\partial \gamma_j \partial \gamma_k} \right], \quad E \left[ \frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} \right], \quad E \left[ \frac{\partial^2 \log L}{\partial \gamma_j \partial \beta_k} \right].
\]
This process is then repeated by replacing \( \hat{\beta}_0 \) and \( \hat{\gamma}_0 \) by the new estimators \( \hat{\beta}_1 = \hat{\beta}_0 + \hat{\beta} - \beta_0 \), \( \hat{\gamma}_1 = \hat{\gamma}_0 + \gamma - \gamma_0 \). This process is continued until the change in the value of the estimates becomes sufficiently small.

A second example is provided by a form of the nonparametric unequal variance model discussed in Subsection A.

\[
y_{ij} = \sum_{k=1}^{p} x_{ijk} \beta_k + e_{ij} \quad i = 1, 2, \ldots, n \\
j = 1, 2, \ldots, m;
\]

\[
E[e_{ij}] = 0; \quad E[e_{ij}e_{rj}] = \sigma_j^2
\]

\[
E[e_{ij}e_{rs}] = 0 \quad s \neq j, \quad s, j = 1, 2, \ldots, m;
\]

The logarithm of the likelihood function is given by

\[
\log L = c - \frac{1}{2} \sum_{j=1}^{m} \left[ \frac{n}{\Sigma} \frac{(y_{ij} - \sum_{k=1}^{p} x_{ijk} \beta_k)^2}{\sigma_j^2} + n \log \sigma_j^2 \right] + n \log \sigma_j^2 \quad (2.2.2)
\]

Differentiating \( \log L \) with respect to \( \beta_k \), \( k = 1, 2, \ldots, p \) and \( \sigma_j^2 \), \( j = 1, 2, \ldots, m \) and setting the derivatives equal to zero we obtain
\begin{align*}
\frac{\partial \log L}{\partial \sigma_j^2} &= \frac{1}{2} \sum_{i=1}^{n} \frac{(y_{ij} - \sum_{k=1}^{p} x_{ijk} \beta_k)^2}{[\sigma_j^2]^2} - \frac{1}{2} \frac{n}{\sigma_j^2} = 0 \\
&\quad j = 1, 2, \ldots, m; \quad (2.2.3)
\end{align*}

\begin{align*}
\frac{\partial \log L}{\partial \beta_r} &= -\frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{(y_{ij} - \sum_{k=1}^{p} x_{ijk} \beta_k)(-2 x_{ijr})}{\sigma_j^2} = 0 \\
&\quad r = 1, 2, \ldots, n. \quad (2.2.4)
\end{align*}

These equations cannot be solved for \( \sigma_j^2 \), \( j = 1, 2, \ldots, m \) independently of \( \beta \).

Ehrenberg (1950) suggested the approximate solution

\begin{equation}
\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} (y_{ij} - \sum_{k=1}^{p} x_{ijk} \hat{\beta}_k)^2}{n} \quad (2.2.5)
\end{equation}

\( j = 1, 2, \ldots, m \)

where \( \hat{\beta} \) is the simple least squares estimator of \( \beta \).

Hartley and Jayatillake (1971) replace \( \sigma_j^2 \) in (2.2.2) by

\begin{equation}
\frac{n}{\sum_{i=1}^{n} (y_{ij} - \sum_{k=1}^{p} x_{ijk} \beta_k)^2} / n \quad j = 1, 2, \ldots, m.
\end{equation}
Equation (2.2.2) becomes

\[
\log L = c - \frac{1}{2} \sum_{j=1}^{m} \left[ n + n \log \left( \frac{1}{n} \sum_{i=1}^{n} \frac{(y_{ij} - \sum_{k=1}^{P} x_{ijk} \beta_k)^2}{n} \right) \right].
\]

Hartley and Jayatillake then employ the method of steepest descent to maximize \( \log L \) with respect to \( \beta \). In the case \( n = 1 \) a positive lower bound \( \delta_j^2 \) is assumed for \( \hat{\sigma_j^2} \) so that \( \hat{\sigma_j^2} \) is given by

\[
\hat{\sigma_j^2} = \left( y_j - \sum x_{jk} \beta_k \right)^2 \quad \left( y_j - \sum x_{jk} \beta_k \right)^2 > \delta_j^2 ;
\]

\[
= \delta_j^2 \quad \text{otherwise} \quad j = 1, 2, \ldots, m.
\]

For \( m \) a finite number and as \( n \) becomes infinite Hartley and Jayatillake show that the estimator of \( \beta \) is consistent and asymptotically efficient. For the case \( n = 1 \) and \( m \) approaching infinity they claim it may not be meaningful to assume that all \( \sigma_j^2 \), \( j = 1, 2, \ldots, m \) are different. Therefore one might assume only \( k \) different variances \( \sigma_j^2 \), \( j = 1, 2, \ldots, k \) and proceed as in the case of fixed \( m \) and \( n \) approaching infinity.

For the random coefficient model discussed by Fisk (1967) outlined in Subsection A the logarithm of the likelihood function is given by

\[
\log L = c - \frac{n}{2} \sum_{j=1}^{m} \log X_j'X_j + \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(y_{ij} - X_j'\beta)^2}{X_j'X_j}.
\]
Fisk states that the asymptotic distribution of \( \sqrt{n} \begin{bmatrix} \beta - \hat{\beta} \\ g - \hat{g} \end{bmatrix} \) is normal with mean 0 and covariance matrix

\[
\begin{bmatrix}
E \left[ \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right]_{ij}^{-1} & E \left[ \left( \frac{\partial \log L}{\partial \theta_i} \right) \left( \frac{\partial \log L}{\partial \theta_j} \right) \right]_{ij} \left( E \left[ \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right]_{ij} \right)^{-1}
\end{bmatrix};
\]

where \( \hat{\beta}, \hat{g} \) are the maximum likelihood estimators of \( \beta \) and \( g \), \( \theta = \begin{bmatrix} \beta \\ g \end{bmatrix} \), and \( g \) is the vector of unknown elements of \( V \).

The likelihood equations for the random coefficient model of Hildreth and Houck (1968) discussed in Subsection A were given by Rubin (1950, p. 419-421). Rubin suggested a Lagrangian approach to assure that the estimates of the variance remain nonnegative. An analytic expression for the maximum likelihood estimators of the unknown parameters does not exist.

In the following section we develop a least squares procedure which yields estimators of \( \beta \) and \( \theta \) that are asymptotically normal and efficient. We then show that an iterative procedure that uses the least squares procedure can be used to obtain a sequence of estimators that converge to a local maximum of the likelihood function.
III. EXISTENCE AND PROPERTIES OF ESTIMATORS FOR A SUBCLASS OF $\mathcal{L}$

In this section we obtain estimators and some of their properties for a subclass of the class $\mathcal{L}$ of linear models of Section I. We concentrate on models with independently distributed errors. Many of the examples given in Section II belong to this subclass.

In Subsection A we list the principal assumptions. Under the assumption of normality we show in Subsection B that the asymptotic distribution of the maximum likelihood estimator is normal. After demonstrating in Subsection C the existence of a consistent simple least squares estimator a joint least squares procedure is developed in Subsection D. This procedure uses a consistent preliminary estimator to obtain an improved estimator. The improved estimator is shown to have the same asymptotic distribution as the maximum likelihood estimator. In Subsection E we show that an iterative procedure can be established using the joint least squares procedure to obtain a sequence of estimators that converges to a local maximum of the likelihood function.

A. The Principal Assumptions

The eight principal assumptions that we require to establish the results of this section are summarized below.

Assumption 1.

$$y_t = \sum_{k=1}^{p} \beta_{tk}^O x_t + e_t , \quad t = 1, 2, \ldots, n$$  \hspace{1cm} (3.1.1)
where the following are true.

\[ \{X_t\}_{t=1}^n \] is a sequence of observed fixed \( p \times 1 \) vectors in \( \mathbb{R}^p \)

whose elements are uniformly bounded in \( \mathbb{R}^p \). \( \beta^0 = \{\beta_k^0\}_{k=1}^p \) is an unknown \( p \times 1 \) constant vector. \( \{y_t\}_{t=1}^n \) is a sequence of observed random variables in \( \mathbb{R} \). \( \{e_t\}_{t=1}^n \) is a sequence of unobserved independent random variables in \( \mathbb{R} \) with corresponding densities \( f(e_t; X_t, \alpha^0, \beta^0) \) where \( \alpha^0 \) is an unknown \( r \times 1 \) constant vector, \( \Theta^0 = \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \in S_0 \subset \Theta \subset \mathbb{R}^{p+r} \), \( S_0 \) is a bounded open sphere, \( \Theta \) is closed and bounded and \( n, r \) and \( p \) are integers such that \( n > r+p \).

\[
\begin{align*}
E[e_t^2] &= g(X_t; \alpha^0, \beta^0), & t = 1, 2, \ldots, n. & (3.1.2) \\
E[e_t^j] &= 0, & j = 1, 3, 5, 7; t = 1, 2, \ldots, n. & (3.1.3) \\
E[e_t^j] &< M < \infty, & M a positive real number, & j = 2, 4, 6, 8; t = 1, 2, \ldots, n. & (3.1.4)
\end{align*}
\]

\( \Theta^0 = \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \), \( \beta^0 \) is a \( q \times 1 \) vector whose elements are a subset of the elements of \( \beta^0 \), \( q \) an integer such that \( 0 \leq q \leq p \), \( 1 \leq q+r \leq p+r \), \( \Theta^0 \) is contained in \( S_0 \) where \( S_0 \subset \Theta \subset \mathbb{R}^{q+r} \). We shall assume without loss of generality that \( \beta^0 \) contains the first \( q \) elements of \( \beta_0 \). The function \( g(X_t; \Theta) \) is a real valued continuous function of \( \Theta \) on \( \Theta \) for all \( X_t \), \( t = 1, 2, \ldots, n \).
It is assumed that

\[ 0 < L_g < g(X_t; \theta) < M_g < \infty \]  

(3.1.5)

for all \( X_t, t = 1, 2, \ldots, n \); and for all \( \theta \in \Theta \) where \( L_g \) and \( M_g \) are positive real constants.

In vector notation we write

\[ Y = X\beta^0 + e; \]  

(3.1.6)

\[ E[e] = 0 \quad E[ee'] = G, \]  

where \( Y_{nx1}, X_{nxp}, \beta_{px1}, e_{nx1} \),

and \( G \) is an \( nxn \) diagonal matrix with diagonal elements

\[ \{g(X_t; \theta^0)\}^n_{t=1}. \]

The \( nx1 \) vector of diagonal elements of \( G \) will be denoted by \( g \). The matrices \( \lim_{n \to \infty} \frac{X'X}{n} \) and \( \lim_{n \to \infty} \frac{X'GX}{n} \) are finite and positive definite.

We sometimes denote the functions \( g(X_t; \theta), f(e_t; X_t, \theta) \) or equivalently \( g(X_t; \alpha, \beta), f(e_t; X_t, \alpha, \beta) \) by \( g_t(\theta), f_t(e_t; \theta), \)

\( g_t(\alpha, \beta), f_t(e_t; \alpha, \beta) \). The derivatives of these functions with respect to an element \( \theta_j \) or \( \theta_j \) of \( \Theta \) or \( \Theta \) evaluated at a point \( \Theta^* \) or \( \theta^* \) given by

\[ \left( \frac{\partial g_t(\theta)}{\partial \theta_j} \right) \theta = \theta^*, \quad \left( \frac{\partial f_t(e_t; \theta)}{\partial \theta_j} \right) \theta = \theta^* \]
will be denoted by

\[
\frac{\partial g_t(\theta^*)}{\partial \theta_j} \quad \text{and} \quad \frac{\partial f_t(e_t; \theta^*)}{\partial \theta_j}.
\]

The densities of \( \{e_t\}_{t=1}^n \), \( \{f_t(e_t; \theta)\}_{t=1}^n \) will also be denoted by \( \{f_t(y_t; \theta)\}_{t=1}^n \) where it is understood that \( e_t \) has been replaced by \( (y_t - X_t' \beta) \).

Assumption 2. For all \( \theta \in S_1 \), \( S_1 \supset \Theta \), \( S_1 \) a bounded open sphere, the first, second and third order partials of \( \{g(X_t; \theta)\}_{t=1}^n \) with respect to the elements of \( \Theta \) exist and are uniformly bounded. The first and second partials are continuous functions of the elements of \( \Theta \) for all \( \theta \in S_1 \).

Assumption 3. The quantity

\[
Q = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^n \left | g(X_t; \theta) - g(X_t; \theta^0) \right |^2
\]

exists and has a unique minimum on the \( S_0 \) of Assumption 1 at \( \theta = \Theta^0 \) for all \( X_t, t = 1, 2, \ldots, n \).

Assumption 4. The matrices \( \lim_{n \to \infty} \left( \frac{H^T \psi H}{n} \right), \lim_{n \to \infty} \left( \frac{F' \tilde{G} F}{n} \right), \lim_{n \to \infty} \left( \frac{H^T \psi^{-1} H}{n} \right) \) and \( \lim_{n \to \infty} \left( \frac{F' \psi F}{n} \right) \) are finite and positive definite for all \( \theta \in S_1 \) where \( S_1 \) is given in Assumption 2. \( H, F \) and \( \psi \) evaluated at \( \Theta^0 \) are given below. The matrices \( H_\Theta, F_\Theta, \psi_\Theta \) will denote \( H, F \) and \( \psi \) with \( \Theta^0 \) replaced by \( \Theta \).
where \( G \) is the matrix whose elements are the squares of the elements of \( G \) given in (3.1.6) and

\[
\begin{align*}
H_1 &= \begin{bmatrix}
\frac{\partial g_1(\varphi^0)}{\partial \alpha_1} & \cdots & \frac{\partial g_1(\varphi^0)}{\partial \alpha_r} \\
\vdots & \ddots & \vdots \\
\frac{\partial g_n(\varphi^0)}{\partial \alpha_1} & \cdots & \frac{\partial g_n(\varphi^0)}{\partial \alpha_r}
\end{bmatrix} \\
H_2 &= \begin{bmatrix}
\frac{\partial g_1(\varphi^0)}{\partial \beta_1} & \cdots & \frac{\partial g_1(\varphi^0)}{\partial \beta_q} & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \ddots & \ddots & \ddots \\
\frac{\partial g_n(\varphi^0)}{\partial \beta_1} & \cdots & \frac{\partial g_n(\varphi^0)}{\partial \beta_q} & 0 & \cdots & 0
\end{bmatrix}
\end{align*}
\]

(3.1.8)
Assumption 5. The \( \{e_t\}_{t=1}^n \) are normally distributed.

Assumption 6. There does not exist \( \left[ \begin{array}{c} \alpha' \\ \beta' \end{array} \right] \) and \( \left[ \begin{array}{c} \alpha'' \\ \beta'' \end{array} \right] \) in \( \Theta \) such that \( f_t(e_t; \alpha', \beta') = f_t(e_t; \alpha'', \beta'') \) for all \( e_t, \ t = 1, 2, \ldots, n \).

Assumption 7. Preliminary estimators, \( \left[ \begin{array}{c} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{array} \right] \) exist

where \( \left[ \begin{array}{c} \hat{\alpha} - \alpha_0 \\ \hat{\beta} - \beta_0 \end{array} \right] = O_p(\frac{1}{\sqrt{n}}) \), and where \( \left[ \begin{array}{c} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{array} \right] \) is in \( \Theta \).

Assumption 8. The preliminary estimators \( \left[ \begin{array}{c} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{array} \right] \) of Assumption 7 lie in the interior of a convex bounded subset \( T \) of \( \Theta \) where \( T \subset \Theta \), such that \( \ell(\hat{\alpha}_0, \hat{\beta}_0) < \inf_{(\alpha, \beta) \in \Theta} \{ \ell(\alpha, \beta) ; \left[ \begin{array}{c} \alpha \\ \beta \end{array} \right] \} \) a boundary point of \( T \),

where

\[
\ell(\alpha, \beta) = \frac{1}{n} \sum_{t=1}^{n} \log g(x_t; \alpha, \beta) + \frac{1}{N} \sum_{t=1}^{N} \frac{(y_t - X_t^T \beta)^2}{g(x_t; \alpha, \beta)}
\]

\( \ell(\alpha, \beta) \) is related to the logarithm of the likelihood function for \( \{e_t\}_{t=1}^n \) under the normality assumption.
and there does not exist \[ \begin{bmatrix} \alpha' \\ \beta' \end{bmatrix} \] and \[ \begin{bmatrix} \alpha'' \\ \beta'' \end{bmatrix} \] such that
\[
\frac{\partial l(\alpha', \beta')}{\partial \alpha_j} = \frac{\partial l(\alpha'', \beta'')}{\partial \alpha_j}, \quad j = 1, 2, \ldots, r;
\]
\[
\frac{\partial l(\alpha', \beta')}{\partial \beta_k} = \frac{\partial l(\alpha'', \beta'')}{\partial \beta_k}, \quad k = 1, 2, \ldots, p;
\]
and \( l(\alpha', \beta') = l(\alpha'', \beta'') \).

For notational convenience, we shall assume throughout the remainder of Section III that all \( p \) elements of \( \beta^0 \) are elements of \( \beta^0 \), and hence that \( \theta^0 = \phi^0 \) and \( q = p \). The results that we obtain under this convention are valid for all \( q \) subject to \( 0 \leq q \leq p \) and \( 1 \leq q \leq p + r \).

**B. The Likelihood Function and Properties of the Maximum Likelihood Estimator**

We begin by giving the logarithm of the likelihood function and the likelihood equations. By Assumptions 1 and 5 we have that the logarithm of the likelihood function for a given sample \( \{y_t\}_{t=1}^n \) at \( \{x_t\}_{t=1}^n \) is given by

\[
\log L(Y; X, \theta) = c - \frac{1}{2} \sum_{t=1}^{n} \log \epsilon_t(\alpha, \beta) - \frac{1}{2} \sum_{t=1}^{n} \frac{(y_t - x_t' \beta)^2}{\epsilon_t(\alpha, \beta)}, \quad (3.2.1)
\]

where \( c \) is a real constant. We define the likelihood equations to be the equations obtained by setting equal to zero the derivatives of \( \log L(Y; X, \theta) \) with respect to the elements of \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \). For this case the equations are given below.
\[ \frac{\partial \log L}{\partial \alpha_k} = -\frac{1}{2} \sum_{t=1}^{n} \frac{\partial g_t(\alpha, \beta)}{\partial \alpha_k} \left( \frac{1}{g_t(\alpha, \beta)} \right) + \]

\[ \frac{1}{2} \sum_{t=1}^{n} (y_t - x_t \beta)^2 \frac{\partial g_t(\alpha, \beta)}{\partial \alpha_k} \left( \frac{\partial g_t(\alpha, \beta)}{\partial \alpha_k} \right) = 0 \]

\[ k = 1, 2, \ldots, r. \]

\[ (3.2.2) \]

\[ \frac{\partial \log L}{\partial \beta_j} = -\frac{1}{2} \sum_{t=1}^{n} \frac{\partial g_t(\alpha, \beta)}{\partial \beta_j} \left( \frac{1}{g_t(\alpha, \beta)} \right) + \]

\[ \frac{1}{2} \sum_{t=1}^{n} (y_t - x_t \beta)^2 \frac{\partial g_t(\alpha, \beta)}{\partial \beta_j} \left( \frac{\partial g_t(\alpha, \beta)}{\partial \beta_j} \right) + \]

\[ \sum_{t=1}^{n} (y_t - x_t \beta) x_{tj} \left( \frac{\partial g_t(\alpha, \beta)}{\partial \beta_j} \right) = 0 \]

\[ j = 1, 2, \ldots, p. \]

\[ (3.2.3) \]

In matrix notation these equations are given by

\[ H_{\hat{\theta}} \psi_{\hat{\theta}}^{-1} z_{\hat{\theta}} = 0 \]

\[ (3.2.4) \]
where

\[ H_\theta = \begin{bmatrix} 0 & X \\ H_{\theta 1} & H_{\theta 2} \end{bmatrix} , \quad \psi_\theta = \begin{bmatrix} G_\theta & 0 \\ 0 & 2G_\theta \end{bmatrix} , \]

\( G_\theta \) is the matrix whose elements are the squares of the elements of \( G_\theta \);

\[ H_{\theta 1} = \begin{bmatrix} \frac{\partial g_1(\theta)}{\partial \alpha_1} & \ldots & \frac{\partial g_1(\theta)}{\partial \alpha_r} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n(\theta)}{\partial \alpha_1} & \ldots & \frac{\partial g_n(\theta)}{\partial \alpha_r} \end{bmatrix} , \]

\[ H_{\theta 2} = \begin{bmatrix} \frac{\partial g_1(\theta)}{\partial \beta_1} & \ldots & \frac{\partial g_1(\theta)}{\partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n(\theta)}{\partial \beta_1} & \ldots & \frac{\partial g_n(\theta)}{\partial \beta_p} \end{bmatrix} , \]
\[ Z_\theta = \begin{bmatrix} Y - X\beta \\ \vdots \\ e - g(\theta) \end{bmatrix} ; \quad e = \begin{bmatrix} (y_1 - X_1\beta)^2 \\ \vdots \\ (y_n - X_n\beta)^2 \end{bmatrix} \]

\[ g(\theta) = \begin{bmatrix} g_1(\alpha, \beta) \\ \vdots \\ g_n(\alpha, \beta) \end{bmatrix} \]

By Assumption 1, \( g(X_t; \alpha, \beta) \) is continuous in \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \) and

\[ 0 < L_g < g(X_t; \alpha, \beta) < M_g < \infty \] for all \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \Theta \) and for all \( \{X_t\}_{t=1}^n \). Therefore \( \log L(Y; X, \theta) \) given by (3.2.1) is continuous in \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \). Since \( \Theta \) is closed and bounded \( \log L(Y; X, \theta) \) has a maximum in \( \Theta \).

Wald (1949) demonstrated that under certain conditions the value of the parameter \( \theta \) that maximizes the likelihood function is a strongly consistent estimator of the true value of \( \theta \). One of the conditions requires that the random variables be independent and identically distribu-
ed. Since under Assumption 1 the densities of the \( \{e_t\}^n_{t=1} \),
\( \{f(e_t; X_t, \theta)\}^n_{t=1} \) or \( \{f(y_t; X_t, \theta)\}^n_{t=1} \) depend on the observed vector \( X_t \),
they are not identically distributed. Silvey (1961) has shown that if
certain conditions are met the maximum likelihood estimator for dependent
random variables is weakly consistent for \( \theta^0 \). We use Silvey's result to
show that under Assumptions 1, 5 and 6 the maximum likelihood estimator is
weakly consistent. We first prove several lemmas.

The following lemma was given by Wald (1949).

**Lemma 3.2.1.** Under Assumptions 1, 5 and 6 for any \( \theta \neq \theta^0 \), \( \theta^0 \) the
ture value of \( \theta \),

\[
E_{\theta^0} \left[ \log L(Y; X, \theta) \right] < E_{\theta^0} \left[ \log L(Y; X, \theta^0) \right]
\]

for all \( \theta \in \Theta \) where \( L(Y; X, \theta) \) is the likelihood of the sample given
by \( Y_{nx1} = \{y_t\}^n_{t=1} \), \( X_{nxp} = \{X_t\}^n_{t=1} \). The expectations are taken
with respect to \( \theta^0 \).

**Proof.** This proof is similar to the proof of Lemma 1 in Wald (1949).

Under Assumption 5

\[
L(Y; X, \theta) = \left( \frac{1}{\sqrt{2\pi}} \right)^{-n} \pi^{\frac{-1}{2}} \prod_{t=1}^{n} \exp \left\{ \frac{-1}{2} (y_t - X_t^\theta)^2 / \sigma_t(\theta) \right\}
\]
\[ \log L(Y; X, \theta) = c - \frac{1}{2} \sum_{t=1}^{n} \log g_t(\theta) - \frac{1}{2} \sum_{t=1}^{n} \frac{(y_t - x_t^T \beta)^2}{g_t(\theta)}. \]

Since by Assumption 1, \(0 < L_g < g_t(\theta) < M_g < \infty, \forall \theta \in \Theta,\)
where \(\Theta\) is closed and bounded, there exist real numbers \(B_1\) and \(B_2\) such that \(E_{\theta_0}[\log L(Y; X, \theta)] < B_1\), \(\forall \theta \in \Theta,\) and

\[ E_{\theta_0}[|\log L(Y; X, \theta)|] < B_2, \quad \forall \theta \in \Theta. \]

If \(E_{\theta_0}[\log L(Y; X, \theta)] = -\infty\) the conclusion is immediate. Therefore we need only consider the case when \(E_{\theta_0}[\log L(Y; X, \theta)] > -\infty\) and hence \(E_{\theta_0}[|\log L(Y; X, \theta)|] < \infty\). Let

\[ u = \log L(Y; X, \theta) - \log L(Y; X, \theta_0), \]

then \(E|u| < \infty\). From Jensen's Inequality it can be shown that for a random variable \(u\) satisfying \(E|u| < \infty\) and such that \(u\) is not equal to a constant with probability 1

\[ E[u] < \log E[e^u]. \]

In this case
\[ B[e^u] = \int \int \int_{-\infty}^{\infty} \frac{L(Y; X, \theta)}{L(Y; X, \theta^0)} \cdot \prod_{t=1}^{n} f_t(y_t; \theta^0) \, dy_1 \ldots dy_n; \]

\[ = \int \int \int_{-\infty}^{\infty} \frac{L(Y; X, \theta)}{L(Y; X, \theta^0)} \cdot L(Y; X, \theta^0) \, dy_1 \ldots dy_n ; \]

\[ \leq 1. \]

By Assumption 6, \( \log L(Y; X, \theta) - \log L(Y; X, \theta^0) \) are nonzero on a set of positive probability. Therefore since \( E[e^u] \leq 1 \) then \( E[u] < 0 \).

The proof is complete. \( \square \)

The following lemma was given by Wald (1949).

**Lemma 3.2.2.** Under Assumptions 1 and 5

\[ \lim_{\rho \rightarrow 0} E[\log L(Y; X, \theta, \rho)] = E[\log L(Y; X, \theta)] \]

where
\[
\log L(Y; X, \theta, \rho) = \sup_{\theta' \in S(\theta, \rho)} \log L(Y; X, \theta')
\]

and \( S(\theta, \rho) \) is the closed sphere \( \| \mathbf{\theta} - \mathbf{\theta}' \| < \rho \), \( \rho \) a positive real number.

\textbf{Proof.} The proof follows the proof of Lemma 2 in Wald (1949).

Let

\[
L^*(Y; X, \theta, \rho) = \begin{cases} 
L(Y; X, \theta, \rho) & \text{when } L(Y; X, \theta, \rho) \geq 1 \\
1 & \text{otherwise .}
\end{cases}
\]

\[
L^*(Y; X, \theta) = \begin{cases} 
L(Y; X, \theta) & \text{when } L(Y; X, \theta) \geq 1 \\
1 & \text{otherwise .}
\end{cases}
\]
Now

\[ L(Y; X, \theta) = (\sqrt{2\pi})^{-n} \prod_{t=1}^{n} \frac{1}{\sqrt{\pi g_t(\theta)}} \exp\left[ -\frac{1}{2} \frac{(y_t - x_t')g_t(\theta)}{g_t(\theta)} \right] \].

By Assumption 1 since 0 < L < g(\theta) < M < \infty and since g_t(\theta) is a continuous function \theta in \Theta, L(Y; X, \theta) is a continuous function of \theta on \Theta. Therefore

\[ \lim_{\rho \to 0} \log L^*(Y; X, \theta, \rho) = \log L^*(Y, X, \theta). \quad (3.2.6) \]

Now \log L^*(Y; X, \theta, \rho) is an increasing function of \rho. In the proof of Lemma 3.2.1 it is shown that \( E [ \log L(Y; X, \theta)] < \infty \) for all \( \theta \in \Theta \).

Therefore by 3.2.6 and the Monotone Convergence Theorem

\[ \lim_{\rho \to 0} E [ \log L^*(Y; X, \theta, \rho)] = \]

\[ \lim_{\rho \to 0} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \log L^*(Y; X, \theta, \rho) L(Y; X, \theta) \, dy_1 \ldots \, dy_n ; \]

\[ = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \lim_{\rho \to 0} \log L^*(Y; x, \theta, \rho) L(Y; X, \theta) \, dy_1 \ldots \, dy_n ; \]
Similarly let

\[
L^{**}(Y; X, \theta, \rho) = \begin{cases} 
L(Y; X, \theta, \rho) & \text{when } L(Y; X, \theta, \rho) \leq 1 \\
1 & \text{otherwise}
\end{cases}
\]

\[
L^{**}(Y; X, \theta) = \begin{cases} 
L(Y; X, \theta) & \text{when } L(Y; X, \theta) \leq 1 \\
1 & \text{otherwise}
\end{cases}
\]

Then since \(-\infty < \log x \leq 0\) for \(0 < x \leq 1\)

\[
| \log L^{**}(Y; X, \theta, \rho) | \leq | \log L^{**}(Y; X, \theta) |
\]

and

\[
\lim_{\rho \to 0} \log L^{**}(Y; X, \theta, \rho) = \log L^{**}(Y; X, \theta)
\]
as in (3.2.6). By the Lebesgue Dominated Convergence Theorem

\[ \lim_{\rho \to 0} E[\log L**(Y; X, \Theta, \rho)] \]

\[ = \lim_{\rho \to 0} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log L**(Y; X, \Theta, \rho) L(Y; X, \Theta) \, dy_1 \cdots dy_n ; \]

\[ = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \lim_{\rho \to 0} \log L**(Y; X, \Theta, \rho) L(Y; X, \Theta) \, dy_1 \cdots dy_n ; \]

\[ = E[\log L**(Y; X, \Theta)] \quad (3.2.8) \]

for the two cases, \( E[\log L**(Y; X, \Theta)] \) finite and \( E[\log L**(Y; X, \Theta)] = -\infty \). The result follows from (3.2.7) and (3.2.8). This completes the proof.

The following lemma due to Silvey (1961) gives sufficient conditions for the weak consistency of the maximum likelihood estimator and is stated without proof.

**Lemma 3.2.3.** The maximum likelihood estimator is consistent if the following conditions are satisfied:

1. \( \Theta \) is compact.

2. \( \left[ \frac{\text{Var}_{\Theta^0}(R_{n, \Theta^0, \Theta})}{\text{E}^R_{\Theta^0}(n, \Theta^0, \Theta)} \right]^{\frac{1}{2}} \to 0 \) as \( n \to \infty \)

uniformly on \( \Theta - \omega \) where \( \omega \) is any open neighborhood of \( \Theta^0 \).
\[ R \neq \log L(Y;X, \theta^0) - \log L(Y;X, \theta), \]

and \( \theta \) is any value of \( \theta \) except \( \theta^0 \).

3. There exists a \( \rho > 0 \) and \( k \) such that \( \{ L(Y;X, \theta^0) \} \) satisfies the following regularity condition on \( \theta - \omega \). For each \( \theta_i \), \( i = 1, 2, \ldots, h \) contained in \( \Theta - \omega \) there exists a real number \( S_{n, \theta_i} \) where

\[ P_{\theta^0} \{ S_{n, \theta_i} < -k \log E_{\theta^0} [R_{n, \theta^0, \theta_i}] \} \to 0 \text{ as } n \to \infty \]

such that for all \( \theta' \), \( \| \theta - \theta' \| < \rho \)

\[ R_{n, \theta_i, \theta'_i} > S_{n, \theta_i}. \]

We now show that Silvey's three conditions are satisfied under Assumptions 1, 5 and 6 and hence that the maximum likelihood estimator is consistent.

**Theorem 3.2.4.** Under Assumptions 1, 5 and 6 Silvey's conditions as given in Lemma 3.2.3 are satisfied and hence the maximum likelihood estimator of \( \theta^0 \) is consistent.
Proof. The proof follows Swamy and Rao (1971). By Assumption 1, \( \mathcal{C} \) is closed and bounded and hence compact. We show that

\[
\left[ \frac{\text{Var}_{\theta_0^n(R, \theta_0^n, \theta)}^{1/2}}{\mathbb{E}_{\theta_0^n(R, \theta_0^n, \theta)}} \right] \to 0 \quad \text{as } n \to \infty
\]

uniformly on \( \mathcal{C} - \omega \) where \( \omega \) and \( n, \theta_0^n, \theta \) are given in Lemma 3.2.3.

We first show that there exist constants \( a_1 \) and \( a_2 \) such that

\[
\lim_{n \to \infty} \frac{1}{n} \mathbb{E}_{\theta_0^n}[\log L(Y; X, \theta)] = a_1.
\]

\[
\lim_{n \to \infty} \frac{1}{n} \text{Var}_{\theta_0^n}[\log L(Y; X, \theta)] = a_2.
\]

By Assumptions 1 and 5

\[
\frac{1}{n} \log L(Y; X, \theta) =
\]

\[
\frac{1}{n} \left[ c - \frac{1}{2} \sum_{t=1}^{n} \log s_t(\theta) - \frac{1}{2} \sum_{t=1}^{n} \frac{(y_t - x_t')\beta^2}{s_t(\theta)} \right]
\]

and there exists a constant \( a_1 \) such that

\[
\lim_{n \to \infty} \frac{1}{n} \mathbb{E}_{\theta_0^n}[\log L(Y; X, \theta)] =
\]

\[
\lim_{n \to \infty} \frac{1}{n} \left[ c - \frac{1}{2} \sum_{t=1}^{n} \log s_t(\theta) - \frac{1}{2} \sum_{t=1}^{n} \frac{\theta_0^n(t)}{s_t(\theta)} \right] = a_1;
\]
since by Assumption 1
\[ 0 < L_t < \xi_t(\theta) < M_t < \infty \quad \forall \theta \in \Theta \]
\[ \forall t = 1, 2, \ldots, n \]
and by Assumption 5
\[ E_{\theta_0}[(y_t - X_t^\prime \beta)^2] = \xi_t(\theta_0) \quad t = 1, 2, \ldots, n . \]

Similarly, there exists a constant \( a_2 \) such that
\[
\lim_{n \to \infty} \frac{1}{n} \text{Var}_{\theta_0} \left[ \log L(Y; X, \Theta) \right] =
\]
\[
\lim_{n \to \infty} \frac{1}{n} \text{Var}_{\theta_0} \left[ - \frac{1}{2} \sum_{t=1}^{n} \frac{(y_t - X_t^\prime \beta)^2}{\xi_t(\theta)} \right] = \frac{1}{4n} \sum_{t=1}^{n} \frac{2\xi_t^2(\theta_0)}{\xi_t^2(\theta)} = a_2
\]

since by Assumption 5,
\[ \text{Var}_{\theta_0}[(y_t - X_t^\prime \beta)^2] = 2\xi_t^2(\theta_0) \quad t = 1, 2, \ldots, n . \]

Therefore it follows that
\[
\sqrt{\frac{\text{Var}_{\theta_0} \left[ \log L(Y; X, \Theta) \right]}{E_{\theta_0} \left[ \log L(Y; X, \Theta) \right]}} \xrightarrow{n \to \infty} 0
\]
and hence

$$\left[ \text{Var } \theta^0 \left[ R_{n, \theta^0, \theta} \right] \right]^{1/2} / \left[ E \theta^0 \left[ R_{n, \theta^0, \theta} \right] \right] \to 0 \text{ as } n \to \infty$$

uniformly on $\mathcal{E} - \omega$. \hspace{1cm} (3.2.9)

We now show that Silvey's third condition is also satisfied. Since $\mathcal{E}$ is closed and bounded then $\mathcal{E} - \omega$ is closed and bounded. Therefore there exists a finite number of open spheres of radius $\rho_\omega$ having centers $\theta_1, \theta_2, \ldots, \theta_h$ that cover $\mathcal{E} - \omega$. By Lemmas 3.2.1 and 3.2.2

$$E \theta^0 \{ \log L(Y; X, \theta_i, \rho_\omega) \} < E \theta^0 \{ \log L(Y; X, \theta^0) \} \hspace{1cm} (3.2.10)$$

$$i = 1, 2, \ldots, h.$$

Let

$$S_{n, \theta_i} = \log L(Y; X, \theta_i) - \log L(Y; X, \theta_i, \rho_\omega).$$
Then

\[ R_n, \theta_i, \theta'_i = \log L(Y; X, \theta_i) - \log L(Y; X, \theta'_i) \geq S_n, \theta_i \]

where \( ||\theta_i - \theta'_i|| < \rho_0 \).

By Chebyshev's Inequality

\[ P_{\theta_0} \left[ S_n, \theta_i > (1-\epsilon)E_{\theta_0} \left[ S_n, \theta_i \right] \right] > 1 - \frac{\text{Var}_{\theta_0} \left[ S_n, \theta_i \right]}{\epsilon^2 E_{\theta_0} \left[ S_n, \theta_i \right]^2} \]

\( \forall \epsilon < 0. \) \hspace{1cm} (3.2.11)

Now

\[ S_n, \theta_i = -R_{n, \theta_0, \theta_i} - [\log L(Y; X, \theta_i, \rho_0) - \log L(Y; X, \theta_0)] \]

and

\[ E_{\theta_0} \left[ S_n, \theta_i \right] = -E_{\theta_0} \left[ R_{n, \theta_0, \theta_i} \right] - E_{\theta_0} \left[ \log L(Y; X, \theta_i, \rho_0) - \log L(Y; X, \theta_0) \right] \]

\[ \geq -E_{\theta_0} \left[ R_{n, \theta_0, \theta_i} \right] \]

by (3.2.10).

By (3.2.9) and (3.2.11) we have
for some

\[ k_w = (1 - c) < 1 \]

Thus Silvey's third condition is satisfied. Therefore the maximum likelihood estimator is consistent. This completes the proof. \( \square \)

We have shown that a maximum of the likelihood function exists and provides a consistent estimator of the true parameter \( \theta^0 \). We now consider the properties of a solution of the likelihood equations.

Cramer (1946, p. 500) demonstrates that under certain conditions the following are true for a scalar parameter \( \theta \).

1. The likelihood equations have a solution \( \tilde{\theta} \) with probability 1 as \( n \to \infty \) which is a consistent estimator of \( \theta^0 \).

2. The asymptotic distribution of \( \sqrt{n}(\tilde{\theta} - \theta^0) \) is normal with mean \( \theta^0 \) and variance

\[
-1 \quad \mathbb{E} \left[ \frac{\partial^2 \log L}{\partial \theta^2} \right]
\]

Huzurbazar (1948) shows that under the same conditions as Cramer the following are also true.

1. The probability that the consistent solution to the likelihood equations is the maximum likelihood estimator converges to 1.
as $n$ approaches $\infty$.

2. There is only one solution of the likelihood equations that is a consistent estimator of $\theta^0$.

The above results are for a scalar parameter $\theta$ and identically distributed densities. Below we extend the above results to multiparameter densities which are nonidentical functions of the same $(p+r) \times 1$ parameter vector $\theta$.

We require that the following regularity conditions be satisfied.

**Regularity Conditions 3.2.5.** The following conditions are satisfied under Assumptions 1, 2, 4, and 5.

1. \[ E_{\theta^0} \left[ \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_i} \right] = 0 ; \]

2. \[ E_{\theta^0} \left[ \frac{\partial^2 \log f_t(y_t; \theta^0)}{\partial \theta_i \partial \theta_j} \right] = -E_{\theta^0} \left[ \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_i} \right] \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_j} \]

\[ i, j = 1, 2, \ldots, p+r ; \]

3. \[ \left\{ E_{\theta^0} \left[ \frac{\partial^2 \log L(Y; X, \theta^0)}{\partial \theta_i \partial \theta_j} \right] \right\}^{p+r} = -H' \psi^{-1} H ; \]

where $H' \psi^{-1} H$ is of full rank $(p+r)$ and where $H$ and $\psi$ are given by equations (3.1.7) - (3.1.9).
Proof. We show that these conditions are met by computing the necessary derivatives and their expectations. By Assumptions 1 and 5 we have

$$\log f_t(y_t; \alpha^0, \beta^0) = c - \frac{1}{2} \log g_t(\alpha^0, \beta^0) - \frac{1}{2} (y_t - x_t^* \beta^0)^2 / g_t(\alpha^0, \beta^0)$$

By Assumption 2 the partials of $g_t(\alpha, \beta)$ exist with respect to the elements of $[\alpha, \beta] \in \mathcal{E}$, $t = 1, 2, \ldots, n$. Therefore,

$$\frac{\partial}{\partial \beta_j} \log f_t(y_t; \alpha^0, \beta^0) = -\frac{1}{2} \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \beta_j} / g_t(\alpha^0, \beta^0) + \frac{x_{tj} (y_t - x_t^* \beta^0)}{g_t(\alpha^0, \beta^0) + \frac{1}{2} (y_t - x_t^* \beta)^2} \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \beta_j} / g_t^2(\alpha, \beta)$$

and

$$\frac{\partial}{\partial \alpha_k} \log f_t(y_t; \alpha^0, \beta^0) = -\frac{1}{2} \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \alpha_k} / g_t(\alpha^0, \beta^0)$$

$$+ \frac{1}{2} (y_t - x_t^* \beta^0)^2 \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \alpha_k} / g_t^2(\alpha, \beta) \quad k = 1, 2, \ldots, r.$$
Since by Assumptions 1 and 5

\[ \mathbb{E}[y_t - X_t' \beta^0] = 0, \quad \mathbb{E}[(y_t - X_t' \beta^0)^2] = \varphi_t(\alpha^0, \beta^0). \quad t = 1, 2, \ldots, n. \]

then

\[
\mathbb{E} \left[ \frac{\partial \log f_t(y_t; \alpha^0, \beta^0)}{\partial \beta_j} \right] = 0 \quad \quad j = 1, 2, \ldots, p. \\
\mathbb{E} \left[ \frac{\partial \log f_t(y_t; \alpha^0, \beta^0)}{\partial \alpha_k} \right] = 0 \quad \quad k = 1, 2, \ldots, r.
\]

By Assumption 2 the second partials of \( \varphi_t(\alpha, \beta) \) with respect to the elements of \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \Theta \) exist \( t = 1, 2, \ldots, n. \)

\[
\frac{\partial^2 \log f_t(y_t; \alpha^0, \beta^0)}{\partial \beta_j \partial \beta_l} = -\frac{1}{2} \left\{ \left( \frac{\partial^2 \varphi_t(\alpha^0, \beta^0)}{\partial \beta_j \partial \beta_l} \right) \varphi_t(\alpha, \beta) - \left( \frac{\partial \varphi_t(\alpha^0, \beta^0)}{\partial \beta_j} \right) \left( \frac{\partial \varphi_t(\alpha^0, \beta^0)}{\partial \beta_l} \right) \right\} \varphi_t^2(\alpha, \beta)
\]
\[
\frac{1}{2} \left\{ (y_t - x_t' \beta^0)^2 (\begin{array}{c}
\frac{\partial \theta_t}{\partial \beta_j} \\
\frac{\partial \theta_t}{\partial \beta_k}
\end{array}) \right\} \bigg/ \theta_t^0 (\alpha, \beta)
\]

\[
\frac{1}{2} \left\{ (y_t - x_t' \beta^0)^2 \left( \frac{\partial \theta_t}{\partial \beta_j} \right) \left( \frac{\partial \theta_t}{\partial \beta_k} \right) \right\} \bigg/ \theta_t^0 (\alpha, \beta)
\]

\[
- \frac{1}{2} \left\{ \left(x_t' x_j \theta_t (\alpha^0, \beta^0) - (y_t - x_t' \beta^0) x_t' \theta_t \left( \frac{\partial \theta_t}{\partial \beta_j} \right) \right\} \bigg/ \theta_t^0 (\alpha, \beta^0)
\]

\[
- x_t' x_j \theta_t (\alpha^0, \beta^0) - \frac{\partial \log f_t(y_t; \alpha^0, \beta^0)}{\partial \beta_j} \left( \frac{\partial \log f_t(y_t; \alpha^0, \beta^0)}{\partial \beta_k} \right)
\]

\[
\frac{\partial^2 \log f_t(y_t; \alpha^0, \beta^0)}{\partial \alpha_k \partial \alpha_l} = - \frac{1}{2} \left\{ \frac{\partial^2 \theta_t (\alpha^0, \beta^0)}{\partial \alpha_k \partial \alpha_l} \bigg/ \theta_t (\alpha^0, \beta^0) \right\}
\]
\[
\begin{align*}
&\left( \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \alpha_k} \right) \left( \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \alpha_l} \right) \right] \bigg/ \varepsilon_t^2(\alpha^0, \beta^0) \\
+ \frac{1}{2} \left\{ (y_t - x_t^0)^2 \left[ \frac{\partial^2 g_t(\alpha^0, \beta^0)}{\partial \alpha_k \partial \alpha_j} \right] \varepsilon_t^2(\alpha^0, \beta^0) \right\} \\
&\left( \frac{\partial u_t(\alpha, \beta)}{\partial \alpha_k} \right) \left( \frac{\partial u_t(\alpha^0, \beta^0)}{\partial \alpha_l} \right) \varepsilon_t(\alpha^0, \beta^0) \bigg] \bigg/ \varepsilon_t^4(\alpha^0, \beta^0) \\
&k = 1, 2, \ldots, r.
\end{align*}
\]

(3.2.13)

\[
E \left[ \frac{\partial^2 \log f_t(y_t; \alpha^0, \beta^0)}{\partial \alpha_k \partial \alpha_l} \right] = - \left( \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \alpha_k} \right) \left( \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \alpha_l} \right) \bigg/ 2\varepsilon_t^2(\alpha^0, \beta^0) \\
k, l = 1, 2, \ldots, r.
\]

\[
= -E \left[ \left( \frac{\partial \log f_t(y_t; \alpha^0, \beta^0)}{\partial \alpha_k} \right) \left( \frac{\partial \log f_t(y_t; \alpha^0, \beta^0)}{\partial \alpha_l} \right) \right].
\]

\[
\frac{\partial^2 \log f_t(y_t; \alpha^0, \beta^0)}{\partial \alpha_k \partial \beta_l} = - \frac{1}{2} \left\{ \left( \frac{\partial^2 g_t(\alpha^0, \beta^0)}{\partial \alpha_k \partial \beta_l} \right) \varepsilon_t(\alpha^0, \beta^0) \right\}.
\]
\[ \begin{align*}
  \cdot \alpha, \ldots, \beta, t & = \alpha \\
  \cdot x, \ldots, \beta, t & = x
\end{align*} \]

\[
\left[ \frac{r_{de}}{(r, \rho, \omega)^{2\gamma}} \right]_2 \left[ \frac{r_{ce}}{(r, \rho, \omega)^{2\gamma}} \right]_3 \equiv \left[ \frac{r_{de}}{(r, \rho, \omega)^{2\gamma}} \right]_5
\]

(4.12.3)

\[
\left( (0 \rho, \rho)^{2\gamma} \right) / \left\{ (0 \rho, \rho)^{2\gamma} \left( \frac{r_{de}}{(r, \rho, \omega)^{2\gamma}} \right) \right\} \equiv \left\{ (0 \rho, \rho)^{2\gamma} \left( \frac{r_{ce}}{(r, \rho, \omega)^{2\gamma}} \right) \right\} \frac{\lambda}{\gamma_1}
\]

\[
\left( (0 \rho, \rho)^{2\gamma} \right) / \left\{ (0 \rho, \rho)^{2\gamma} \left( \frac{r_{de}}{(r, \rho, \omega)^{2\gamma}} \right) \right\} \equiv \left\{ (0 \rho, \rho)^{2\gamma} \left( \frac{r_{ce}}{(r, \rho, \omega)^{2\gamma}} \right) \right\} \frac{\lambda}{\gamma_1} + \left( \frac{r_{ce}}{(r, \rho, \omega)^{2\gamma}} \right) \left( (0 \rho, \rho)^{2\gamma} \right)
\]}
Therefore

\[
\left\{ E_\theta^0 \left[ \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right] \right\}_{i,j}^{p+r} = \left\{ E_\theta^0 \left[ \sum_{t=1}^{n} \frac{\partial^2 \log f_t(y_t; \theta^0)}{\partial \theta_i \partial \theta_j} \right] \right\}_{i,j}^{p+r}
\]

\[
= \left\{ - \sum_{t=1}^{n} \left( \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \theta_i} \right) \left( \frac{\partial g_t(\alpha^0, \beta^0)}{\partial \theta_j} \right) \right\}_{i,j=2}^{p+r} \frac{1}{2 g_t^2(\alpha^0, \beta^0)}
\]

by equations 3.2.12 - 3.2.14

\[
= H' \psi^{-1} H
\]

and where \( H \) and \( \psi \) are defined by equations (3.1.7) - (3.1.9). By Assumption 4 \( H' \psi^{-1} H \) is of full rank \((p+r)\). This completes the proof.

We require the following mean value theorem given in Jenrich (1969) which we state without proof.

**Lemma 3.2.6.** Given that:

1. \( f \) is a real valued function on \( E \times \Omega \) where \( E \) is a measurable space and \( \Omega \) is a convex compact subset of a Euclidean space.
2. For each \( \theta \) in \( \Omega \), let \( f(e; \theta) \) be a measurable function of \( e \); and for each \( e \) in \( E \) let \( f(e; \theta) \) be a continuously differentiable function of \( \theta \).
3. \( \theta_1 \) and \( \theta_2 \) are measurable functions from \( E \) into \( \mathcal{C} \).

Then there exists a measurable function \( \bar{\theta} \) from \( E \) into \( \mathcal{C} \) such that

\[
\Phi(e; \theta_1(e)) - \Phi(e; \theta_2(e)) = \frac{3}{\bar{\theta}(e)} \Phi(e; \bar{\theta}(e))(\theta_1(e) - \theta_2(e))
\]

where \( \bar{\theta}(e) \) lies on the segment joining \( \theta_1(e) \) and \( \theta_2(e) \).

The next two theorems establish consistency properties of a solution of the likelihood equations. We show that: (1) the likelihood equations have a root \( \bar{\theta} \) with probability 1 as \( n \to \infty \) which is consistent for \( \theta^0 \); (2) there is only one root of the likelihood equations that is consistent for \( \theta^0 \); and (3) the probability that the consistent root corresponds to a maximum of the likelihood function converges to 1 as \( n \to \infty \).

**Theorem 3.2.7.** Given that Assumptions 1, 2, 5 and 6 are satisfied the likelihood equations (3.2.2) and (3.2.3) have a root with probability 1 as \( n \to \infty \) which is consistent for the true parameter \( \theta^0 \).

**Proof.** This proof follows C.R. Rao (1965a, p. 300). By Lemma 3.2.1 for \( \theta \neq \theta^0, \theta \in \mathcal{C} \),

\[
\mathbb{E}_{\theta^0} [\log f_t(e_t; \theta)] < \mathbb{E}_{\theta^0} [\log f_t(e_t; \theta^0)]
\]

Consider the points \( \theta \) on the boundary \( S^* \) of \( S(\theta^0, \rho) \) where \( S(\theta^0, \rho) \) is the sphere \( |\theta - \theta^0| < \rho \), \( \rho \) an arbitrary positive number and \( S(\theta^0, \rho) \subset \mathcal{C} \). The existence of such a sphere is guaranteed by Assumption 1.
Then

\[ E_{\theta^0} [\log f_t(e_t; \theta)] < E_{\theta^0} [\log f_t(e_t; \theta^0)] \]

\( \forall \theta \text{ on } S^* \)

and

\[ E_{\theta^0} [\log f_t(e_t; \theta) - \log f_t(e_t; \theta^0)] < 0 \]

\( \forall \theta \text{ on } S^* \)

(3.2.12)

Consider

\[ \frac{1}{n} \sum_{t=1}^{n} \left[ \log f_t(e_t; \theta) - \log f_t(e_t; \theta^0) \right], \quad \theta \text{ on } S^* \]

By Assumptions 1 and 5 since

\[ 0 < L_g < g_t(\theta) < M_g < \infty \]

and

\[ f_t(e_t; \theta) = (2\pi)^{-1/2} g_t^{-1/2}(\theta) \exp\left[ -\frac{1}{2} \frac{e_t^2}{g_t(\theta)} \right] \]

\( \forall \theta \in \Theta, \)

\( \forall t = 1, 2, \ldots, n ; \)
there exists a real number $B_2$ such that

$$E[(\log f_t(e_t; \theta))^2] < B_2 \quad \forall \theta \in \Theta, \; \forall t = 1, 2, \ldots, n.$$  

Therefore by Kolmogorov's sufficient condition for the strong law of large numbers and (3.2.12) there exists a positive real number $\gamma$ such that

$$\frac{1}{n} \sum_{t=1}^{n} [\log f_t(e_t; \theta) - \log f_t(e_t; \theta^0)] \xrightarrow{a.s.} -\gamma$$

for all $\theta$ on $S^\ast$.

Therefore for $n$ sufficiently large

$$P[\frac{1}{n} \sum_{t=1}^{n} (\log \frac{f_t(e_t; \theta)}{f_t(e_t; \theta^0)}) < 0] = 1$$

for all $\theta$ on $S^\ast$. Thus for all $\theta$ on $S^\ast$ and for $n$ sufficiently large the value of the likelihood function will be greater at $\theta^0$ than at any $\theta$ on $S^\ast$ for almost all sample sequences. By Assumptions 1 and 5 $\log L(Y; X, \theta)$ is a continuous function of $\theta$, $\theta \in S(\theta^0, \rho)$ and by Assumption 2 $\log L(Y; X, \theta)$ is differentiable in $\theta$, $\theta \in S(\theta^0, \rho)$.

Therefore there is a relative maximum at some $\tilde{\theta}$ in $S(\theta^0, \rho)$ and hence

$$\frac{\partial \log L(Y; X, \theta)}{\partial \theta_i} \bigg|_{\theta = \tilde{\theta}} = 0, \quad i = 1, 2, \ldots, p+r.$$
Since \( \rho \) is an arbitrary positive real number \( \tilde{\theta} \) is consistent for \( \theta^0 \).

This completes the proof. \( \Box \)

**Theorem 3.2.8.** Given that Assumptions 1, 2, 4, 5 and 6 are satisfied the likelihood equations have only one consistent root, \( \tilde{\theta} \), and the probability that the likelihood function is a maximum at \( \tilde{\theta} \), converges to 1 as \( n \to \infty \).

**Proof.** The proof follows Huzurbazar (1948). We first show that

\[
\left\{ \frac{1}{n} \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right) \right\}_{\theta = \theta^0}^{p+r}
\]

converges in probability to the negative definite matrix

\[
- \frac{H'}{n} \psi^{-1} H - \left\{ \mathbb{E}_\theta \left[ \frac{1}{n} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \theta)}{\partial \theta_i} \left( \frac{\partial \log f_t(y_t; \theta)}{\partial \theta_j} \right) \right] \right\}^{p+r}
\]

Now the elements of

\[
\left\{ \frac{\partial^2 \log f_t(y_t; \theta)}{\partial \theta_i \partial \theta_j} \right\}_{i,j=1}^{p+r}
\]

are functions of \( \xi_t(\theta) \), the first and second partials of \( g_t(\theta) \) and
\((y_t - X_t' \beta)\). By Assumption 5 the moments of \((y_t - X_t' \beta)\) are functions of \(g_t(\theta)\). Since by Assumptions 1 and 2 \(g_t(\theta)\) and its first and second partials are uniformly bounded \(\forall \theta \in \Theta\), the moments of

\[ \begin{bmatrix} \frac{\partial^2 \log f_t(y_t; \theta)}{\partial \theta_i \partial \theta_j} \end{bmatrix}^{p+r} \]

for \(t = 1, 2, \ldots, n; i, j = 1\)

are uniformly bounded \(\forall \theta \in \Theta\). Therefore by Chebyshev's Weak Law of Large Numbers and Regularity Conditions 3.2.5

\[ \frac{1}{n} \sum_{t=1}^{n} \left( \frac{\partial^2 \log f_t(y_t; \theta)}{\partial \theta_i \partial \theta_j} \right) \]

converges in probability to

\[ \mathbb{E}_\theta \left[ \frac{1}{n} \sum_{t=1}^{n} \left( \frac{\partial^2 \log f_t(y_t; \theta)}{\partial \theta_i \partial \theta_j} \right) \right] \]

\[ = -\frac{1}{n} \sum_{t=1}^{n} \left( \frac{\partial g_t(\theta_0)}{\partial \theta_i} \right) \left( \frac{\partial g_t(\theta_0)}{\partial \theta_j} \right) / 2 g_t^2(\theta_0). \]

\((3.2.13)\)

We show that if \(\theta^*\) is any consistent estimator of \(\theta_0, \theta^* \in \Theta\), the matrix
\[
\left\{ \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right) \right\}_{\theta = \theta^*}^{p+r} \quad i, j = 1
\]
converges in probability to

\[
\left\{ \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right) \right\}_{\theta = \theta^0}^{p+r} \quad i, j = 1
\]

By the mean value theorem (Lemma 3.2.6) and by the properties of the partials of \( \{g_t(\theta)\}_{t=1}^n \) given by Assumption 2

\[
\frac{1}{n} \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right)_{\theta = \theta^*} - \frac{1}{n} \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right)_{\theta = \theta^0} = \frac{1}{n} \sum_{k=1}^{p+r} \left( \frac{\partial^3 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right)_{\theta = \theta^*} \left( \theta_k^* - \theta_k^0 \right) \tag{3.2.14}
\]

where \( \theta' \) lies between \( \theta^0 \) and \( \theta^* \).
Now since \( \theta^* \) is consistent \( (\theta^0 - \theta^*) \xrightarrow{p} 0 \) and by Assumptions 1 and 2

\[
\frac{1}{n} \sum_{k=1}^{p+r} \left( \frac{\partial^3 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right)
\]

is uniformly bounded in \( \Theta \). Therefore from (3.2.14)

\[
\left\{ \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right) \right\}_{\theta = \theta^*}^{p+r}
\]

converges in probability to

\[
\left\{ \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right) \right\}_{\theta = \theta^0}^{p+r} = -H' \psi^{-1} H.
\]

Since the root \( \tilde{\theta} \) is a consistent estimator of \( \theta^0 \) then

\[
\left\{ \left( \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right) \right\}_{\theta = \tilde{\theta}}^{p+r}
\]
converges in probability to the negative definite matrix $-H'\psi^{-1}H$.

The root $\tilde{\theta}$ corresponds to a relative maximum if the matrix

$$\left\{ \begin{array}{c} \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \\ \theta = \tilde{\theta} \end{array} \right\}_{i,j=1}^{p+r}$$

is negative definite. Therefore the probability that the likelihood function is a relative maximum at $\tilde{\theta}$ converges to 1 as $n \to \infty$.

Let $\tilde{\theta}_1$ and $\tilde{\theta}_2$ be two distinct consistent roots of the likelihood equations.

Then we know that

$$\left( \begin{array}{c} \frac{\partial \log L(Y; X, \theta)}{\partial \theta_i} \\ \theta = \tilde{\theta}_1 \end{array} \right) = \left( \begin{array}{c} \frac{\partial \log L(Y; X, \theta)}{\partial \theta_i} \\ \theta = \tilde{\theta}_2 \end{array} \right) = 0$$

$$i = 1, 2, \ldots, p+r.$$ 

By Assumptions 1, 2 and 5 $0 < L_g < g_t(\theta) < M_g < \infty$ and the first and second partials of $g_t(\theta)$ with respect to the elements of $\theta$ are continuous $\forall \theta \in \Theta$, $t = 1, 2, \ldots, n$. Therefore the elements of

$$\left\{ \begin{array}{c} \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \\ \theta = \tilde{\theta}_1 \end{array} \right\}_{i,j=1}^{p+r}$$
are continuous functions of $\theta$ on $C$ for all $t = 1, 2, \ldots, n$. Since $\tilde{\theta}_1$ and $\tilde{\theta}_2$ converge in probability to $\theta^0$ then

$$\left\{ \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right\}_{\theta = \tilde{\theta}_1}^{p+r} i, j = 1$$

and

$$\left\{ \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right\}_{\theta = \tilde{\theta}_2}^{p+r} i, j = 1$$

converge in probability to

$$\left\{ \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right\}_{\theta = \theta^0}^{p+r} i, j = 1$$

Let $\tilde{\theta}_3$ be any point on the line joining $\tilde{\theta}_1$ and $\tilde{\theta}_2$ then

$$\tilde{\theta}_3 = \tilde{\theta}_1 + \lambda (\tilde{\theta}_1 - \tilde{\theta}_2) \text{ where } 0 < \lambda < 1.$$ The directional derivative is given by

$$\frac{\partial \log L(Y; X, \theta)}{\partial \lambda}.$$
We have
\[ \frac{\partial \log L(Y; X, \theta)}{\partial \lambda} = \sum_{i=1}^{p+r} \frac{\partial \log L(Y; X, \theta)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \lambda}. \]

Therefore \( \frac{\partial \log L(Y; X, \theta)}{\partial \lambda} = 0 \) at \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \). By Rolle's Theorem since \( \frac{\partial \log L(Y; X, \theta)}{\partial \lambda} = 0 \) at \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \) there exists a point \( \tilde{\theta}_3 \) on the line joining \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \) such that \( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \lambda^2} = 0 \).

Now
\[ \frac{\partial^2 \log L}{\partial \lambda^2} = \sum_{i=1}^p \sum_{j=1}^p \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right) \left( \frac{\partial \theta_i}{\partial \lambda} \frac{\partial \theta_j}{\partial \lambda} \right) = 0 \]

which implies that the matrix
\[ \left\{ \left( \frac{\partial^2 \log L(Y; X, \theta)}{\partial \theta_i \partial \theta_j} \right) \right\}_{i,j=1}^{p+r} \]  

(3.2.15)

is not negative definite. Since \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \) converge in probability to \( \theta^0 \) and since \( \tilde{\theta}_3 \) lies on a line joining \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \) then \( \tilde{\theta}_3 \) converges in probability to \( \theta^0 \). But if \( \tilde{\theta}_3 \) converges in probability to \( \theta^0 \) then
matrix (3.2.15) must be negative definite, a contradiction. The conclusion follows.

The following three lemmas are used to obtain a central limit theorem for \( \frac{1}{\sqrt{n}} A' \epsilon \) where \( \epsilon \) is a vector of independent random variables and \( A \) is a matrix of real constants. These results will be used on several occasions throughout Section III.

Lemma 3.2.9 is given as Theorem 1 in Eicker (1963, p. 440) and is stated below without proof.

**Lemma 3.2.9.** Given that the following conditions are satisfied:

1. \( \max_{k=1,2,\ldots,k_n} \left[ \frac{\sum_{k=1}^{n} a_{nk}^2}{\sum_{k=1}^{n} a_{nk}^2} \right] \rightarrow 0 \) as \( n \rightarrow \infty \).

2. \( \sup_{G_k} \int_{|x| > c} x^2 dG_k(x) \rightarrow 0 \) as \( c \rightarrow \infty \).

3. \( \inf_{G_k} \int x^2 dG_k(x) > 0 \).

then

\[ \frac{b^{-1}}{\sqrt{n}} \sum_{k=1}^{n} a_{nk} \epsilon_k \xrightarrow{\mathcal{D}} N(0,1) \]
where

1. \( \{a_{nk}\} \) is a double sequence of real constants
   \[ n = 1, 2, \ldots; k = 1, 2, \ldots k_n; \]

2. \( \{k_n\} \) is a sequence of positive integers such that \( a_{nk_n} \neq 0 \) and
   \[ a_{nk} = 0 \quad \text{for} \quad k > k_n \quad n = 1, 2, \ldots; \]

3. \( \{\varepsilon_k\} \) is a sequence of independent random variables with mean 0,
   positive finite variances, \( \sigma_k^2 \), and corresponding distribution functions.
   \( \{G_k\}, k = 1, 2, \ldots; \)

4. \( B_n^2 = \text{Var} \left[ \sum_{k=1}^{k_n} a_{nk} \varepsilon_k \right] = \sum_{k=1}^{k_n} a_{nk}^2 \sigma_k^2. \)

Lemma 3.2.10 is given in Meyer (1966, p. 16) and is stated below without proof.

**Lemma 3.2.10.** Given that \( \{\varepsilon_k\} \) is a sequence of random variables
with finite second moments and corresponding distribution functions \( \{F_k\} \)
\( k = 1, 2, \ldots \) and that there exists a positive real valued increasing
function \( h(t) \) such that:

\[
\lim_{t \to \infty} \frac{h(t)}{t} = +\infty \quad \text{and} \quad \sup_{F_k} \mathbb{E} \left[ h(|\varepsilon_k|) \right] < \infty; \]

then

\[
\sup_{F_k} \int_{|x| > c} |x| \, dF_k \quad \longrightarrow \quad 0 \quad \text{as} \quad c \quad \longrightarrow \quad \infty. \]
Meyer recommends using \( h(t) = t^p \), \( p > 1 \), in applications of this theorem.

Lemma 3.2.11 is given in C.R. Rao (1965a, p. 103) and is stated below without proof.

**Lemma 3.2.11.** Given a sequence of \( p \times 1 \) vector random variables \( \{e_n^t\} \), \( n = 1, 2, \ldots \), an arbitrary \( p \times 1 \) vector \( \lambda \) and a \( p \times 1 \) vector \( e \) such that
\[
\lambda' e_n^t \xrightarrow{\alpha} \lambda' e
\]
then \( G_n \rightarrow G \) where \( G_n \) and \( G \) are the distribution functions of \( e_n \) and \( e \) respectively \( n = 1, 2, \ldots \).

Lemmas 3.2.9, 3.2.10 and 3.2.11 will be used to prove Lemma 3.2.12 below.

**Lemma 3.2.12.** Given that:

1. \( \{e_t^t\}_{t=1}^n \) is a sequence of independent random variables with means \( 0 \), finite variances \( \{\sigma_t^2\} \) and distribution functions \( \{G_t\}_{t=1}^n \).

2. There exists real positive numbers \( L \) and \( M \) such that \( 0 < L < \sigma_t^2 \) and \( E[e_t^t] < M ; \ t = 1, 2, \ldots, n \).

3. \( A \) is an \( n \times p \) matrix of real constants uniformly bounded for all \( n \) such that
\[
\lim_{n \to \infty} \left[ \frac{A'A}{n} \right] \text{ exists and is positive definite}.
\]
4. \( \lim_{n \to \infty} \frac{1}{n} \left[ \frac{A'DA}{n} \right] = Q \) where \( E[\epsilon \epsilon'] = D \), \( \epsilon' = (\epsilon_1, \ldots, \epsilon_n) \), and \( Q \) is a positive definite matrix of real constants, then

\[
\frac{1}{\sqrt{n}} A' \epsilon \xrightarrow{\alpha} N(0, Q) .
\]

**Proof.** Let \( \lambda \) be an arbitrary \( p \times 1 \) vector of real constants. We first show that

\[
\frac{1}{\sqrt{n}} \lambda' A' \epsilon \xrightarrow{\alpha} N(0, \lambda' Q \lambda)
\]

using Lemma 3.2.9.

\[
\frac{1}{\sqrt{n}} \lambda' A' \epsilon = \frac{1}{\sqrt{n}} \sum_{j=1}^{p} \lambda_j \left( \sum_{t=1}^{n} a_{tj} \epsilon_t \right) ;
\]

\[
= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \epsilon_t \left[ \sum_{j=1}^{p} \lambda_j a_{tj} \right] ;
\]

\[
= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} b_t \epsilon_t ;
\]

where

\[
b_t = \left( \frac{1}{\sqrt{n}} \sum_{j=1}^{p} \lambda_j a_{tj} \right) .
\]
We want to show that

\[
\max_{t=1, 2, \ldots, n} \left[ \frac{b_t^2}{\sqrt{\frac{n}{\sum b_t^2}}} \right] \to 0 \quad \text{as} \quad n \to \infty.
\]

Let

\[
\lambda_0 = \max_{j=1, 2, \ldots, p} \{ |\lambda_j| \},
\]

and let

\[
a_{t0} = \max_{j=1, 2, \ldots, p} \{|a_{tj}|\},
\]

then

\[
b_t \leq \frac{p}{\sqrt{n}} \lambda_0 a_{t0}.
\]

Let

\[
a_{00} = \max_{t=1, 2, \ldots, n} \{ a_{t0} \},
\]

then

\[
\max_{t=1, 2, \ldots, n} \left[ \frac{b_t^2}{\sqrt{\frac{n}{\sum b_t^2}}} \right] \leq \max_{t=1, 2, \ldots, n} \left[ \frac{p^2 \lambda_0^2 a_{00}^2 \frac{1}{n}}{\sqrt{\frac{n}{\sum b_t^2}}} \right] = O \left( \frac{1}{\sqrt{n}} \right).
\]
since $B$ is positive definite \( \exists \rho > 0 \) such that

\[
\sum_{t=1}^{n} \left( \frac{1}{n} \sum_{t=1}^{n} \left( \sum_{j=1}^{p} \lambda_{j} a_{tj} \right)^{2} \right) \longrightarrow \rho .
\]

To show that

\[
\sup_{G_t} \int_{|x| > c} x^2 \, dG_t(x) \longrightarrow 0 \quad \text{as} \quad c \longrightarrow \infty
\]

we use Lemma 3.2.10 above. We require a positive real valued increasing function $h(t)$ such that

\[
\lim_{t \to \infty} \frac{h(t)}{t} = \infty \quad \text{and} \quad \sup_{G_t} E[h(|e_t^2|)] < \infty .
\]

We let $h(t) = t^2$ then since by assumption $E[|e_t^2|] < M$, $t = 1, 2, \ldots, n$; we have the required result.

We also require that

\[
\inf_{G_t} \int_{G_t} x^2 \, dG_t(x) > 0 \quad \text{for} \quad t = 1, 2, \ldots, n ;
\]
which follows from the fact that $E[e_t] = 0$ and

$$0 < L < \sigma_t^2, \quad t = 1, 2, \ldots, n.$$ 

Thus

$$\frac{1}{\sqrt{n}} \lambda' A' \varepsilon \xrightarrow{\alpha} N(0, \lambda' Q \lambda).$$

By Lemma 3.2.11 since $\lambda$ is arbitrary

$$\frac{1}{\sqrt{n}} A' \varepsilon \xrightarrow{\alpha} N(0, Q).$$

This completes the proof. []

Lemma 3.2.13. Given that Assumptions 1, 2, 4 and 5 are satisfied the asymptotic distribution of the vector

$$c = \left\{ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_j} \right\}_{j=1}^{p+r}$$

is normal with mean 0 and covariance matrix
\[ \frac{1}{n} H' \psi^{-1} H = \left\{ E_0 \left[ \frac{1}{n} \sum_{t=1}^{n} \left( \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_i} \right) \left( \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_j} \right) \right] \right\}_{i,j=1}^{p+r} \]

and where \( H \) and \( \psi \) are given by (3.1.7) - (3.1.9).

**Proof.** This proof follows Wilks (1962, p. 358). Consider the element

\[ c_j = \frac{1}{\sqrt{n}} n \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_j} \]

of the vector \( c \).

By the Regularity Conditions 3.2.5

\[ E_0 \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_j} \right] = 0 \quad j = 1, 2, \ldots, p+r ; \]

\[ \left\{ E_0 \left[ \frac{1}{n} \sum_{t=1}^{n} \left( \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_i} \right) \left( \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_j} \right) \right] \right\}_{i,j=1}^{p+r} = H' \psi^{-1} H . \]

Let \( \lambda \) be an arbitrary \((p+r) \times 1\) vector of real constants and consider \( \lambda'C \).
where

\[ e_t = \sum_{i=1}^{p+r} \lambda_i \left( \frac{\partial \log f_t(y_t; \theta_0)}{\partial \theta_i} \right) \quad t = 1, 2, \ldots, n. \]

The \( \{ e_t \}_{t=1}^n \) are independent random variables with mean 0 and variances

\[
\left\{ \sum_{i=1}^{p+r} \sum_{j=1}^{p+r} \lambda_i \lambda_j E \left[ \left( \frac{\partial \log f_t(\theta_0)}{\partial \theta_i} \right) \left( \frac{\partial \log f_t(\theta_0)}{\partial \theta_j} \right) \right] \right\}_{t=1}^n
\]

By Assumptions 1, 2 and 5 these variances are uniformly bounded \( t = 1, 2, \ldots, n \) and the fourth moments of the \( \{ e_t \}_{t=1}^n \) are uniformly bounded.

By Assumption 4 \( \lim_{n \to \infty} \left( \frac{H^{-1}}{n} \right) \) exists and is positive definite.
Thus

$$\inf_{\varepsilon_t} E[\varepsilon_t^2] > 0.$$ 

Therefore by Lemma 3.2.12

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t \xrightarrow{\alpha} N(0, \lim_{n \to \infty} \frac{1}{n} (\lambda^\prime H^{-1} H \lambda))$$

where

$$\lambda^\prime C = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t$$

and

$$\frac{1}{n} \lambda^\prime H^{-1} H \lambda = \frac{1}{n} \sum_{t=1}^{n} \sum_{i=1}^{P+r} \sum_{j=1}^{P+r} \lambda_i \lambda_j E[\frac{\partial \log f_t(y_t; \Theta^0)}{\partial \theta_i} \frac{\partial \log f_t(y_t; \Theta^0)}{\partial \theta_j}].$$

Therefore by Lemma 3.2.11

$$C \xrightarrow{\alpha} N(0, \lim_{n \to \infty} \frac{1}{n} H^\prime \psi^{-1} H)$$

This completes the proof.

We now determine the asymptotic distribution of $\sqrt{n} (\tilde{\Theta} - \Theta^0)$.

**Theorem 3.2.14.** Given that Assumptions 1, 2, 4, 5 and 6 are satisfied then the asymptotic distribution of $\sqrt{n} (\tilde{\Theta} - \Theta^0)$ is normal with mean
0 and covariance matrix

\[
\lim_{n \to \infty} \left( \frac{H'\psi^{-1}H}{n} \right)^{-1}
\]

where \( \tilde{\theta} \) is the consistent root of the likelihood equations given in Theorem 3.2.7, and \( H \) and \( \psi \) are given in (3.1.7) - (3.1.9).

This proof follows Wilks (1962, p. 361).

Proof. By Assumptions 1 and 2 we have that

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \theta)}{\partial \theta_i} , \quad i = 1, 2, \ldots, p+r
\]

is continuous and has continuous partial derivatives, \( \forall \theta \in \Theta, \ t = 1, 2, \ldots, n \). By Assumptions 1 and 2 \( \tilde{\theta} \) is a measurable function of \( \{ y_t \}_{t=1}^{n} \) into \( \Theta \).

Then by Lemma 3.2.6

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_i} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \tilde{\theta})}{\partial \theta_i} \\
+ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sum_{j=1}^{p+r} \frac{\partial^2 \log f_t(y_t; \tilde{\theta})}{\partial \theta_i \partial \theta_j} (\theta_{0j} - \tilde{\theta}_j) \quad i = 1, 2, \ldots, p+r
\]
Now since $\tilde{\theta}$ is a solution to the likelihood equations

$$
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \tilde{\theta})}{\partial \theta_i} = 0 \quad i = 1, 2, \ldots, p+r .
$$

Thus we have

$$
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_i} = \frac{1}{n} \sum_{t=1}^{n} \sum_{j=1}^{p+r} \frac{\partial^2 \log f_t(y_t; \tilde{\theta})}{\partial \theta_i \partial \theta_j} \sqrt{n} (\theta_j - \tilde{\theta}_j)
$$

(3.2.16)

As in Theorem 3.2.13 let $C$ be the $(p+r) \times 1$ vector of elements

$$
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log f_t(y_t; \theta^0)}{\partial \theta_i} \quad i = 1, 2, \ldots, (p+r) ;
$$

then by Theorem 3.2.13

$$
C \overset{\mathcal{D}}{\rightarrow} N(0, \lim_{n \to \infty} \frac{1}{n}(H' \psi^{-1} H)) .
$$
Since \( \bar{\Theta} \to \Theta^0 \) then \( \hat{\Theta} \to \Theta^0 \). As in the proof of Theorem 3.2.8 by the continuity of

\[
\frac{\partial^2 \log f_t(y_t; \Theta)}{\partial \Theta_i \partial \Theta_j}, \quad \Theta \in C \\
i, j = 1, 2, ..., p+r
\]

by Assumptions 1 and 2

\[
\frac{1}{n} \sum_{t=1}^{p+r} \frac{\partial^2 \log f_t(y_t; \Theta)}{\partial \Theta_i \partial \Theta_j} \to \mathbb{E} \left[ \sum_{j=1}^{p+r} \frac{\partial^2 \log f_t(y_t; \Theta^0)}{\partial \Theta_i \partial \Theta_j} \right], \\
i = 1, 2, ..., p+r.
\]

(3.2.17)

Therefore by the Regularity Conditions 3.2.5

\[
\left\{ \frac{1}{n} \sum_{t=1}^{p+r} \frac{\partial^2 \log f_t(y_t; \Theta)}{\partial \Theta_i \partial \Theta_j} \right\}^{p+r} \to - \frac{H' \Psi^{-1} H}{n}.
\]

(3.2.18)
Since
\[ c \xrightarrow{\alpha} N(0, \lim_{n \to \infty} \frac{H'\psi^{-1}H}{n}) \]
then from (3.2.16) and (3.2.18)
\[ \sqrt{n}(\hat{\theta} - \theta^0) \xrightarrow{\alpha} \left( \frac{H'\psi^{-1}H}{n} \right)^{-1} c \]
\[ \xrightarrow{\alpha} N(0, \lim_{n \to \infty} \left( \frac{H'\psi^{-1}H}{n} \right)^{-1}) \]

This completes the proof. □

Thus by Theorems 3.2.7, 3.2.8 and 3.2.14 we have shown that under Assumptions 1, 2, 4, 5 and 6:

1. The likelihood equations have a root \( \hat{\theta} \) with probability 1 as \( n \to \infty \) which is consistent for the true parameter \( \theta^0 \).

2. There is only one root of the likelihood equations that is a consistent estimator of \( \theta^0 \) and the probability that the likelihood function is a maximum at this root converges to 1 as \( n \to \infty \).

3. The asymptotic distribution of \( \sqrt{n}(\hat{\theta} - \theta^0) \) is normal with mean 0 and covariance \( \lim_{n \to \infty} \left( \frac{H'\psi^{-1}H}{n} \right) \) where \( H \) and \( \psi \) are given by (3.1.7) - (3.1.9).
It was shown that under Assumptions 1, 5 and 6 the maximum likelihood estimator exists and is a consistent estimator of \( \theta^0 \). If the maximum of the likelihood function occurs in the interior of \( \Theta \), then the maximum will correspond to a root of the likelihood equations. Under Assumptions 1, 2, 4, 5 and 6 we have shown that there is only one consistent root of the likelihood equations and hence we may conclude that this consistent root is the maximum likelihood estimator. If the maximum of the likelihood function is not an interior point of \( \Theta \) then a root of the likelihood equations may not correspond to the maximum likelihood estimator.

We have shown however that under Assumptions 1, 2, 4, 5 and 6 the probability is 1 that as \( n \to \infty \) there exists a root of the likelihood equations which is consistent for \( \theta^0 \) and that the probability that the likelihood function is a maximum at this root converges to 1 as \( n \to \infty \).

The solution of the likelihood equations (3.2.4) is a nontrivial problem. There is no analytic expression for the solutions of the equations and hence iterative procedures must be used. In Subsection E we develop an iterative procedure to find a local maximum of the likelihood function.
C. Existence and Properties of Simple Least Squares Estimators

It is well known that the simple least squares estimator of $\beta$

$$\hat{\beta} = (X'X)^{-1} X'Y$$  \hspace{1cm} (3.3.1)

eexists and is unbiased under Assumption 1. Some properties of $\hat{\beta}$ are given in Theorem 3.3.6 below. To prove Theorem 3.3.6 we shall require the following lemmas and notation. Lemma 3.3.1 is concerned with almost sure convergence.

**Lemma 3.3.1.** Given that:

1. $\{e_t\}_{t=1}^n$ is a sequence of independent random variables with means 0 and finite variances $\{\sigma^2\}_{t=1}^n$;

2. $\sum_{t=1}^n \sigma^2_t / t^2 < \infty$ for all $n$;

3. $\{a_t\}_{t=1}^n$ a sequence of uniformly bounded real numbers; then

$$\frac{1}{n} \sum_{t=1}^n a_t e_t \xrightarrow{a.s.} 0.$$

**Proof.** Let $Z_t = a_t e_t$ and apply Kolmogorov's sufficient condition for the strong law of large numbers. □

Since the orders of convergence notation of Mann and Wald (1943) is used throughout we give the following definitions and properties which
are given in Fuller (1969).

**Definition 3.3.2.** Given that \( \{a_n\} \) is a sequence of real numbers and \( \{g_n\} \) a sequence of positive integers and that there exists a real number \( M, 0 < M < \infty \), such that \( \frac{|a_n|}{g_n} \leq M \) for all \( n \), then the sequence \( \{a_n\} \) is convergent to order \( \{g_n\} \). This will be denoted by \( a_n = O(g_n) \).

**Definition 3.3.3.** Given that \( \{x_n\} \) is a sequence of random variables, \( \{g_n\} \) a sequence of positive real numbers, and that for all \( \epsilon > 0 \) there exists a real number \( M_\epsilon \) and an integer \( N_\epsilon \) such that

\[
P[ |x_n| > M_\epsilon g_n ] < \epsilon \text{ for all } n > N_\epsilon,
\]

then the sequence \( \{x_n\} \) is of probability order \( g_n \). This will be denoted by \( x_n = O_p(g_n) \).

The following properties of orders of convergence are useful.

**Properties 3.3.4.** Given that:

1. \( \{g_n\}, \{f_n\}, \{r_n\} \) are sequences of positive real numbers;
2. \( \{a_n\}, \{b_n\} \) are sequences of real numbers;
3. \( \{x_n\}, \{y_n\} \) are sequences of random variables;
4. \( \{H_n(z)\} \) is a sequence of measurable functions of the random variables \( \{x_n\} \) and the real numbers \( \{a_n\} \);

then the following are true:

1. If \( a_n = O(g_n) \) and \( b_n = O(f_n) \) then

\[
\frac{a_n b_n}{g_n} = O(f_n g_n) \text{ and } a_n + b_n = O(\max [g_n, f_n]);
\]
2. If \( x_n = o_p(g_n) \) and \( y_n = o_p(f_n) \) then

\[
x_n y_n = o_p(f_n g_n) \quad \text{and} \quad a_n b_n = o_p(\max[g_n, f_n]);
\]

3. If \( a_n = o(g_n) \) implies that \( H_n(a_n) = o(r_n) \) then

\[
x_n = o_p(g_n) \quad \text{implies that} \quad H_n(x_n) = o_p(r_n);
\]

To show that \( \hat{\beta} - \beta = O_p(\frac{1}{\sqrt{n}}) \) we shall require the following lemma.

**Lemma 3.3.5.** Given that:

1. \( \{\epsilon_t\} \) is a sequence of random variables with \( E[\epsilon_t] = \mu_t \),

\[
V(\epsilon_t) = \sigma_t^2, \quad Cov(\epsilon_t, \epsilon_s) = \sigma_{st}, \quad t \neq s, \quad t, s = 1, 2, \ldots, n;
\]

2. \( \mu_t = o(n^{-1}); \quad \sigma_t^2 = o(n^{-s}); \quad \sigma_{st} = o(n^{-t}), \quad t \geq s + 1, \)

\( t, s = 1, 2, \ldots, n; \)

3. \( \{a_t\} \) a sequence of uniformly bounded real constants; then

\[
\frac{1}{n} \sum_{t=1}^{n} a_t \epsilon_t = o_p(n^{-k}) \quad \text{where} \quad k = \min\{r, \frac{s+1}{2}\}.
\]

**Proof.** The proof follows Fuller (1969)

Let

\[
X_n = \frac{1}{n} \sum_{t=1}^{n} a_t \epsilon_t.
\]
\[ E[X_n] = \frac{1}{n} \sum_{t=1}^{\infty} a_t \mu_t = \frac{1}{n} \cdot n \cdot O(n^{-x}) = O(n^{-x}). \]

\[ V[X_n] = \frac{1}{n^2} \sum_{t=1}^{\infty} a_t^2 \sigma_t^2 + \frac{1}{n^2} \sum_{t=1}^{\infty} \sum_{s=1}^{\infty} a_t a_s \sigma_t \sigma_s, \]

\[ = \frac{1}{n^2} \cdot n \cdot O(n^{-s}) + \frac{1}{n^2} (n)(n-1) O(n^{-t}). \]

\[ = O(n^{-s-1}). \]

By Chebyshev's Inequality

\[ P \left[ \left| X_n - \frac{1}{n} \sum_{t=1}^{\infty} a_t \mu_t \right| \geq \delta \right] \leq \frac{V(X_n)}{\delta^2} \quad \forall \delta > 0. \]

Let \( \delta = \varepsilon [V(X_n)]^{1/2} \), where \( \varepsilon > 0 \) is arbitrary, then

\[ P \left[ \left| X_n - \frac{1}{n} \sum_{t=1}^{\infty} a_t \mu_t \right| \geq \varepsilon [V(X_n)]^{1/2} \right] \leq \frac{1}{\varepsilon^2}. \]

Since \( [V(X_n)]^{1/2} = O(n^{-\left(\frac{s+1}{2}\right)}) \) then

\[ X_n - \frac{1}{n} \sum_{t=1}^{\infty} a_t \mu_t = O_p \left( n^{-\left(\frac{s+1}{2}\right)} \right). \]

\[ \therefore X_n = O_p \left( n^{-\left(\frac{s+1}{2}\right)} \right) + O(n^{-x}). \]
\[ = 0_p(n^{-k}) \text{ where } k = \min\{r, \frac{s+1}{2}\} \].

This completes the proof. \[\Box\]

We now obtain some properties of \(\hat{\beta}\).

**Theorem 3.3.6.** Given that Assumption 1 is satisfied then the following are true:

1. \(\hat{\beta} - \beta^0 \overset{a.s.}{\longrightarrow} 0\);

2. \((\hat{\beta} - \beta^0) = 0_p\left(\frac{1}{\sqrt{n}}\right)\);

3. \(\sqrt{n}(\hat{\beta} - \beta^0) \overset{\alpha}{\longrightarrow} N[0, \lim_{n\to\infty} \left(\frac{X'X}{n}\right)^{-1} \frac{X'G_X}{n} \left(\frac{X'X}{n}\right)^{-1}]\).

**Proof.** From (3.3.1) and Assumption 1 it follows that

\[(\beta - \beta^0) = (X'X)^{-1} X'e\;\]

\[= \left(\frac{X'X}{n}\right)^{-1} \frac{X'e}{n} .\]

By Assumption 1, \(\left(\frac{X'X}{n}\right)^{-1} = O(1)\). Consider \(X'e\).
By Lemmas 3.3.1, 3.3.5 and 3.2.12 and Assumption 1 it follows that:

1. \( \frac{X'e}{n} \xrightarrow[n \to \infty]{a.s.} 0 \);

2. \( \frac{X'e}{n} = 0_p \left( \frac{1}{\sqrt{n}} \right) \);

3. \( \frac{(X'X)^{-1} X'e}{\sqrt{n}} \xrightarrow{\alpha} N[0, \lim_{n \to \infty} \frac{X'X}{n}(X'G X)(X'X)^{-1}] \).

This completes the proof. □

Estimation of the unknown covariance matrix \( G \) involves estimation of the unknown parameters \( \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} = \theta^0 \) in the functions \( \{ g_t(\alpha^0, \beta^0) \}_{t=1}^n \).

To obtain a simple least squares estimator of \( \theta^0 \) we use the residuals \( Y - \hat{X}\beta \), from the simple least squares estimator \( \hat{X}\beta \). We have

\[
\hat{e}_t = y_t - \hat{x}_t \beta; \quad t = 1, 2, \ldots, n.
\]

\[
= y_t - x_t (X'X)^{-1} X'y;
\]

\[
= e_t - x_t (X'X)^{-1} X'e;
\]

\[
= M'_t e;
\]

where \( M'_t = 1 - x_t [X'X]^{-1} X' \) is a row of the matrix \( M = I - X(X'X)^{-1} X' \).
In vector notation we have

\[ \hat{e} = Me. \]

From the properties of the \( \{ e_t \} \) given in Assumption 1 we have

\[ E[ee'] = ME[ee']M; \]

\[ = MG M; \]

since \( E[ee'] = G \) where \( G \) is the \( n \times n \) diagonal matrix with diagonal elements \( \{ g_t(\theta^0) \}_{t=1}^n \).

Now consider only the diagonal elements of \( E[ee'] \).

\[ E[\hat{e}] = Mg; \quad (3.3.2) \]

where \( \hat{e} \) is the vector of diagonal elements of \( ee' \), \( M \) is the matrix whose elements are the squares of the elements of \( M \) and \( g \) is the vector of diagonal elements of \( G \).

For each \( t, \ t = 1, 2, \ldots, n \) we have

\[ E[\hat{e}_t^2] = \hat{M}_t g; \]

where \( \hat{e}' = (\hat{e}_2, \ldots, \hat{e}_n) \) and \( \hat{M}_t \) is the \( t \)th row of \( M \).
We define
\[ w_t = e_t^2 - E[e_t^2] \quad t = 1, 2, \ldots, n; \]
and write
\[ e_t^2 = \hat{M}_t g + w_t \quad t = 1, 2, \ldots, n. \quad (3.3.3) \]

We note that these equations are equivalent to the equations (2.1.9) given in Section II.

Below we shall show that \( \hat{M} = I + O\left(\frac{1}{n}\right) \). Since we are interested in obtaining large sample properties of the estimators we use
\[ g_t(\theta^0) + O\left(\frac{1}{n}\right), \]
in place of \( \hat{M}_t g \) in the equation (3.3.2) to obtain
\[ e_t^2 = g_t(\theta^0) + O\left(\frac{1}{n}\right) + w_t. \quad (3.3.4) \]

In matrix notation we write
\[ \hat{e} = g(\theta^0) + W + O\left(\frac{1}{n}\right) \]
where \( \hat{e}, g(\theta^0) \) and \( W \) are the corresponding \( n \times 1 \) vectors. We will also require that the covariance matrix of the \( \{w_t\}_{t=1}^n \) be known and that the fourth moments of the \( \{w_t\}_{t=1}^n \) be uniformly bounded. The following three lemmas establish the above requirements.
Lemma 3.3.7. Given that Assumption 1 is satisfied then the following are true:

1. \( \hat{M} = I + A \) where the diagonal elements of \( A \) are \( O(\frac{1}{n}) \) and the off diagonal elements are \( O(\frac{1}{n^2}) \);

2. \( E[^\wedge e_t^2] = E[e_t^2] + O(\frac{1}{n}) \); \( E[^\wedge e_t e_s] = O(\frac{1}{n}) \).

Proof.

\[
M = I - XX'X^{-1}X' = I + 0(\frac{1}{n}) , \text{ since } (X'X)^{-1} = O(\frac{1}{n})
\]

and \( X = O(1) \).

Therefore the diagonal elements of \( \hat{M} \) are \( 1 + O(\frac{1}{n}) \) and the off diagonal elements are \( O(\frac{1}{n^2}) \).

\[
E[^\wedge ee'] = M E[ee'] M ;
\]

\[
= E[ee'] + O(\frac{1}{n}) , \text{ since } M = I + O(\frac{1}{n})
\]

and \( E[ee'] \) is diagonal.

This completes the proof. \( \square \)
In the following we obtain the higher moments of a bivariate normal in terms of the second moments.

**Lemma 3.3.8.** Given that \( \begin{pmatrix} X \\ Y \end{pmatrix} \) is bivariate normal with mean \( \begin{pmatrix} 0 \\ 0 \end{pmatrix} \) and covariance matrix \( \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \) then the following are true:

1. \( E[(X^2 - E(X^2))^2] = 2[\sigma_X^2]^2 \);

2. \( E[(X^2 - E(X^2))(Y^2 - E(Y^2))] = 2[\sigma_{XY}]^2 \);

3. \( E[(X^2 - E(X^2))^2 - E[(X^2 - E(X^2))^2]]^2] = 56[\sigma_X^4] \);

4. \( E[(X^2 - E(X^2))^2 - E[(X^2 - E(X^2))^2]](Y^2 - E(Y^2))^2 - E[(Y^2 - E(Y^2))^2]) = 32\sigma_X^2 \sigma_Y^2 \sigma_{XY}^2 + 24\sigma_{XY}^4 \).

**Proof.** These results can be obtained by matching the coefficients in series expansions of the moment generating function and cumulant generating function for the bivariate normal. 

In the following lemma we use the above two lemmas to establish that the higher moments of the distribution of \( \{w_t\}_{t=1}^n \) can be expressed in terms of the second moments. We may then use the upper bounds on the second moments to provide upper bounds on the higher moments.
Lemma 3.3.9. Given that Assumptions 1 and 5 are satisfied then the following are true:

1. \( E[w_t^2] = 2 g_t^2(\theta_0^0) + o\left(\frac{1}{n}\right), \quad t = 1, 2, \ldots, n; \)

2. \( E[w_t w_s] = O\left(\frac{1}{n^2}\right), \quad t, s = 1, 2, \ldots, n; \)

3. \( E[(w_t^2 - E[w_t^2])^2] = 56 g_t^4(\theta) + o\left(\frac{1}{n}\right), \quad t = 1, 2, \ldots, n; \)

4. \( E[(w_t^2 - E[w_t^2])(w_s^2 - E[w_s^2])] = O\left(\frac{1}{n^2}\right), \quad t, s = 1, 2, \ldots, n. \)

Proof. This result follows from Lemmas 3.3.7 and 3.3.8. By assumption 5, the \( \{e_t^0\}_{t=1}^n \) are normally distributed.

We are now ready to show the existence and properties of a simple least squares estimator of \( \alpha^0 \) in \( g(X_t; \alpha^0, \beta^0) \). We now show that:

1. A simple least squares estimator \( \hat{\alpha}_n^0 \) of \( \alpha^0 \) in \( g_t^0(\theta) \) exists;

2. \( (\hat{\alpha}_n^0 - \alpha^0) \) converges almost surely;

3. \( \sqrt{n}(\hat{\alpha}_n^0 - \alpha^0) \) is asymptotically normal;

4. \( (\hat{\alpha}_n^0 - \alpha^0) = o_p\left(\frac{1}{\sqrt{n}}\right). \)

To prove these results we follow the proofs of Jennrich (1969).
We obtain a least squares estimator $\hat{\theta}_0$ of $\theta^0$ in

$$e_t^2 = M_t g(\theta^0) + w_t \quad t = 1, 2, \ldots, n;$$

by minimizing

$$Q_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} (e_t^2 - g_t(\theta))^2$$

(3.3.3)

with respect to $\theta \in \Theta$.

The following lemma regarding the existence of a least squares estimator is given as Lemma 2 in Jennrich (1969) and is stated without proof.

**Lemma 3.3.10.** Given that:

1. $Q$ is a real valued function on $\Theta \times Y$ where $\Theta$ is a compact subset of a Euclidean space and $Y$ is a measurable space.

2. For each $\theta$ in $\Theta$, $Q(\theta, y)$ is a measurable function of $y$ and for each $y$ in $Y$ $Q(\theta, y)$ is a continuous function of $\theta$.

Then there exists a measurable function $\hat{\theta}$ from $Y$ into $\Theta$ such that for all $y$ in $Y$

$$Q(\hat{\theta}(y), y) = \inf_{\theta \in \Theta} Q(\theta, y).$$
By Lemma 3.3.10 and Assumption 1 there exists a measurable function \( \hat{\theta}_0 \) of \( y \) such that \( Q_n(\theta) \) is minimized with respect to \( \theta \) in \( \mathcal{G} \).

Before we can show that the least squares estimator is a strongly consistent estimator of \( \theta^0 \) we require the following lemma.

**Lemma 3.3.11.** Given that the following are satisfied:

1. The \( \{e_t\}_{t=1}^n \) are a sequence of random variables with mean \( 0 \), uniformly bounded second moments and \( E[e_t e_s] = O(\frac{1}{n}) \), \( t \neq s \)

\( t, s = 1, 2, \ldots, n \); 

2. \( \{g_t(\theta)\}_{t=1}^n \) is a sequence of continuous functions of \( \theta \), \( \theta \in \Theta \), \( \Theta \) a compact set;

3. \( \frac{1}{n} \sum_{t=1}^{n} g_t^2(\theta) \) converges uniformly to a finite limit \( \gamma \theta \in \Theta \); 

then \( \frac{1}{n} \sum_{t=1}^{n} g_t(\theta) e_t \overset{a.s.}{\longrightarrow} 0 \) uniformly for all \( \theta \in \Theta \).

**Proof.** We first show that

\[ \frac{1}{n} \sum_{t=1}^{n} e_t \overset{a.s.}{\longrightarrow} 0 \]

This part of the proof follows Theorem 5.1.2 in Chung (1968).
Let \( S_n = \sum_{t=1}^{n} \epsilon_t \) then

\[
E[S_n^2] = \sum_{t=1}^{n} E[\epsilon_t^2] + 2 \sum_{t=1}^{n} \sum_{s=1}^{t} E[\epsilon_t \epsilon_s]
\]

Since \( E[\epsilon_t \epsilon_s] = o\left(\frac{1}{n}\right) \) \( \exists \) a \( K \) such that

\[
|E[\epsilon_t \epsilon_s]| n \leq K \quad \forall t \neq s.
\]

Also, since \( E[\epsilon_t^2] \) is uniformly bounded \( \exists \) and \( M \) such that \( E[\epsilon_t^2] < M \forall t \).

Therefore,

\[
E[S_n^2] \leq nM + (n-1)K \leq n(M + K).
\]

\[.\] \[
E[S_n^2] \leq nM_K,
\]

where \( M_K = M + K \).

By Chebyshev's Inequality,

\[
P\left[ \left| \frac{S_n}{n} \right| > \delta \right] \leq \frac{nM_K}{n^2\delta^2} = \frac{M_K}{n\delta^2}.
\]
Consider the subsequence of the sequence of integers \( \{ n \} \) given by \( \{ n^2 \} \). Summing over this subsequence

\[
\sum_{n=1}^{\infty} P \left[ \left| \frac{S_n^2}{n^2} \right| > \delta \right] < \sum_{n=1}^{\infty} \frac{M_k}{n^2 \delta^2} < \infty .
\]

Hence by the Borel-Cantelli Lemma

\[
P \left[ \left| \frac{S_n^2}{n^2} \right| > \delta \text{ i.o.} \right] = 0 .
\]

\[\therefore \quad \frac{S_n^2}{n^2} \xrightarrow{a.s.} 0 \]

To show that \( \frac{S_n}{n} \xrightarrow{a.s.} 0 \) we must show that \( S_k \) does not differ greatly from the nearest \( S_{n^2} \).

For \( n \geq 1 \) let

\[
D_n = \max_{n^2 \leq k \leq (n+1)^2} | S_k - S_{n^2} | .
\]

Using the fact that if \( \{ Z_i \}_{i=1}^{n} \) are positive

\[
(\max Z_i)^2 \leq \sum_{i=1}^{n} Z_i^2
\]
we have

\[ D_n^2 \leq \sum_{k=n^2}^{(n+1)^2} |S_k - S_{n^2}|^2. \]

\[ \mathbb{E}[D_n^2] \leq \sum_{k=n^2}^{(n+1)^2} \mathbb{E}|S_k - S_{n^2}|^2 = \sum_{k=n^2}^{(n+1)^2} \mathbb{E}\left[ \left( \sum_{t=n^2+1}^{k} \varepsilon_t \right)^2 \right]. \]

Now

\[ \mathbb{E}\left[ \left( \sum_{t=n^2+1}^{k} \varepsilon_t \right)^2 \right] = \sum_{t=n^2+1}^{k} \mathbb{E}[\varepsilon_t^2] + 2 \sum_{t=n^2+1}^{k} \sum_{s=n^2+1}^{k} \mathbb{E}[\varepsilon_t \varepsilon_s]; \]

\[ \leq \left[ k - (n^2 + 1) \right] M + 2 \left[ k - (n^2 + 1) \right] \left[ k - 1 - (n^2 + 1) \right] \frac{K}{n}; \]

\[ < 2nM + 2 \left[ 2n \right] \left[ 2n \right] \frac{K}{n} = 2nM + 8nK. \]

\[ \mathbb{E}[D_n^2] \leq (2n+1) (2nM + 8nK); \]

\[ < n^2Q \quad Q = 6M + 24K. \]

Therefore by Chebyshev's Inequality

\[ P[D_n > n^2 \delta] \leq \frac{Q}{\delta^2 n^2} \]

and hence by the Borel-Cantelli Lemma

\[ \frac{D_n}{n^2} \overset{a.s.}{\longrightarrow} 0. \]

Now \( \frac{S_{n^2}}{n^2} \overset{a.s.}{\longrightarrow} 0 \) and \( \frac{D_n}{n^2} \overset{a.s.}{\longrightarrow} 0 \) imply that

\[ \frac{S_n}{n} \overset{a.s.}{\longrightarrow} 0 \text{ since} \]
\[
\frac{|S_k|}{K} \leq \frac{|S_{n^2}| + D_n}{n^2},
\]

where

\[
n^2 \leq k \leq (n+1)^2.
\]

Now to show that

\[
\frac{1}{n} \sum_{t=1}^{n} g_t(\theta) \varepsilon_t \xrightarrow{a.s.} 0
\]

we follow the proof of Theorem 4 in Jennrich (1969).

We have assumed that

\[
\frac{1}{n} \sum_{t=1}^{n} g_t^2(\theta)
\]

converges uniformly to a finite limit \( \psi(\theta) \in C \), \( C \) a compact set.

Therefore,

\[
\left[ \frac{1}{n} \sum_{t=1}^{n} [g_t(\theta_1) - g_t(\theta_2)]^2 \right]^{1/2}
\]

converges uniformly to a finite limit \( \psi(\theta_1, \theta_2) \in C \).
Since \( g_t(\theta) \) is a continuous function of \( \theta, \theta \in \Theta \), then for every \( \delta > 0 \), and every \( \theta_2 \) in \( \Theta \), \( \exists \) a neighborhood \( N \) of \( \theta_2 \) such that

\[
\left[ \frac{1}{n} \sum_{t=1}^{n} \left[ g_t(\theta_1) - g_t(\theta_2) \right]^2 \right]^{\frac{1}{2}} < \delta \quad \forall \theta_1 \text{ in } N
\]

and all \( n \) sufficiently large.

Now

\[
\left| \frac{1}{n} \sum_{t=1}^{n} g_t(\theta_1) \epsilon_t \right| \leq \left[ \frac{1}{n} \sum_{t=1}^{n} \left[ g_t(\theta_1) - g_t(\theta_2) \right]^2 \right]^{\frac{1}{2}} \left[ \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2 \right]^{\frac{1}{2}}
\]

\[+ \left| \frac{1}{n} \sum_{t=1}^{n} g_t(\theta_2) \epsilon_t \right| .
\]

Now let

\[ Y_t = g_t(\theta_2) \epsilon_t \quad t = 1, 2, \ldots, n. \]

Since

\[ 0 < L < g_t(\theta_2) < M, \]

then \( E[Y_t^2] \) is uniformly bounded and \( E[Y_t Y_s] = O(\frac{1}{n}), \ t \neq s \), since the second moments of \( \epsilon_t \) are uniformly bounded and \( E(\epsilon_t \epsilon_s) = O(\frac{1}{n}), \ t \neq s, \ t, \ s = 1, 2, \ldots, n \). Then by the first part of this proof

\[
\frac{1}{n} \sum_{t=1}^{n} g_t(\theta_2) \epsilon_t \xrightarrow{a.s.} 0 .
\]

Therefore for almost every \( \epsilon_t \) and for \( n \) sufficiently large \( \Theta \) is covered
by neighborhoods $N$ such that

$$
\left| \frac{1}{n} \sum_{t=1}^{n} g_t(\theta) \varepsilon_t \right| < \delta \quad \forall \theta \text{ in } N.
$$

Since $\Theta$ is compact there is a finite collection of such neighborhoods which cover $\Theta$. Therefore for almost every $\varepsilon_t$ and $n$ sufficiently large

$$
\frac{1}{n} \sum_{t=1}^{n} g_t(\theta) \varepsilon_t < \delta \text{ for all } \theta \text{ in } \Theta.
$$

Since $\delta$ arbitrary,

$$
\frac{1}{n} \sum_{t=1}^{n} g_t(\theta) \varepsilon_t \xrightarrow{a.s.} 0.
$$

This completes the proof. \(\square\)

Using the above lemma we now prove strong consistency of the least squares estimator of $\theta^0$.

**Theorem 3.3.12** Given that Assumptions 1, 3 and 5 are satisfied and that $\hat{\theta}$ is the minimum of

$$
\frac{1}{n} \sum_{t=1}^{n} (\varepsilon_t^2 - g_t(\theta))^2
$$

with respect to $\theta \in \Theta$ then $\hat{\theta} \xrightarrow{a.s.} \theta^0$. 
Proof. The proof follows Jennrich (1969), Theorem 6. As in (3.3.3) we have

\[ Q_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} [ \hat{e}_t^2 - \hat{e}_t(\theta)]^2 \]

\[ = \frac{1}{n} \sum_{t=1}^{n} [ \hat{M}_t g_t(\theta^0) - g_t(\theta)]^2 + \frac{1}{n} \sum_{t=1}^{n} w_t^2 \]

\[ + \frac{2}{n} \sum_{t=1}^{n} [ \hat{M}_t g_t(\theta^0) - g_t(\theta)] w_t , \]

where

\[ w_t = \hat{e}_t^2 - E(\hat{e}_t^2) , \quad t = 1, 2, \ldots, n. \]

By Lemma 3.3.9,

\[ E[w_t^2] = 2\hat{e}_t^2 + O\left(\frac{1}{n}\right) , \]

\[ E[w_t w_s] = 0\left(\frac{1}{n^2}\right) , \quad t \neq s , \quad t, s = 1, 2, \ldots, n ; \]

and by Lemma 3.3.7

\[ \hat{M} = I + O\left(\frac{1}{n}\right) . \]

Therefore by Lemma 3.3.11
Now consider

\[ \frac{1}{n} \sum_{t=1}^{n} [ \hat{\beta}_t (\theta^0) - \hat{\beta}_t (\hat{\theta}) ] w_t \xrightarrow{\text{a.s.}} 0. \]

By Lemma 3.3.9

\[ E[(\hat{w}_t^2 - E[\hat{w}_t^2])^2] = 56 g_t^4 (\theta^0) + o \left( \frac{1}{n} \right); \]

and

\[ E[(\hat{w}_t^2 - E[\hat{w}_t^2])(\hat{w}_s^2 - E[\hat{w}_s^2])] = o \left( \frac{1}{n^2} \right); \]

hence by Lemma 3.3.11

\[ \frac{1}{n} \sum_{t=1}^{n} (\hat{w}_t^2 - E[\hat{w}_t^2]) \xrightarrow{\text{a.s.}} 0. \]

Now by Lemma 3.3.9

\[ E[\hat{w}_t^2] = 2 g_t^2 (\theta^0) + o \left( \frac{1}{n} \right) \]

and since by Assumption 1 the \( \{\hat{\beta}_t (\theta^0)\}_{t=1}^{n} \) are uniformly bounded
exists a real number \( \rho < \infty \) such that

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} g_t^2(\theta) \longrightarrow \rho \quad \text{as} \quad n \longrightarrow \infty.
\]

\[
\therefore \quad \frac{1}{\sqrt{n}} \sum_{t=1}^{n} w_t^2 \quad \text{a.s.} \quad \longrightarrow \rho
\]

Since \( M = I + O\left(\frac{1}{n}\right) \), by Lemma 3.3.7,

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left[ \frac{\partial}{\partial \theta} g_t(\theta^0) - g_t(\theta) \right]^2 = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left[ \frac{\partial}{\partial \theta} g_t(\theta^0) - g_t(\theta) \right]^2 + O\left(\frac{1}{n}\right).
\]

By Assumption 3 there exists a \( Q(\theta) \) such that

\[
\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left[ \frac{\partial}{\partial \theta} g_t(\theta^0) - g_t(\theta) \right]^2
\]

converges uniformly to \( Q(\theta) \).

Therefore

\[
Q_n(\theta) \quad \text{a.s.} \quad \longrightarrow Q(\theta) + \rho
\]

uniformly \( \theta \in \Theta \).
Let

$$\{\hat{\theta}_n\} = \{\hat{\theta}_n(e_t^2)\}$$

be the sequence of least squares estimators, $n = 1, 2, \ldots$; let $\theta'$ be a limit point of the sequence $\{\theta_n\}$ and let $\{\theta_{n_t}\}$ be any subsequence which converges to $\theta'$. Since $Q_n(\theta)$ converges uniformly to $Q(\theta) + \rho$ and $Q(\theta)$ is continuous then

$$Q_{n_t}(\theta_{n_t}) \rightarrow Q(\theta') + \rho \quad \text{as } t \rightarrow \infty.$$

Since $\theta_{n_t}$ is a least squares estimator

$$Q_{n_t}(\theta_{n_t}) \leq Q_{n_t}(\theta^0) = \frac{1}{n} \sum_{t=1}^{n} [w_t^2] + O\left(\frac{1}{n}\right)$$

and hence

$$Q_{n_t}(\theta_{n_t}) \leq Q_{n_t}(\theta^0) \rightarrow \rho \quad \text{as } t \rightarrow \infty.$$

Thus as $t \rightarrow \infty$

$$Q_{n_t}(\theta_{n_t}) \rightarrow Q(\theta') + \rho \leq \rho.$$

Therefore,

$$Q(\theta') = 0.$$

Since by Assumption 3 the minimum of $Q(\theta)$ is unique $\theta' = \theta^0$. Thus

$$\hat{\theta}_0 \xrightarrow{a.s.} \theta^0.$$
This completes the proof. \[ \square \]

We now show that \( \sqrt{n} (\hat{\theta}_0 - \theta_0) \) is asymptotically normal and that

\[
(\hat{\theta}_0 - \theta_0) = O_p\left( \frac{1}{\sqrt{n}} \right).
\]

**Theorem 3.3.13.** Given that Assumptions 1, 2, 3 and 5 are satisfied then

1. \( \sqrt{n} (\hat{\theta}_0 - \theta_0) \xrightarrow{a} N[0, \lim_{n \to \infty} \frac{A^{-1}F'[2\hat{\theta}]FA^{-1}}{n}] \);
2. \( (\hat{\theta}_0 - \theta_0) = O_p\left( \frac{1}{\sqrt{n}} \right) \);

where \( \hat{\theta}_0 \) is the least squares estimator of \( \theta \) and

\[
\frac{1}{n} \sum_{t=1}^{n} \left( \frac{\partial g_t(\theta^0)}{\partial \theta_i} \right) \left( \frac{\partial g_t(\theta^0)}{\partial \theta_j} \right) = \left\{ \frac{1}{n} F'F \right\}_{ij} \xrightarrow{i,j=1} A_{ij}
\]

\( i,j = 1, \ldots, p+r \);

and

\[
A = \left\{ A_{ij} \right\}_{i,j=1}^{p+r}.
\]

**Proof.** The proof of (1) follows Jennrich's (1969) Theorem 7 proof.

Since \( \hat{\theta}_0 \xrightarrow{a.s.} \theta^0 \) there is a sequence \( \{ \tilde{\theta} \} \) which is equivalent
to \{ \theta \} almost everywhere and which takes its values in a convex compact neighborhood of \( \theta^0 \) which is interior to \( \Theta \). By the law of the mean given by Lemma 3.2.6 there exists a measurable \( \Theta \) valued function such that

\[
\sum_{t=1}^{n} \frac{\partial g_t(\theta)}{\partial \theta_i} (e_t^2 - g_t(\theta)) - \sum_{t=1}^{n} \frac{\partial g_t(\theta^0)}{\partial \theta_i} (e_t^2 - g_t(\theta^0))
\]


\[
\sum_{j=1}^{p+r} \left[ \sum_{t=1}^{n} \frac{\partial^2 g_t(\theta)}{\partial \theta_i \partial \theta_j} (e_t^2 - g_t(\theta)) - \sum_{t=1}^{n} \frac{\partial g_t(\theta)}{\partial \theta_i} \frac{\partial g_t(\theta)}{\partial \theta_j} \right] \begin{bmatrix} \tilde{\theta}_j - \theta^0_j \end{bmatrix}
\]

(3.3.4)

where

\[
| \tilde{\theta}_j - \theta^0_j | \leq | \tilde{\theta}_j - \theta^0_j | , \quad j = 1, 2, \ldots, p+r.
\]

Now

\[
\hat{e}_t^2 = \dot{M}_t g(\theta^0) + w_t
\]

where \( g(\theta^0) \) is the vector of diagonal elements of \( G \), \( \{g_t(\theta^0)\}_{t=1}^{n} \).

\[
\hat{e}_t^2 = g_t(\theta^0) + w_t + 0(\frac{1}{n}) .
\]

Replace \( \hat{e}_t^2 - g_t(\theta^0) \) by \( w_t + 0(\frac{1}{n}) \) in (3.3.4), multiply by \( \frac{1}{\sqrt{n}} \) and solve for
\[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial g_t(\theta^0)}{\partial \theta_i} \left( x_t + O\left( \frac{1}{n} \right) \right), \]

to obtain

\[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial g_t(\theta^0)}{\partial \theta_i}. \left[ w_t + O\left( \frac{1}{n} \right) \right] = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial g_t(\theta)}{\partial \theta_i} \left( e_t^2 - g_t(\theta) \right) \]

\[ - \sqrt{n} \sum_{j=1}^{p+r} \left[ \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 g_t(\theta)}{\partial \theta_i \partial \theta_j} \left( e_t^2 - g_t(\theta) \right) \right] \]

\[ + \frac{1}{n} \sum_{t=1}^{n} \frac{\partial g_t(\theta)}{\partial \theta_i} \frac{\partial g_t(\theta)}{\partial \theta_j} \left[ \hat{\theta}_j - \theta_j^0 \right]. \quad (3.3.5) \]

Consider the terms on the right in (3.3.5). Since

\[ \frac{\partial q(\theta^0)}{\partial \theta_i} = 0 \]

when \( \theta^0 \in \Theta \), and \( \hat{\theta} = \theta^0 \) almost everywhere for \( n \) sufficiently large, and the term

\[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial g_t(\theta)}{\partial \theta_i} \left( e_t^2 - g_t(\theta) \right) \quad (3.3.6) \]
is a multiple of
\[ \frac{\partial q(\tilde{\theta})}{\partial \theta_i} , \]

then (3.3.6) converges to zero almost everywhere.

Since \( \tilde{\theta} \overset{a.s.}{\to} \theta^0 \) and since \( \tilde{\theta} \) lies on the segment joining \( \theta^0 \) and \( \tilde{\theta} \) by Lemmas 3.3.7, 3.3.9, and 3.3.11 and Assumption 2 we have,

\[ \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 g_t(\tilde{\theta})}{\partial \theta_i \partial \theta_j} (e_t^2 - g_t(\tilde{\theta})) \overset{a.s.}{\to} 0 ; \]

and in addition since \( \frac{\partial g_t(\theta)}{\partial \theta_i} \) is uniformly bounded \( \forall \theta \in \Theta \), \( i = 1, \ldots, p+r \) there exists a matrix

\[ A = \{ A_{ij} \}_{i,j=1}^{p+r} = \lim_{n \to \infty} \frac{1}{n} F'F \]

such that

\[ \frac{1}{n} \sum_{t=1}^{n} \frac{\partial g_t(\tilde{\theta})}{\partial \theta_i} \frac{\partial g_t(\tilde{\theta})}{\partial \theta_j} \overset{a.s.}{\to} A_{ij} . \]

Thus the right hand side of (3.3.5) converges almost everywhere to

\[ \frac{1}{\sqrt{n}} \sum_{j=1}^{p+r} A_{ij} (\tilde{\theta}_j - \theta^0_j) \quad i = 1, 2, \ldots, p+r . \]
By Assumption 2 the \( \frac{\partial g_i(\theta^0)}{\partial \theta_i} \), \( i = 1, 2, \ldots, p+r \) are uniformly bounded and by Assumption 4 there exists a positive definite matrix \( A \),

\[
A = \lim_{n \to \infty} \left( \frac{F'F}{n} \right)
\]

By Lemma 3.3.9 and Assumption 1, \( E[w_t^2] \) and \( E[w_t^4] \) are uniformly bounded, \( E[w_t^2] \) is bounded above zero, \( t = 1, 2, \ldots, n \); and

\[
E[WW'] = 2G + o\left(\frac{1}{n}\right)
\]

where \( 2G \) is a diagonal matrix with elements, \( 2g_t^2(\theta^0) \), \( t = 1, 2, \ldots, n \); and \( W \) is the vector of elements \( \{w_t\}_{t=1}^n \).

Therefore by Lemma 3.2.12

\[
\frac{1}{\sqrt{n}} F'W \xrightarrow{\text{a}} N[0, \lim_{n \to \infty} \frac{F'2GF}{n}]
\]

where

\[
F = \begin{bmatrix}
\frac{\partial g_1(\theta^0)}{\partial \theta_1} & \cdots & \frac{\partial g_1(\theta^0)}{\partial \theta_{p+r}} \\
\cdots & \ddots & \cdots \\
\frac{\partial g_n(\theta^0)}{\partial \theta_1} & \cdots & \frac{\partial g_n(\theta^0)}{\partial \theta_{p+r}}
\end{bmatrix}; \quad W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}
\]
Now for sufficiently large \( n \) and almost every sequence \( \{w_t\}_{t=1}^n \)

\[
\frac{1}{\sqrt{n}} F'W = A(\hat{\phi}_n - \phi^0) \sqrt{n} \tag{3.3.7}
\]

hence

\[
\sqrt{n} (\hat{\phi}_n - \phi^0) = A^{-1} \frac{1}{\sqrt{n}} F'W \overset{\alpha}{\rightarrow} N[0, \lim_{n \to \infty} \frac{A^{-1}F'\sigma F A^{-1}}{n}].
\]

By Lemma 3.3.5,

\[
\frac{1}{n} F'W = O_p \left( \frac{1}{\sqrt{n}} \right);
\]

and since \( A^{-1} = O(1) \) then from (3.3.7)

\[
(\hat{\phi}_n - \phi^0) = O_p \left( \frac{1}{\sqrt{n}} \right).
\]

Since \( \hat{\phi} = \hat{\phi}_0 \) almost everywhere the proof is complete. \( \square \)

Thus we have shown that simple least squares estimators \( \hat{\beta} \) of \( \beta^0 \) in \( Y = X\beta^0 + e \) given by (3.3.1) and \( \hat{\phi}_0 \) of \( \phi^0 \) in \( g(X_t; \phi^0), t = 1, 2, \ldots, n \) given by (3.3.3) exist, are strongly consistent, asymptotically normal, and

\[
(\hat{\beta} - \beta^0) = O_p \left( \frac{1}{\sqrt{n}} \right), \quad (\hat{\phi}_0 - \phi^0) = O_p \left( \frac{1}{\sqrt{n}} \right).
\]

Under conditions contained in our Assumptions 1, 2, and 3, Gallant (1971) has shown that the modified Gauss Newton Procedure due to Hartley
(1961) can be used to construct the least squares estimator for nonlinear models. This procedure could be employed to get the least squares estimator of \( \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \) in \( g(X_t; \alpha^0, \beta^0) \). A similar procedure will be discussed in connection with obtaining the maximum likelihood estimator in Subsection III E.

D. A Joint Least Squares Procedure

And Its Properties

In this section we assume we have available preliminary estimators

\[
\begin{bmatrix}
\hat{\alpha}_0 \\
\hat{\beta}_0
\end{bmatrix}
\]

satisfying Assumption 7. These preliminary estimators may be the simple least squares estimators obtained in III C. We develop a joint least squares procedure which improves the estimator of \( \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \) by revising the preliminary estimator \( \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{bmatrix} \). The estimator obtained by the joint least squares procedure will be shown to have the same asymptotic distribution as the maximum likelihood estimator given in III B.

We have the following system

\[
y_t = \sum_{k=1}^{p} \beta_k^0 x_{tk} + e_t \quad t = 1, 2, \ldots, n. \quad (3.4.1)
\]
\[ e_t^2 = \hat{M}_t \, \hat{w}_t \quad t = 1, 2, \ldots, n; \]

\[ = g_t(\alpha^0, \beta^0) + o(\frac{1}{n}) + \hat{w}_t, \quad \text{by Lemma 3.3.7} \]

\[ t = 1, 2, \ldots, n; \quad (3.4.2) \]

where

\[ e_t = y_t - X_t^\prime \hat{\beta}, \quad \hat{w}_t = e_t^2 - E[e_t^2], \]

\[ \hat{\beta} = (X^\prime X)^{-1} X^\prime y, \]

\[ M = I - X(X^\prime X)^{-1} X^\prime, \]

\( \hat{M} \) is the matrix whose elements are the squares of the elements of \( M \), and \( \hat{M}_t \) is the \( t \)-th row of \( \hat{M} \). Expanding the right hand sides of (3.4.1) and (3.4.2) in a Taylor Series expansion about the preliminary estimators

\[ \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{bmatrix} \]

we obtain

\[ y_t - \sum_{k=1}^{p} \hat{\beta}_0 k x_{tk} = \sum_{k=1}^{p} (\beta^0_k - \hat{\beta}_0 k) x_{tk} + e_t \quad t = 1, 2, \ldots, n; \]

\[ e_t^2 - g_t(\alpha^0, \beta^0) = \sum_{k=1}^{p} (\beta^0_k - \hat{\beta}_0 k) \frac{\partial g_t(\hat{\theta}_0)}{\partial \beta_k} + \]

\[ r \sum_{j=1}^{r} (\alpha^0_j - \hat{\alpha}_0 j) \frac{\partial g_t(\hat{\theta}_0)}{\partial \alpha_j} + r_t + \hat{w}_t \quad t = 1, 2, \ldots, n; \]
where \( r_t \) is the Taylor Series remainder. We show that \( r_t = o_p \left( \frac{1}{n} \right) \).

\[
r_t = \sum_{k=1}^{p} \sum_{l=1}^{r} (\beta_k^0 - \bar{\beta}_k)(\beta_l^0 - \bar{\beta}_l) \frac{\partial^2 g_t(\bar{\Theta})}{\partial \beta_k \partial \beta_l}
\]

\[
+ \sum_{k=1}^{p} \sum_{j=1}^{r} (\alpha_k^0 - \bar{\alpha}_k)(\alpha_j^0 - \bar{\alpha}_j) \frac{\partial^2 g_t(\bar{\Theta})}{\partial \beta_k \partial \alpha_j}
\]

\[
+ \sum_{j=1}^{r} \sum_{l=1}^{r} (\alpha_j^0 - \bar{\alpha}_j)(\alpha_l^0 - \bar{\alpha}_l) \frac{\partial^2 g_t(\bar{\Theta})}{\partial \alpha_j \partial \alpha_l}, \quad t = 1, 2, \ldots, n
\]

where \( \bar{\Theta} = \begin{bmatrix} \bar{\alpha} \\ \bar{\beta} \end{bmatrix} \) lies between \( \Theta^0 \) and \( \hat{\Theta}_0 \). By Assumptions 1 and 2, the functions

\[
\frac{\partial^2 g_t(\Theta)}{\partial \Theta_i \partial \Theta_j}
\]

are continuous and uniformly bounded for \( \Theta \in \Theta \). Thus since

\[
(\Theta^0 - \hat{\Theta}_0) = \begin{bmatrix} \alpha^0 - \hat{\alpha}_0 \\ \beta^0 - \hat{\beta}_0 \end{bmatrix} = o_p \left( \frac{1}{\sqrt{n}} \right), \quad r_t = o_p \left( \frac{1}{n} \right).
\]

We write the system as
\[ y_t = \frac{p}{\sum_{k=1}^{p} \beta_{ok}} x_{tk} = \frac{p}{\sum_{k=1}^{p} (\beta_{k}^{0} - \beta_{ok})} x_{tk} + e_t; \]

\[ e_t^2 - g_t(\alpha_0, \beta_0) = \frac{p}{\sum_{k=1}^{p} (\beta_{k}^{0} - \beta_{ok})} \frac{\partial g_t(\theta_0)}{\partial \beta_k} \]

\[ + \sum_{j=1}^{r} (\alpha_{j}^{0} - \alpha_{oj}) + \frac{\partial g_t(\theta_0)}{\partial \alpha_j} \]

\[ + o_p \left( \frac{1}{n} \right) + w_t \quad t = 1, 2, \ldots, n. \]

In matrix notation we write

\[ \hat{z}_0 = \frac{\hat{e}^0}{\delta_0} + U; \quad (3.4.3) \]

where

\[
\hat{z}_0 = \begin{bmatrix}
(y_1 - x_1' \beta_0)
\vdots
\vdots
(y_n - x_n' \beta_0)
\end{bmatrix}, \quad U = \begin{bmatrix}
e_1 \\
\vdots \\
e_n \\
w_1 + o_p \left( \frac{1}{n} \right) \\
w_n + o_p \left( \frac{1}{n} \right)
\end{bmatrix},
\]
\[ \delta_0 = \begin{bmatrix} \alpha^0 - \alpha_0 \\ \beta^0 - \beta_0 \end{bmatrix}, \quad U = \begin{bmatrix} e \\ W + O_p \left( \frac{1}{n} \right) \end{bmatrix} \]

\[ \delta_0 = \begin{bmatrix} 0 & X \\ \delta_0 & \delta_0 \end{bmatrix} \]

\[ \delta_0 \]

\[ \delta_0 \]

\[ H_{01} = \begin{bmatrix} \frac{\partial g_1(\theta_0)}{\partial \alpha_1} & \cdots & \frac{\partial g_1(\theta_0)}{\partial \alpha_r} \\ \cdots & \cdots & \cdots \\ \frac{\partial g_n(\theta_0)}{\partial \alpha_1} & \cdots & \frac{\partial g_n(\theta_0)}{\partial \alpha_r} \end{bmatrix} \]

\[ H_{02} = \begin{bmatrix} \frac{\partial g_1(\theta_0)}{\partial \beta_1} & \cdots & \frac{\partial g_1(\theta_0)}{\partial \beta_p} \\ \cdots & \cdots & \cdots \\ \frac{\partial g_n(\theta_0)}{\partial \beta_1} & \cdots & \frac{\partial g_n(\theta_0)}{\partial \beta_p} \end{bmatrix} \]
This notation is consistent with Assumption 4 and Subsection B.

From Assumption 4 we have

\[
\psi = \begin{bmatrix} E[ee'] & 0 \\ 0 & E[WW'] \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & 2\hat{G} \end{bmatrix}.
\]

We define the estimator

\[
\delta_1^* = (H_0^T \hat{\psi}_0^* H_0) (H_0^T \hat{\psi}_0^* Z_0)
\]

(3.4.4)

where \( \hat{\psi}_0^* \) is obtained by replacing \( \theta^0 \) by \( \hat{\theta}_0^* = \begin{bmatrix} \hat{\alpha}_0^* \\ \hat{\beta}_0^* \end{bmatrix} \) in the elements \( g_t^*(\theta^0) \) of \( G \) and the elements \( 2g_t^2(\theta^0) \) of \( 2\hat{G} \). A new estimator

\[
\begin{bmatrix} \hat{\alpha}_1^* \\ \hat{\beta}_1^* \end{bmatrix}
\]

which we shall call the joint least squares estimator is defined by

\[
\begin{bmatrix} \hat{\alpha}_1^* \\ \hat{\beta}_1^* \end{bmatrix} = \delta_1^* + \begin{bmatrix} \hat{\alpha}_0^* \\ \hat{\beta}_0^* \end{bmatrix}
\]

(3.4.5)

We show that
\[
\sqrt{n} \begin{bmatrix}
\hat{\alpha}_1 - \alpha^0 \\
\hat{\beta}_1 - \beta^0 
\end{bmatrix} \xrightarrow{\alpha} N(0, \lim_{n \to \infty} \left( \frac{H'\psi^{-1}H}{n} \right))
\]

which is the same limiting distribution as that of

\[
\sqrt{n} \begin{bmatrix}
\hat{\alpha} - \alpha^0 \\
\hat{\beta} - \beta^0 
\end{bmatrix} \quad \text{where} \quad \begin{bmatrix}
\hat{\alpha} \\
\hat{\beta}
\end{bmatrix}
\]

is the consistent root of the likelihood equations discussed in Subsection IIIB. Before obtaining the asymptotic distribution of \(\sqrt{n} (\theta_1 - \theta^0) = \begin{bmatrix}
\hat{\alpha}_1 - \alpha^0 \\
\hat{\beta}_1 - \beta^0 
\end{bmatrix}\) we require the following three lemmas.

Lemma 3.4.1 below is required to show that the central limit theorem holds for a vector of independent observations from a bivariate distribution where the variables are correlated.

**Lemma 3.4.1.** Given that

1. \((x_t, y_t), t = 1, \ldots, n\) is a sequence of bivariate independent random variables with zero mean and covariance matrix

\[
\begin{bmatrix}
\sigma_{xx} & \sigma_{xy} \\
\sigma_{yx} & \sigma_{yy}
\end{bmatrix};
\]

2. \(0 < L < E[x_t^i y_t^j] < M\), \(M\) and \(L\) are positive real numbers, 

\[i, j = 0, 1, 2, 3, \quad 1 \leq (i+j) \leq 4;\]
3. \( \{ c_{1t} \}_{t=1}^{n}, \{ c_{2t} \}_{t=1}^{n} \) are sequences of uniformly bounded real constants;

4. \( A = \lim_{n \to \infty} \left[ \frac{1}{n} c_{1}^{\prime} \hat{x} x c_{1} + \frac{2}{n} c_{2}^{\prime} \hat{x} y c_{2} + \frac{1}{n} c_{2}^{\prime} \hat{y} y c_{2} \right] \) exists;

5. \( \lim_{n \to \infty} \frac{1}{n} \hat{x} x, \lim_{n \to \infty} \frac{1}{n} \hat{x} y, \lim_{n \to \infty} \frac{1}{n} \hat{y} y \) exist and define positive definite matrices;

then

\[
Z \xrightarrow{\alpha} N(0, A);
\]

where

\[
Z = \frac{1}{\sqrt{n}} C_{1}^{\prime} X + \frac{1}{\sqrt{n}} C_{2}^{\prime} Y;
\]

\[
X = \begin{bmatrix} x_{1} \\ \vdots \\ x_{n} \end{bmatrix}; \quad Y = \begin{bmatrix} y_{1} \\ \vdots \\ y_{n} \end{bmatrix}; \quad C_{1} = \begin{bmatrix} c_{11} \\ \vdots \\ c_{1n} \end{bmatrix}; \quad C_{2} = \begin{bmatrix} c_{21} \\ \vdots \\ c_{2n} \end{bmatrix}.
\]

\[
E[XX'] = \frac{1}{n} \hat{x} x; \quad E[YY'] = \frac{1}{n} \hat{y} y; \quad E[XY'] = \frac{1}{n} \hat{x} y.
\]

Proof.

\[
Z = \frac{1}{\sqrt{n}} C_{1}^{\prime} X + \frac{1}{\sqrt{n}} C_{2}^{\prime} Y;
\]

\[
= \frac{1}{\sqrt{n}} l'P
\]

where \( P \) is an \( n \times 1 \) vector of elements \( p_{t} = c_{1t}^{\prime} x_{t} + c_{2t}^{\prime} y_{t} \), \( t = 1, 2, \ldots, n \)

and \( l' = (1, 1, \ldots, 1)_{n \times 1} \).
Now

\[ 0 < 4c_t^2 < E[w_t^2] = c_{1t}^2c_{2t}^2 + c_{2t}^2c_{1t}^2 + 2c_{1t}c_{2t}c_{Xyt} \leq 4c_m^2 M \]

and

\[ E[w_t^h] = c_{1t}^h E[X_t^h] + c_{2t}^h E[Y_t^h] + 4c_{1t}c_{2t}E[X_t^3Y_t] \]

\[ + 6c_{1t}^3c_{2t}^3 E[X_t^2Y_t^2] + 4c_{1t}c_{2t}E[X_t^3Y_t^3] \leq 16c_m^4 M \]

where

\[ c_m = \max_{i=1,2} \max_{t=1,2,...,n} \{ c_{it} \} \quad \text{and} \quad c_L = \min_{i=1,2} \min_{t=1,2,...,n} \{ c_{it} \} \]

By Lemma 3.2.12

\[ Z \overset{\alpha}{\rightarrow} N(0, A). \]

The proof is complete.

**Lemma 3.4.2.** Given that Assumptions 1, 2, 4, 5, 6 and 7 are satisfied then the vector \( U \) given in (3.4.3) can be written as
\[
U = \begin{bmatrix} e \\ \hat{e} - E[e] \end{bmatrix} + \begin{bmatrix} 0 \\ d \end{bmatrix} + O_p \left( \frac{1}{n} \right)
\]

where

\[
d = O_p \left( \frac{1}{\sqrt{n}} \right), \quad E[d] = 0,
\]

\(e\) is the \(n \times 1\) vector of elements \(\{e_t^2\}_{t=1}^n\)

and

\[
E[d_t^2] = O \left( \frac{1}{n} \right); \quad E[d_t d_s] = O \left( \frac{1}{n^2} \right), \quad t \neq s,
\]

\(t, s = 1, 2, \ldots, n.\)

**Proof.**

\[
U = \begin{bmatrix} e \\ \hat{w} + O_p \left( \frac{1}{n} \right) \end{bmatrix} \quad \text{where} \quad w_t = \hat{e}_t^2 - E[\hat{e}_t^2]
\]

\(t = 1, 2, \ldots, n.\)

\[
w_t = \hat{e}_t^2 - E[\hat{e}_t^2] \quad t = 1, 2, \ldots, n;
\]

\[
= (e_t - X_t'(X'X)^{-1} X'e)^2 - E(e_t - X_t'(X'X)^{-1} X'e)^2.
\]
where $X_t$ is $t$th row of $X$,

$$
e^2_t - 2e_t X_t' (X'X)^{-1} X'e + X_t' (X'X)^{-1} X'e'X(X'X)^{-1} X_t
$$

$$
- E[e^2_t - 2e_t X_t' (X'X)^{-1} X'e + X_t' (X'X)^{-1} X'e'X(X'X)^{-1} X_t];
$$

$$
e^2_t - E[e^2_t] - 2e_t X_t' (\hat{\beta} - \beta^0) + X_t' (\hat{\beta} - \beta^0) (\hat{\beta} - \beta^0) X_t
$$

$$
+ 2X_t' (X'X)^{-1} X_t' E[e^2_t] - X_t' (X'X)^{-1} X'GX(X'X)^{-1} X_t;
$$

$$
e^2_t - E[e^2_t] + O_p \left( \frac{1}{\sqrt{n}} \right) + O_p \left( \frac{1}{n} \right)
$$

$$
+ O \left( \frac{1}{n} \right) + O \left( \frac{1}{n} \right);
$$

$$
e^2_t - E[e^2_t] + O_p \left( \frac{1}{\sqrt{n}} \right)
$$

Therefore

$$d = O_p \left( \frac{1}{\sqrt{n}} \right)
$$

since

$$d = \hat{w} - [ \hat{e} - E(\hat{e}) ] ,$$
and
\[ E[d] = E[W] - E(\hat{e} - E(\hat{e})) = 0 - 0 = 0. \]

By Lemma 3.3.9
\[ E[w_t^2] = 2g_t^2(\theta^0) + O\left(\frac{1}{n}\right) \quad t = 1, 2, \ldots, n; \]
\[ E[w_tw_s] = O\left(\frac{1}{n^2}\right) \quad t \neq s \quad t, s = 1, 2, \ldots, n. \]

By Lemma 3.3.8 and Assumption 1
\[ E[(e_t^2 - E(e_t^2))^2] = 2g_t^2(\theta^0); \]
\[ E[(e_t^2 - E(e_t^2))(e_s^2 - E(e_s^2))] = 0. \]

Then,
\[ E[d_t^2] = E[w_t^2 - (e_t^2 - E(e_t^2))^2]; \]
\[ = E[w_t^2] - E(e_t^2 - E(e_t^2))^2 = O\left(\frac{1}{n}\right); \]
and for \( t \neq s \)

\[
E[d_t d_s] = E[(w_t - (e_t^2 - E(e_t^2))(w_s - (e_s^2 - E(e_s^2)))];
\]

\[
= E[w_t w_s] - E(e_t^2 - E(e_t^2))(e_s^2 - E(e_s^2));
\]

\[
= 0\left(\frac{1}{n^2}\right).
\]

This completes the proof. \(\square\)

**Lemma 3.4.3.** Given that Assumptions 1, 2, 4, 6 and 7 are satisfied then the following are true:

1. \((\hat{\psi}_0^{-1} - \psi^{-1}) = O_p\left(\frac{1}{\sqrt{n}}\right)\);

2. \((\hat{H_0} - H) = O_p\left(\frac{1}{\sqrt{n}}\right)\);

3. \((\frac{\hat{H_0} \hat{\psi}_0^{-1} \hat{H_0}}{n}) - (\frac{H' \psi^{-1} H}{n}) = O_p\left(\frac{1}{\sqrt{n}}\right)\);

\[
(\frac{\hat{H_0} \hat{\psi}_0^{-1} \hat{H_0}}{n})^{-1} - (\frac{H' \psi^{-1} H}{n})^{-1} = O_p\left(\frac{1}{\sqrt{n}}\right);
\]
4. \( \frac{H'\psi^{-1}U}{n} - \frac{H'\psi^{-1}\left[ e - E[e] \right]}{n} = O_p \left( \frac{1}{n} \right) \);

\[ H'\psi^{-1}\left[ e - E[e] \right] = O_p \left( \frac{1}{\sqrt{n}} \right); \]

5. \( \frac{H'\hat{\psi}^{-1}U}{n} - \frac{H'\psi^{-1}U}{n} = O_p \left( \frac{1}{n} \right) \).

**Proof.**

\[ \psi^{-1} = \begin{bmatrix} G^{-1} & 0 \\ 0 & [2\hat{G}]^{-1} \end{bmatrix}, \quad G^{-1} = \begin{bmatrix} 1/g_1(\theta^0) & 0 \\ \vdots & \vdots \end{bmatrix}, \quad [2\hat{G}]^{-1} = \begin{bmatrix} \frac{1}{2}g_1^2(\theta^0) & 0 \\ \vdots & \vdots \\ 0 & \frac{1}{2}g_n^2(\theta^0) \end{bmatrix}. \]

By Assumptions 1 and 7, \( g_t(\theta^0) \) and \( g_t(\hat{\theta}_0) \) are uniformly bounded in \( \theta \) such that

\[ 0 < L_g < g_t(\theta^0), \quad g_t(\hat{\theta}_0) < M_g < \infty \]

\( t = 1, 2, \ldots, n. \)

We let \( \hat{\psi}^{-1} \) be the matrix obtained by replacing \( \theta^0 \) in \( \psi^{-1} \) by the least squares estimate \( \hat{\theta}_0 \). A typical element of \( \hat{\psi}^{-1} - \psi^{-1} \) is given by
\[
\frac{1}{\xi_t(\hat{\theta}_0)} - \frac{1}{\xi_t(\theta^0)}.
\]

Using a Taylor Series expansion we have

\[
\frac{1}{\xi_t(\hat{\theta}_0)} - \frac{1}{\xi_t(\theta^0)} = -\frac{1}{\xi_t^2(\theta^0)} \sum_{j=1}^{p+r} \left( \frac{\partial^2 \xi_t^\wedge (\theta^0)}{\partial \theta_j^2} \right) \left( \hat{\theta}_0 - \theta^0_j \right) + \text{remainder},
\]

since by Assumption 2 \( \xi_t(\theta^0) \) has continuous first and second partials.

Then

\[
\frac{1}{\xi_t(\hat{\theta}_0)} - \frac{1}{\xi_t(\theta^0)} = O_p\left( \frac{1}{\sqrt{n}} \right)
\]

since by Assumption 7

\[
(\hat{\theta} - \theta^0) = O_p\left( \frac{1}{\sqrt{n}} \right).
\]

Therefore,

\[
\psi_0^{-1} - \psi^{-1} = O_p\left( \frac{1}{\sqrt{n}} \right).
\]
To prove the second result, recall that

\[ H = \begin{bmatrix}
\frac{\partial g_1(\theta^0)}{\partial \theta_1} & \cdots & \frac{\partial g_1(\theta^0)}{\partial \theta_{p+r}} \\
\vdots & \ddots & \vdots \\
\frac{\partial g_n(\theta^0)}{\partial \theta_1} & \cdots & \frac{\partial g_n(\theta^0)}{\partial \theta_{p+r}}
\end{bmatrix} \]

and \(^\wedge H\) is obtained from \(H\) by replacing \(\theta^0\) by \(^\wedge \theta^0\). A typical element of \(^\wedge H - H\) is given by

\[
\frac{\partial g_t(\theta^0)}{\partial \theta_j} - \frac{\partial g_t(\theta^0)}{\partial \theta_j} = \sum_{k=1}^{p+r} \frac{\partial^2 g_t(\theta^0)}{\partial \theta_j \partial \theta_k} (\theta^0_k - \theta^0_k) + \text{remainder},
\]

since by Assumption 2 \(g_t(\theta)\) has continuous and uniformly bounded first and second partial derivatives

\[
\frac{\partial g_t(\theta^0)}{\partial \theta_j} - \frac{\partial g_t(\theta^0)}{\partial \theta_j} = O_p \left( \frac{1}{\sqrt{n}} \right) \quad \text{since} \quad (\hat{\theta}_0 - \theta^0) = O_p \left( \frac{1}{\sqrt{n}} \right).
\]

The third result follows directly from the first two results.
To show result 4 we have

\[
\frac{H' \psi^{-1} U}{n} = \frac{H' \psi^{-1} \left[ e \right]}{n} + \frac{H' \psi^{-1} \left[ d \right]}{n} + O_p \left( \frac{1}{n} \right).
\]

We employ Lemma 3.3.5 to show that

\[
\frac{H' \psi^{-1} \left[ e \right]}{n} = O_p \left( \frac{1}{\sqrt{n}} \right)
\]

and

\[
\frac{H' \psi^{-1} \left[ d \right]}{n} = O_p \left( \frac{1}{n} \right).
\]

Now by Assumptions 1 and 2 the elements of \( H \) and \( \psi^{-1} \) are uniformly bounded since by Assumption 1 and Lemma 3.3.9 the elements of \( \left[ e \right] \) have variances of \( O(1) \) and means 0 and since the elements of \( \left[ d \right] \) have variances of \( O\left(\frac{1}{n}\right) \) and means 0 then the results follow by Lemma 3.3.5.

To show the final result that

\[
\frac{H' \psi_{O}^{-1} U}{n} - \frac{H' \psi^{-1} U}{n} = O_p \left( \frac{1}{n} \right)
\]

we consider a typical element

\[
\frac{1}{n} \sum_{t=1}^{n} \left( h_t \gamma_t - h_t \gamma_t \right) u_t,
\]

where \( h_t, h_t, \gamma_t, \gamma_t \) are elements of \( H^0, H, \psi_{O}^{-1} \) and \( \psi^{-1} \) respectively.
Now

\[ \frac{1}{n} \sum_{t=1}^{n} \left( h_{tk} \gamma_t - h_{tk} \gamma_t \right) u_t = \frac{1}{n} \sum_{t=1}^{n} h_{tk} \left( \gamma_t - \gamma_t \right) u_t \]

\[ + \frac{1}{n} \sum_{t=1}^{n} \gamma_t \left( h_{tk} - h_{tk} \right) u_t + \frac{1}{n} \sum_{t=1}^{n} (\gamma_t - \gamma_t) (h_{tk} - h_{tk}) u_t. \]

Since by results 1 and 2 proved above

\[ \hat{h}_{tk} - h_{tk} = o_p \left( \frac{1}{\sqrt{n}} \right) \quad \text{and} \quad \gamma_t - \gamma_t = o_p \left( \frac{1}{\sqrt{n}} \right) \]

then the last of the three terms is \( o_p \left( \frac{1}{n} \right) \).

Consider the first term. We use a Taylor Series expansion

\[ (\gamma_t - \gamma_t) = \frac{1}{g_t(\theta)} - \frac{1}{g_t(\theta^0)} - \sum_{j=1}^{p+r} \frac{\partial g_t(\theta^0)}{\partial \theta_j} (\hat{\theta}_j - \theta_j^0) \frac{1}{g_t(\theta^0)} + o_p \left( \frac{1}{n} \right). \]

We can write

\[ \frac{1}{n} \sum_{t=1}^{n} h_{tk} (\gamma_t - \gamma_t) u_t = - \sum_{j=1}^{p+r} (\theta_0 j - \theta_j^0) \left[ \frac{1}{n} \sum_{t=1}^{n} \frac{\partial g_t(\theta^0)}{\partial \theta_j} \frac{1}{g_t(\theta^0)} h_{tk} u_t \right] + o_p \left( \frac{1}{n} \right). \]

As above using Lemma 3.3.5 we have
\[
\frac{1}{n} \sum_{t=1}^{n} \frac{\partial g_t(\theta^0)}{\partial \theta_j} \frac{1}{\hat{g}_t(\theta^0)} h_{tk} u_t = O_p\left(\frac{1}{\sqrt{n}}\right);
\]

and since \( \hat{\theta}_{o_j} - \theta_j^0 = O_p\left(\frac{1}{\sqrt{n}}\right) \) by Assumption 7 then

\[
\frac{1}{n} \sum_{t=1}^{n} h_{tk} (\hat{\gamma}_t - \gamma_t) u_t = O_p\left(\frac{1}{n}\right).
\]

Similarly to show that the second term is \( O_p\left(\frac{1}{n}\right) \) we use the Taylor expansion

\[
(h_{tk} - h_{tk}) = \frac{\partial g_t(\theta)}{\partial \theta_k} - \frac{\partial g_t(\theta^0)}{\partial \theta_k} ;
\]

\[
= \sum_{j=1}^{p+r} \frac{\partial^2 g_t(\theta^0)}{\partial \theta_k \partial \theta_j} (\theta_{o_j} - \theta_j^0) + O_p\left(\frac{1}{n}\right);
\]

and

\[
\frac{1}{n} \sum_{t=1}^{n} \gamma_t (h_{tk} - h_{tk}) u_t = \frac{p+r}{n} \sum_{j=1}^{p+r} \gamma_t \left[ \frac{1}{n} \sum_{t=1}^{n} \frac{\partial^2 g_t(\theta^0)}{\partial \theta_k \partial \theta_j} \gamma_t u_t \right] = O_p\left(\frac{1}{n}\right)
\]
as above.
Therefore

\[ \frac{\hat{H}_0^{\psi^{-1}} U}{n} - \frac{H^{\psi^{-1}} U}{n} = \mathcal{O}_p \left( \frac{1}{n} \right). \]

This completes the proof. \( \square \)

We now prove the main result of this section.

**Theorem 3.4.4.** Given that Assumptions 1, 2, 4, 5, 6 and 7 are satisfied then

1. \((\hat{\Theta}_1 - \Theta_0) = \mathcal{O}_p \left( \frac{1}{\sqrt{n}} \right); \)

2. \(\sqrt{n} (\hat{\Theta}_1 - \Theta_0) \rightarrow \mathcal{N} \left[ 0, \lim_{n \to \infty} \left( \frac{H^{\psi^{-1}} H}{n} \right)^{-1} \right]. \)

**Proof.** By (3.4.3) and (3.4.4)

\[
(\hat{\Theta}_1 - \Theta_0) = \left( \hat{H}_0^{\psi^{-1}} \hat{H}_0 \right) \left( H_0^{\psi^{-1}} U \right);
\]

\[
= \left[ \left( H_0^{\psi^{-1}} H_0 \right)^{-1} - (H^{\psi^{-1}} H) + (H^{\psi^{-1}} H) \right]
\]

\[
\left[ H_0^{\psi^{-1}} U - H^{\psi^{-1}} U + H^{\psi^{-1}} U \right];
\]
\[
\begin{align*}
&= (H'\psi^{-1}H)^{-1} (H'\psi^{-1}U) + [(H'\psi^{-1}H_0)^{-1} \\
&\quad - (H'\psi^{-1}H)^{-1} H'\psi^{-1}U + (H'\psi^{-1}H)^{-1} [H_0'\psi^{-1}U - H'\psi^{-1}U] \\
&\quad + [(H_0'\psi^{-1}H_0)^{-1} - (H'\psi^{-1}H)^{-1}] [H_0'\psi^{-1}U - H'\psi^{-1}U] \\
&\quad - (H'\psi^{-1}H_0)^{-1} H'\psi^{-1}U + (H'\psi^{-1}H)^{-1} [H_0'\psi^{-1}U - H'\psi^{-1}U]].
\end{align*}
\]

Therefore by Lemma 3.4.3 we have

\[
\hat{\Theta}_1 - \Theta_0 = (\frac{H'\psi^{-1}H}{n})^{-1} (\frac{H'\psi^{-1}U}{n}) + o_p(\frac{1}{n}).
\]

By Assumption 4,

\[
(\frac{H'\psi^{-1}H}{n})^{-1} = O(1),
\]

and by Lemma 3.4.3

\[
(\frac{H'\psi^{-1}U}{n}) = o_p(\frac{1}{\sqrt{n}}).
\]

Therefore

\[
\hat{\Theta}_1 - \Theta_0 = o_p\left(\frac{1}{\sqrt{n}}\right).
\]

We now show that

\[
\sqrt{n} \left(\hat{\Theta}_1 - \Theta_0\right) \xrightarrow{\mathcal{D}} N(0, \lim_{n \to \infty} (\frac{H'\psi^{-1}H}{n})^{-1}).
\]

Now by Lemma 3.4.2
\[
\frac{1}{\sqrt{n}} \mathbf{H}' \psi^{-1} \mathbf{U} = \frac{1}{\sqrt{n}} \mathbf{H}' \psi^{-1} \left[ \begin{array}{c} \mathbf{e} \\ \mathbf{e} - \mathbf{E}[\mathbf{e}] \end{array} \right] + o_p \left( \frac{1}{\sqrt{n}} \right);
\]

where

\[
\mathbf{U} = \left[ \begin{array}{c} \mathbf{e} \\ \mathbf{w} \end{array} \right] = \left[ \begin{array}{c} \mathbf{e} \\ \mathbf{e} - \mathbf{E}[\mathbf{e}] \end{array} \right] + o_p \left( \frac{1}{\sqrt{n}} \right);
\]

\[
\frac{1}{\sqrt{n}} \mathbf{H}' \psi^{-1} \left[ \begin{array}{c} \mathbf{e} \\ \mathbf{e} - \mathbf{E}[\mathbf{e}] \end{array} \right] = \frac{1}{\sqrt{n}} \left[ \begin{array}{c} \mathbf{0} \\ \mathbf{X}' \mathbf{G}^{-1} \mathbf{e} \end{array} \right] + \frac{1}{\sqrt{n}} \mathbf{F}' [2\mathbf{G}]^{-1} (\mathbf{e} - \mathbf{E}[\mathbf{e}])
\]

where

\[
\mathbf{F} = [\mathbf{H}_1, \mathbf{H}_2]; \quad \mathbf{H} = \left[ \begin{array}{cc} \mathbf{0} & \mathbf{X} \\ \mathbf{H}_1 & \mathbf{H}_2 \end{array} \right]; \quad \mathbf{\psi} = \left[ \begin{array}{cc} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{G} \end{array} \right].
\]

Let \( \lambda = \left[ \begin{array}{c} \lambda_1 \\ \lambda_2 \end{array} \right] \), \( \lambda_1 \) \( \mathbf{r} \times 1 \), \( \lambda_2 \) \( \mathbf{p} \times 1 \), be an arbitrary \((p+r) \times 1\) vector and consider
\[ q = \frac{1}{\sqrt{n}} \lambda' H' \psi^{-1} \begin{bmatrix} e \\ e - E[e] \end{bmatrix} \]

\[ = \frac{1}{\sqrt{n}} \lambda' \Sigma^{-1} e + \frac{1}{\sqrt{n}} \lambda' F'[2\hat{g}]^{-1}(\hat{e} - E[\hat{e}]). \]

By Assumptions 1 and 2 the elements of \( H \) and \( \psi \) are uniformly bounded and by Assumption \( 1 \)

\[ E[e_t(e_t^2 - E(e_t^2))] = 0 \quad t = 1, 2, \ldots, n. \]

Therefore by Lemma 3.4.1

\[ q \xrightarrow{\alpha} N(0, \lim_{n \to \infty} (\frac{1}{n} \lambda' H' \psi^{-1} H \lambda)) \]

and by Lemma 3.2.11

\[ \frac{1}{\sqrt{n}} H' \psi^{-1} \begin{bmatrix} e \\ e - E[e] \end{bmatrix} \xrightarrow{\alpha} N(0, \lim_{n \to \infty} \left( \frac{H' \psi^{-1} H}{n} \right)). \]

Now since

\[ \left( \frac{H' \psi^{-1} H}{n} \right)^{-1} = O(1) \]

then

\[ \sqrt{n} (\hat{\phi}_1 - \phi^0) = \sqrt{n} \left( \frac{H' \psi^{-1} H}{n} \right)^{-1} \left( \frac{H' \psi^{-1} U}{n} \right) + O_p \left( \frac{1}{\sqrt{n}} \right) \]

\[ \xrightarrow{\alpha} N(0, \lim_{n \to \infty} \left( \frac{H' \psi^{-1} H}{n} \right)^{-1}). \]

This completes the proof. \( \square \)
Therefore we have shown that

\[ \sqrt{n} \left( \hat{\theta}_1 - \theta^0 \right) = \sqrt{n} \begin{pmatrix} \hat{\alpha}_1 - \alpha^0 \\ \hat{\beta}_1 - \beta^0 \end{pmatrix} \]

has the same asymptotic distribution as

\[ \sqrt{n} \begin{pmatrix} \hat{\alpha} - \alpha^0 \\ \hat{\beta} - \beta^0 \end{pmatrix} \]

where \( \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \end{pmatrix} \) is defined by (3.4.5) and \( \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \) is the consistent root of the likelihood equations discussed in Section III.B.

We have shown in Lemma 3.4.3 above that

\[ \left( \frac{H_0' \psi_0^{-1} H_0}{n} \right)^{-1} \left( \frac{H_0' \psi_0^{-1} H_0}{n} \right)^{-1} = O_p \left( \frac{1}{\sqrt{n}} \right) \]

therefore

\[ \left( \frac{H_0' \psi_0^{-1} H_0}{n} \right)^{-1} \]
provides a consistent estimator of

\[ \left( \frac{H'\psi^{-1}H}{n} \right)^{-1}, \]

the covariance matrix of the asymptotic distribution of

\[ \sqrt{n} (\hat{\theta}_1 - \theta^0) = \sqrt{n} \left[ \begin{array}{c} \hat{\alpha}_1 - \alpha \\ \hat{\beta}_1 - \beta \end{array} \right]. \]

Using proofs similar to those used in Lemma 3.4.3 and Theorem 3.4.4 it can be shown that for the estimator,

\[ \beta = (X'X)^{-1} X'G_{0}^{-1}Y, \]

\[ \sqrt{n} (\beta - \beta) \xrightarrow{\alpha} N(0, \lim_{n \to \infty} \frac{X'G_{0}^{-1}X^{-1}}{n}). \]

Therefore the estimator \( \bar{\beta} \) is asymptotically equivalent to

\[ \beta^* = (X'G^{-1}X)^{-1} X'G^{-1}Y. \]

If \( \beta \) is not an unknown parameter of the covariance matrix \( G \), then the joint least squares estimator is obviously equivalent to \( \bar{\beta} \). We show below that in the case that \( \beta \) is an unknown parameter of \( G \) the joint least squares estimator of \( \beta \) is at least as efficient as \( \bar{\beta} \).
Theorem 3.4.5. Under Assumptions 1, 2, 4, 5, 6 and 7 the joint least squares estimator \( \hat{\beta}_1 \) defined in (3.4.5) is at least as efficient asymptotically for \( \beta_0 \) as the estimator

\[
\overline{\beta} = \left( X'G_0^{-1}X \right)^{-1}X'G_0^{-1}y.
\]

**Proof.** The asymptotic covariance matrix of

\[
(\hat{\theta}_1 - \theta^0) = \left( \begin{array}{c}
\hat{\alpha}_1 - \alpha^0 \\
\hat{\beta}_1 - \beta^0
\end{array} \right)
\]

is given by

\[
(H'\psi^{-1}H)^{-1} = \begin{bmatrix}
H_1' [\tilde{\omega}]^{-1} H_1 & H_1' [\tilde{\omega}]^{-1} H_2 \\
H_2' [\tilde{\omega}]^{-1} H_1 & H_2' [\tilde{\omega}]^{-1} H_2 + X'G^{-1}X
\end{bmatrix}^{-1};
\]

\[
= \begin{bmatrix}
B_1 & B_2 \\
B_2' & B_3
\end{bmatrix}
\]

The asymptotic covariance matrix of \((\overline{\beta} - \beta)\) is \((X'G^{-1}X)^{-1}\). We want to show that

\[
(X'G^{-1}X)^{-1} - B_3
\]

is positive semi definite. Since by Assumption 4 \((H'\psi^{-1}H)^{-1}\) is positive definite, \(B_3\) is positive definite. Now by 2.7 in C.R. Rao (1965a, p. 29)
\[ B_3 \text{ can be written as } \left\{ H_2' [ \mathbf{2G} ]^{-1} H_2 + X'G^{-1}X + H_2' [ \mathbf{2G} ]^{-1} H_1 \left( H_1' [ \mathbf{2G} ]^{-1} H_1 \right)^{-1} H_1 [ \mathbf{2G} ]^{-1} H_2 \right\}^{-1}. \]

We want to show that
\[
(X'G^{-1}X)^{-1} - \left\{ (X'G^{-1}X) + \left[ H_2' [ \mathbf{2G} ]^{-1} H_2 + H_2' [ \mathbf{2G} ]^{-1} H_1 \left( H_1' [ \mathbf{2G} ]^{-1} H_1 \right)^{-1} H_1 [ \mathbf{2G} ]^{-1} H_2 \right] \right\}^{-1}
\]

is positive semi definite.

By a special case of 2.0 in C. R. Rao (1965a, p. 29) if \( A \) and \( D \) are \( mxm \) nonsingular matrices then \((A+D)^{-1} = A^{-1} - A^{-1}(A^{-1}+D^{-1})A^{-1}\) hence \((3.4.8)\) becomes

\[
(X'G^{-1}X)^{-1} \left\{ (X'G^{-1}X)^{-1} + \left[ H_2' [ \mathbf{2G} ]^{-1} H_2 \\
+ H_2' [ \mathbf{2G} ]^{-1} H_1 \left( H_1' [ \mathbf{2G} ]^{-1} H_1 \right)^{-1} H_1 [ \mathbf{2G} ]^{-1} H_2 \right] \right\} (X'G^{-1}X)^{-1}
\]

which is positive semi definite since by Assumption 4 \((X'G^{-1}X)^{-1}\) and

\[
\left[ H_2' [ \mathbf{2G} ]^{-1} H_2 + H_2' [ \mathbf{2G} ]^{-1} H_1 \left( H_1' [ \mathbf{2G} ]^{-1} H_1 \right)^{-1} H_1 [ \mathbf{2G} ]^{-1} H_2 \right]^{-1}
\]

are positive definite. This completes the proof. \(\square\)
In the case that $\beta$ is not an unknown parameter of the covariance matrix $G$ we can show that

$$\bar{\beta} = (X'G_0^{-1}X')^{-1}X'G_0^{-1}Y$$

is an unbiased estimator of $\beta^0$. The following theorem is given in Kakwani (1967).

**Theorem 3.4.6.** For the class $\mathcal{L}$ of linear models of Section I if the following conditions are satisfied:

1. $\mathbf{e}$ is symmetrically distributed about 0;
2. $\hat{G}$ is an even function of $\mathbf{e}$ in the sense that $\hat{G}$ is invariant with respect to a simultaneous change in sign of all the elements of $\mathbf{e}$;
3. $E[\bar{\beta}]$ exists, where $\bar{B} = (X'G_0^{-1}X')^{-1}X'G_0^{-1}Y$ then $\bar{\beta}$ is an unbiased estimator of $\beta^0$.

**Proof.** This proof follows Kakwani (1967)

$$\bar{\beta} - \beta = (X'G_0^{-1}X')^{-1}X'G_0^{-1}\mathbf{e}$$

$$= [f(e)]\mathbf{e}$$

where

$$f(e) = (X'G_0^{-1}X')^{-1}X'G_0^{-1}$$
Since $G_0$ is an even function of $e$ then $f(e)$ is an even function of $e$. Thus $f(e) = f(-e)$. Therefore if the sign of all the elements of $e$ are changed simultaneously then the signs on all the elements of $(\bar{\beta} - \beta)$ are changed simultaneously. By condition 1), $e$ is symmetrically distributed about 0 which implies that the probability density function of $e$ and $-e$ are the same. Therefore $(\bar{\beta} - \beta^0)$ and $(\beta^0 - \bar{\beta})$ have the same probability density function, and $\bar{\beta}$ is symmetrically distributed about $\beta^0$. Since by condition 3, $E[\bar{\beta}]$, exists $\bar{\beta}$ is unbiased for $\beta^0$. This completes the proof.

We now show that under Assumptions 1, 2, 3, 4, 5, and 6 the conditions of Theorem 3.4.6 are met. By Assumption 5 $e$ is multivariate normal with mean 0 and hence is symmetrically distributed about 0. We now show that $G_0$ is an even function of $e$. The diagonal elements of $G_0$ are given by

$$\left\{ g_t(\hat{\Theta}_0) \right\}_{t=1}^n$$

where $\hat{\Theta}_0$ is the preliminary estimator of $\Theta$. In the case of $\beta$ not being an unknown parameter in $G$, $\hat{\Theta}_0 = \hat{\alpha}_0$. Now $\hat{\alpha}_0$ is a function of $\left\{ e^2_t \right\}_{t=1}^n$ where

$$e^2_t = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{i} m_{j} e_i e_j$$

then clearly $e^2_t$ is an even function of the elements of $e$ and hence $\hat{\alpha}_0$ is an even function of the elements of $e$. $E[\bar{\beta}]$ exists since by
Assumption 4 \( \lim_{n \to \infty} \left( \frac{X'G_0X}{n} \right)^{-1} \) exists, and by Assumption 1 the elements of \( X \) and the elements of \( G_0 \) are uniformly bounded for \( \forall \theta \in \Theta, \ t = 1, 2, \ldots, n \). Therefore by Theorem 3.4.6 \( \bar{\beta} \) is an unbiased estimator of \( \beta \) when \( \beta \) is not an unknown parameter of \( G \).

E. Convergence of the Joint Least Squares Estimator to the Maximum Likelihood Estimator

In Section III B we showed that the maximum likelihood estimator exists and is a consistent estimator of \( \theta^0 \). The likelihood equations were shown to have a root, \( \tilde{\theta} \), with probability 1 as \( n \to \infty \) which is consistent for \( \theta^0 \). This root is the only consistent root of the likelihood equations. The asymptotic distribution of \( \sqrt{n} (\tilde{\theta} - \theta^0) \) is normal with mean 0 and covariance matrix

\[
\lim_{n \to \infty} \left( \frac{H'\psi^{-1}H}{n} \right)^{-1}.
\]

We also observed that there is no known analytic expression for the solution of the likelihood equations.

In Section III D we defined a joint least squares estimator and showed that it has the same asymptotic distribution as the consistent root of the likelihood equations. We now demonstrate that by iterating the joint least squares procedure in connection with the modification used by Hartley
(1961) for nonlinear least squares estimation we obtain a sequence of estimators that converges to a local maximum of the likelihood function.

The joint least squares estimator developed in IID is given by

$$\begin{bmatrix}
\hat{\alpha}_1 \\
\hat{\beta}_1
\end{bmatrix} =
\begin{bmatrix}
\hat{\alpha}_0 \\
\hat{\beta}_0
\end{bmatrix} + \delta^*_1$$

where

$$\delta^*_1 = \left(H_0^\prime \psi_0^{-1} h_0\right)^{-1} \left(H_0^\prime \psi_0^{-1} Z_0\right)$$

as given in (3.5.1). In Section III B the matrix form of the likelihood equations is given by

$$H_0^\prime \psi_0^{-1} Z_0 = 0$$

as in (3.2.4). We obtain an iterative procedure such that

$$H_0^\prime \psi_0^{-1} Z_0$$

converges to zero and at the same time the likelihood function converges to a local maximum, where $H_0^\prime$, $\psi_0^\prime$ and $Z_0^\prime$ are functions of the current estimator.

We begin by reviewing Hartley's (1961) Modified Gauss Newton Procedure for obtaining the least squares estimator of $\gamma_0$ in the model

$$y_t = f(X_t; \gamma_0) + e_t, \quad t = 1, 2, \ldots, n;$$
where $\gamma^0$ is an unknown p x 1 parameter vector. We want to determine the value of $\gamma$ that minimizes the quantity

$$P(\gamma) = \sum_{t=1}^{n} (y_t - f(X_t; \gamma))^2 .$$

The necessary assumptions are given below.

The function $f(X_t; \gamma)$ and its first and second derivatives are continuous functions of $\gamma$ for all $\{ X_t \}_{t=1}^{n} \text{ and for all } \gamma \in A \:$.

The matrix $D'D$ where

$$D = \begin{bmatrix} \frac{\partial f(X_t; \gamma)}{\partial \gamma_i} \end{bmatrix}_{n \times p}$$

$$i = 1, 2, \ldots, p$$

$$t = 1, 2, \ldots, n ; \text{ (3.5.2)}$$

is positive define for all $\gamma$ in a set $A$ containing the true value $\gamma^0$. $P(\gamma^0)$ is less than $P(\gamma)$ for any $\gamma$ on the boundary of $A$.

There does not exist $\gamma'$ and $\gamma''$ in $A$ such that
\[
\frac{\partial P(Y')}{\partial \gamma_i} = \frac{\partial P(Y'')}{\partial \gamma_i} \quad i = 1, 2, \ldots, p;
\]

and

\[P(Y') = P(Y'').\]

A preliminary estimate \(^\hat{\gamma}_0\) of \(\gamma\) exists in the interior of \(A\).

We replace \(f(X_t; \gamma)\) by a first order Taylor Series approximation obtained by expanding \(f(X_t; \gamma)\) about \(^\hat{\gamma}_0\)

\[f(X_t; \gamma) = f(X_t; ^\hat{\gamma}_0) + \sum_{i=1}^{p} \frac{\partial f(X_t; ^\hat{\gamma})}{\partial \gamma_i} (\gamma_i - ^\hat{\gamma}_0) + \text{remainder}.\]

We define

\[P'(\gamma, ^\hat{\gamma}_0) = \sum_{t=1}^{n} \left[ y_t - f(X_t; ^\hat{\gamma}_0) - \sum_{i=1}^{p} \frac{\partial f(X_t; ^\hat{\gamma}_0)}{\partial \gamma_i} (\gamma_i - ^\hat{\gamma}_0) \right]^2.\]

Minimizing \(P'(\gamma, ^\hat{\gamma}_0)\) with respect to \(\gamma\) and solving for \((\gamma - ^\hat{\gamma}_0)\) the following matrix equation is obtained:

\[\left( ^\hat{\gamma}_1 - ^\hat{\gamma}_0 \right) = \left( D_0' D_0 \right)^{-1} D_0' V_0; \quad (3.5.3)\]

where \(^\hat{\gamma}_1\) is the value of \(\gamma\) that minimizes \(P'(\gamma, ^\hat{\gamma}_0)\) and
where

\[ D_0 \text{ is given by (3.5.2) with } \gamma \text{ replaced by } \gamma^* \text{ and } V \text{ is the } n \times 1 \text{ vector of elements} \]

\[ \{y_t - f(X_t; \gamma) \}_{t=1}^n. \]

A new estimator \( \gamma_1 \) of \( \gamma \) is obtained from

\[ \gamma_1 = \lambda_1 \gamma^*_1 + \gamma_0, \quad 0 \leq \lambda_1 \leq 1, \]

where \( \lambda_1 \) is obtained by substituting \( \lambda \gamma^*_1 + \gamma_0 \) for \( \gamma \) in \( P(\gamma) \) and minimizing with respect to \( \lambda \). Equation (3.5.3) is then used to obtain

\[ (\gamma^*_2 - \gamma_1) = (D_1' D_1)^{-1} D_1' V_1 \]

where \( D_1 \) and \( V_1 \) are obtained from \( D_0, V_0 \) by replacing \( \gamma_0 \) by \( \gamma_1 \).

A new estimator

\[ \gamma_2, \quad \gamma_2 = \lambda_2 \gamma^*_2 + \gamma_1, \quad 0 \leq \lambda_2 \leq 1 \]

is obtained as above by minimizing \( P(\lambda \gamma^*_2 + \gamma_1) \) with respect to \( \lambda \).
This process is repeated until the change in the estimator becomes arbitrarily small.

We shall obtain an iterative procedure similar to the Gauss Newton Procedure to obtain the maximum likelihood estimator of \[ \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \]. The logarithm of the likelihood function is given by (3.2.1). Maximizing this function with respect to \[ \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \] is equivalent to minimizing the function

\[
\ell(\alpha, \beta) = \sum_{t=1}^{n} \log g_t(\alpha, \beta) + \sum_{t=1}^{n} \frac{(y_t - x_t' \beta)^2}{g_t(\alpha, \beta)}
\] (3.5.4)

with respect to \[ \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \]. We now construct an iterative estimation procedure.

Construction 3.5.1.

\[
\frac{\partial \ell}{\partial \alpha_s} = - \sum_{t=1}^{n} \frac{\partial g_t(\alpha, \beta)}{\partial \alpha_s} \frac{[(y_t - x_t' \beta)^2 - g_t(\alpha, \beta)]}{g_t^2(\alpha, \beta)} = 0 \quad (3.5.5)
\]

\[
s = 1, 2, \ldots, r.
\]

\[
\frac{\partial \ell}{\partial \beta_u} = - \sum_{t=1}^{n} \frac{\partial g_t(\alpha, \beta)}{\partial \beta_u} \frac{[(y_t - x_t' \beta)^2 - g_t(\alpha, \beta)]}{g_t^2(\alpha, \beta)}
\]

\[
- 2 \sum_{t=1}^{n} \frac{(y_t - x_t' \beta) x_{tu}}{g_t(\alpha, \beta)} = 0 \quad , \quad u = 1, 2, \ldots, p.
\] (3.5.6)
Let \( \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{bmatrix} \) be a preliminary estimator of \( \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \) with the properties given in Assumption 8. Assume also that Assumptions 1, 2, 4, and 5 are satisfied. To solve the system of equations given by (3.5.5), (3.5.6) we expand \( g_t(\alpha, \beta) \) and \( X_t' \beta \) in Taylor Series about \( \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{bmatrix} \) as follows:

\[
X_t' \beta = X_t' \hat{\beta}_0 + X_t' (\beta - \hat{\beta}_0), \quad t = 1, 2, \ldots, n. \tag{3.5.7}
\]

\[
g_t(\alpha, \beta) = g_t(\hat{\alpha}_0, \hat{\beta}_0) + \frac{\partial g_t(\hat{\alpha}_0, \hat{\beta}_0)}{\partial \beta_k} (\beta_k - \hat{\beta}_0) + \sum_{k=1}^{p} \frac{\partial g_t(\hat{\alpha}_0, \hat{\beta}_0)}{\partial \beta_k} (\beta_k - \hat{\beta}_0) + \sum_{j=1}^{r} \frac{\partial g_t(\hat{\alpha}_0, \hat{\beta}_0)}{\partial \alpha_j} (\alpha_j - \hat{\alpha}_j) + \text{remainder} \quad t = 1, 2, \ldots, n. \tag{3.5.8}
\]
We replace $x_t^r$ in $(y_t - x_t^r)$ in (3.5.6) by

$$
(X_t^r e_0 + x_t^r (\beta - \beta_0^*))\quad t = 1, 2, \ldots, n.
$$

We replace $[(y_t - x_t^r)^2 - e_t(\alpha, \beta)]$ in (3.5.5) and (3.5.6) by

$$
[(y_t - x_t^r e_0)^2 - e_t(\alpha_0, \beta_0)]
$$

$$
- \frac{p}{\Sigma_{k=1}^{p} \frac{\partial e_t(\alpha_0, \beta_0)}{\partial \beta_k} (\beta_k - \beta_{0k})} - \frac{r}{\Sigma_{j=1}^{r} \frac{\partial e_t(\alpha_0, \beta_0)}{\partial \alpha_j} (\alpha_j - \alpha_{0j})].
$$

The functions $\frac{\partial e_t(\alpha, \beta)}{\partial \alpha_s}$, $\frac{\partial e_t(\alpha, \beta)}{\partial \beta_u}$ and $e_t(\alpha, \beta)$ are replaced by

$$
\frac{\partial e_t(\alpha_0, \beta_0)}{\partial \alpha_s}, \quad \frac{\partial e_t(\alpha_0, \beta_0)}{\partial \beta_u} \quad \text{and} \quad e_t(\alpha_0, \beta_0) \quad s = 1, 2, \ldots, r;
$$

$$
u = 1, 2, \ldots, v; \quad t = 1, 2, \ldots, n.
$$

The equations (3.5.5) and (3.5.6) become

$$
\Sigma_{t=1}^{n} \left( \frac{\partial e_t(\alpha_0, \beta_0)}{\partial \alpha_s} [(y_t - x_t^r e_0)^2 - e_t(\alpha_0, \beta_0)] - \frac{p}{\Sigma_{k=1}^{p} \frac{\partial e_t(\alpha_0, \beta_0)}{\partial \beta_k} (\beta_k - \beta_{0k})} - \frac{r}{\Sigma_{j=1}^{r} \frac{\partial e_t(\alpha_0, \beta_0)}{\partial \alpha_j} (\alpha_j - \alpha_{0j})] \right) \big/ 2 e_t(\alpha_0, \beta_0) = 0, \quad s = 1, 2, \ldots, r.
\[
\sum_{t=1}^{n} \left( \frac{\partial g_t(\alpha_0, \beta_0)}{\partial \beta_k} \right) \left[ (y_t - x_t^\top \beta_0)^2 - g_t(\alpha_0, \beta_0) \right] - \frac{\partial g_t(\alpha_0, \beta_0)}{\partial \alpha_j} (\alpha_j - \alpha^*_j) \right] / \partial g_t(\alpha_0, \beta_0) +
\]

\[
\sum_{t=1}^{n} (y_t - x_t^\top \beta_0 - x_t^\top (\beta - \beta^*_0) ) x_{tj} / g_t(\alpha_0, \beta_0) = 0
\]

\[
u = 1, 2, \ldots, p.
\]

In matrix notation these equation are given by

\[
(H_0^\top \psi_0^{-1} H_0) \delta_1 - H_0^\top \psi_0^{-1} Z_0 = 0; \quad (3.5.9)
\]

where $H_0$, $\psi_0$, and $\delta_1$ are given by (3.4.3) and (3.4.4) of IID.

We solve (3.5.9) for $\delta_1$

\[
\delta_1^* = (H_0^\top \psi_0^{-1} H_0) (H_0^\top \psi_0^{-1} Z_0)
\]

Since by Assumption 7, $[\alpha_0, \beta_0]^\top$ is in $S_1$ then by Assumption 4 $\delta_1^*$ exists. We then compute the value of $\lambda, \lambda_1$, $0 \leq \lambda \leq 1$, that minimizes
with respect to \( \lambda \) on \( \mathcal{T} \) given by Assumption 8, where

\[
\delta_1 = \begin{bmatrix}
\delta_{1\alpha}^* \\
\delta_{1\beta}^*
\end{bmatrix},
\]

A new estimator is then obtained by computing

\[
\begin{bmatrix}
\hat{\alpha}_1 \\
\hat{\beta}_1
\end{bmatrix} = \lambda_1 \delta_1 + \begin{bmatrix}
\hat{\alpha}_0 \\
\hat{\beta}_0
\end{bmatrix}.
\]

This process is repeated by replacing \( \begin{bmatrix}
\hat{\alpha}_0 \\
\hat{\beta}_0
\end{bmatrix} \) by \( \begin{bmatrix}
\hat{\alpha}_1 \\
\hat{\beta}_1
\end{bmatrix} \) in \( H_0, \bar{\psi}, Z_0 \) to obtain \( \hat{H}_1, \hat{\psi}_1, \hat{Z}_1 \) and

\[
\delta_2 = (\hat{H}_1 \hat{\psi}_1^{-1} \hat{H}_1^{-1}) (\hat{H}_1 \hat{\psi}_1^{-1} \hat{Z}_1).
\]
A new value of \( \lambda_2 \) is obtained by minimizing

\[
el( \hat{\alpha}_1 + \lambda \hat{\delta}_{2x}, \hat{\beta}_1 + \lambda \hat{\delta}_{2\beta} ) \text{ on } \overline{R}
\]

with respect to \( \lambda, 0 \leq \lambda \leq 1 \). The new estimator of \( \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \) is given by

\[
\begin{bmatrix}
\hat{\alpha}_2 \\
\hat{\beta}_2
\end{bmatrix} = \lambda \hat{\delta}_2 + \begin{bmatrix}
\hat{\alpha}_1 \\
\hat{\beta}_1
\end{bmatrix}.
\]

This process is repeated until the change in the estimator becomes arbitrarily small. At the \( i^{th} \) iteration we have

\[
\hat{\delta}_i = (H_{i-1} \psi_{i-1}^{-1} H_{i-1})^{-1} \hat{\alpha}_i^{-1} \hat{\beta}_i^{-1} ; \quad (3.5.10)
\]

\[
\begin{bmatrix}
\hat{\alpha}_i \\
\hat{\beta}_i
\end{bmatrix} = \lambda \hat{\delta}_i + \begin{bmatrix}
\hat{\alpha}_{i-1} \\
\hat{\beta}_{i-1}
\end{bmatrix} . \quad (3.5.11)
\]
We now show that the sequence of estimators \( \left\{ \begin{bmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{bmatrix} \right\} \) obtained by this procedure are members of the set \( T \) of Assumption 8. We also show that the sequence converges to a limit \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \) an interior point of \( T \) where

\[
\frac{\partial L(\alpha, \beta)}{\partial \alpha_s} = 0 \quad s = 1, 2, \ldots, r
\]

\[
\frac{\partial L(\alpha, \beta)}{\partial \beta_u} = 0 \quad u = 1, 2, \ldots, p.
\]

**Theorem 3.5.2.** Given that Assumptions 1, 2, 4, 5, 7 and 8 are satisfied then the sequence of estimators given by Construction 3.5.1 satisfy the following:
1. The estimators in the sequence \( \left\{ \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \right\}_{i=1}^{\infty} \) are interior points of the set \( T \) of Assumption 8;

2. The sequence of estimators \( \left\{ \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \right\}_{i=1}^{\infty} \) converge to a limit \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \) which is interior to \( T \) and where

\[
\frac{\partial \ell(\tilde{\alpha}, \tilde{\beta})}{\partial \alpha_s} = 0 \quad s = 1, 2, \ldots, r
\]

\[
\frac{\partial \ell(\tilde{\alpha}, \tilde{\beta})}{\partial \beta_u} = 0 \quad u = 1, 2, \ldots, p.
\]

**Proof.** The proof follows Gallant (1971), page 36.

We first show that the sequence of estimators generated by the process described in Construction 3.5.1 are members of \( T \). We shall use...
proof by induction. Assume \( \alpha_1 \) \( \beta_1 \) \( \gamma_1 \) where is the interior of \( T \). We minimize

\[
\ell (\alpha_i + \lambda \delta_{i\beta}, \beta_i + \lambda \delta_{i\beta})
\]

with respect to \( \lambda, \ 0 \leq \lambda \leq 1 \), and obtain \( \lambda_i \). Consider the set

\[
\hat{T} = \left\{ \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \overline{T} : \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} + \lambda \begin{bmatrix} \delta_{i\alpha} \\ \delta_{i\beta} \end{bmatrix}, \ 0 \leq \lambda \leq 1 \right\}.
\]

Since \( \overline{T} \) the closure of \( T \) is convex by Assumption 8, \( \hat{T} \) is a closed bounded line segment contained in \( \overline{T} \). By Assumptions 1 and 2 \( \ell \) is continuous over \( \hat{T} \). Since \( \hat{T} \) is closed and bounded and \( \ell \) is continuous over \( \hat{T} \) there is an \( \begin{bmatrix} \alpha'' \\ \beta'' \end{bmatrix} \) minimizing \( \ell \) over \( \hat{T} \). Therefore there is a \( \lambda_i \), such that

\[
\begin{bmatrix} \alpha'' \\ \beta'' \end{bmatrix} = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} + \lambda_k \begin{bmatrix} \delta_{i\alpha} \\ \delta_{i\beta} \end{bmatrix},
\]

minimizes

\[
\ell (\alpha_i + \lambda \delta_{i\alpha}, \beta_i + \lambda \delta_{i\beta}) \quad \text{over} \quad 0 \leq \lambda \leq 1.
\]
Now \[
\begin{bmatrix}
\alpha'' \\
\beta''
\end{bmatrix}
\] is either a boundary point of \( \overline{T} \) or an interior point. Since \( T \) is convex, \( T \) and \( \overline{T} \) necessarily have the same boundary points. Let 

\[ \tilde{\ell} \]

be the

\[ \inf \{ \ell(\alpha, \beta); \begin{bmatrix}
\alpha \\
\beta
\end{bmatrix} \text{ a boundary point of } T \} . \]

If \[
\begin{bmatrix}
\alpha'' \\
\beta''
\end{bmatrix}
\]
is a boundary point of \( T \) we have by Assumption 8

\[ \tilde{\ell} \leq \ell(\alpha'', \beta'') \leq \ell(\alpha, \beta) < \tilde{\ell} \]

which is a contradiction. Hence \[
\begin{bmatrix}
\alpha'' \\
\beta''
\end{bmatrix}
\]
is an interior point of \( T \).

Since by Construction 3.5.1

\[
\begin{bmatrix}
\alpha_{i+1}^\wedge \\
\beta_{i+1}^\wedge
\end{bmatrix}
= \begin{bmatrix}
\alpha'' \\
\beta''
\end{bmatrix}
\]

we have established conclusion 1.

We now establish conclusion 2. By Construction 3.5.1

\[ 0 \leq \ell(\alpha_{i+1}^\wedge, \beta_{i+1}^\wedge) \leq \ell(\alpha_i^\wedge, \beta_i^\wedge) \]

and
\( \ell(\alpha_i, \beta_i) \) goes to a finite limit \( \ell^* \) as \( k \to \infty \). Since \( \overline{T} \) is closed and bounded the sequence \( \left\{ \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \right\} \) in \( \overline{T} \) must have a convergent subsequence \( \left\{ \begin{bmatrix} \alpha_{i_n} \\ \beta_{i_n} \end{bmatrix} \right\} \) with limit \( \begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \) in \( \overline{T} \). By the continuity of \( \ell \) over \( \overline{T} \),

\[
\ell(\alpha_{i_n}, \beta_{i_n}) \to \ell(\alpha^*, \beta^*)
\]

hence

\[
\ell^* = \ell(\alpha^*, \beta^*)
\]

Now \( \begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \) is either an interior point or a boundary point of \( \overline{T} \) and hence of \( T \). If \( \begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \) were a boundary point of \( T \) then by Assumption 8

\[
\tilde{\ell} \leq \ell(\alpha^*, \beta^*) \leq \ell(\alpha^0, \beta^0) < \tilde{\ell}
\]
which is a contradiction hence \[
\begin{bmatrix}
\alpha^* \\
\beta^*
\end{bmatrix}
\] is an interior point of \( T \). By Assumptions 1, 2, and 4 the function

\[
\delta_i^* = \left( H_{i-1} H_{i-1} \right)^{-1} \delta_i
\]

is continuous on \( T \). Thus for the convergent subsequence

\[
\left\{ \begin{bmatrix}
\alpha_{i-1}^* \\
\beta_{i-1}^*
\end{bmatrix} \right\}, \quad \lim_{n \to \infty} \delta_i^* = \delta^*
\]

where \( \delta^* \) is obtained by replacing \( \begin{bmatrix}
\alpha_i \\
\beta_i
\end{bmatrix} 
\) by \( \begin{bmatrix}
\alpha_{i-1} \\
\beta_{i-1}
\end{bmatrix} \) in the equation above for \( \delta_i^* \).

We now show that

\[
\delta^* = 0.
\]
Suppose that $\delta^* \neq 0$ and consider the function

$$q(\lambda) = \varepsilon(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)$$

where $\varepsilon(\alpha, \beta)$ is given by (3.5.4). For $\lambda \in [-\eta, \eta]$ where $0 < \eta \leq 1$

$$\begin{bmatrix}
\alpha^* + \eta \delta^*_\alpha \\
\beta^* + \eta \delta^*_\beta
\end{bmatrix}$$

are interior points of $T$.

Now

$$q'(\lambda) = -\sum_{t=1}^{n} \frac{\partial \varepsilon_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta) \left( y_t - \sum_{k=1}^{p} x_{tk}(\beta^*_k + \lambda \delta^*_\beta_k) \right)^2 - \varepsilon_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)^2}{\varepsilon_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)^2}$$

$$- 2 \sum_{t=1}^{n} \frac{\left( y_t - \sum_{k=1}^{p} x_{tk}(\beta^*_k + \lambda \delta^*_\beta_k) \right) \sum_{k=1}^{p} x_{tk} \delta^*_k}{\varepsilon_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)}.$$
Also,

\[
\frac{\partial g_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)}{\partial \lambda} = \sum_{j=1}^{r} \frac{\partial g_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial \lambda} \\
+ \sum_{k=1}^{p} \frac{\partial g_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)}{\partial \beta_k} \frac{\partial \beta_k}{\partial \lambda} \quad \forall t = 1, 2, ..., n;
\]

where

\[
\alpha = \alpha^* + \lambda \delta^*_\alpha \quad \text{and} \quad \beta = \beta^* + \lambda \delta^*_\beta.
\]

Further

\[
\frac{\partial g_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)}{\partial \lambda} \bigg|_{\lambda = 0} = \sum_{j=1}^{r} \frac{\partial g_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)}{\partial \alpha_j} \bigg|_{\lambda = 0} \frac{\partial \alpha_j}{\partial \lambda} \\
+ \sum_{k=1}^{p} \frac{\partial g_t(\alpha^* + \lambda \delta^*_\alpha, \beta^* + \lambda \delta^*_\beta)}{\partial \beta_k} \bigg|_{\lambda = 0} \frac{\partial \beta_k}{\partial \lambda} \\
= \sum_{j=1}^{r} \frac{\partial g_t(\alpha^*, \beta^*)}{\partial \alpha_j} \delta^*_\alpha_j + \sum_{k=1}^{p} \frac{\partial g_t(\alpha^*, \beta^*)}{\partial \beta_k} \delta^*_\beta_k.
\]
Thus

\[ q'(0) = -2 \sum_{t=1}^{n} \frac{(y_t - \sum_{k=1}^{p} x_{tk} \beta_k^*) \sum_{k=1}^{p} x_{tk} \delta_k^*}{[\varepsilon_t(\alpha^*, \beta^*)]} \]

\[ = -2z^* \psi^{-1} h^* \delta^*; \]

where

\[ z^*, \psi^{-1}, h^*, \text{ and } \delta^* \]

correspond to

\[ \hat{z}_i, \hat{\psi}_i, \hat{h}_i, \text{ and } \hat{\delta}_i \]

with \[ \begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \] replacing \[ \begin{bmatrix} \alpha_i^* \\ \beta_i^* \end{bmatrix} \].
Now
\[ \delta^* = (H^*, \psi^{-1} H^*)^{-1} (H^*, \psi^{-1} Z^*) \]

hence
\[ (H^*, \psi^{-1} H^*) \delta^* = (H^*, \psi^{-1} Z^*) \] (3.5.12)

and
\[ -q'(0) = 2 Z^* \psi^{-1} H^* \delta^* = 2 \delta^* (H^*, \psi^{-1} H^*) \delta^* > 0 . \]

The last inequality holds since \((H^*, \psi^{-1} H^*)\) is of full rank by Assumption 4.

\[ \therefore q'(0) < 0 . \]

Choose an \( \varepsilon > 0 \) so that \( \varepsilon < -q'(0) \). By the definition of the derivative there is a \( \lambda^* \in (0, 1/2 \eta) \) such that

\[ \lambda (\alpha^* + \lambda^* \delta^*, \beta^* + \lambda^* \delta^*) - \lambda (\alpha^*, \beta^*) = q(\lambda^*) - q(0) < [q'(0) + \varepsilon] \lambda^* . \]

Since \( \lambda \) is continuous for \( \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in T \) we may choose \( \gamma > 0 \) such that
- γ > [q'(0) + ε] λ*

and there is a ρ > 0 such that

$$|\alpha_{i_n} + \dot{\alpha}^* \delta_{\alpha_{i_n}} - \alpha_j^* - \lambda \delta_{\alpha_j^*}| < \rho \quad \forall j = 1, 2, \ldots, r;$$

and

$$|\beta_{i_n} + \dot{\beta}^* \delta_{\beta_{i_n}} - \beta_j^* - \lambda \delta_{\beta_j^*}| < \rho \quad \forall j = 1, 2, \ldots, r;$$

therefore,

$$\ell(\alpha_{i_n} + \lambda^* \delta_{\alpha_{i_n}}, \beta_{i_n} + \lambda^* \delta_{\beta_{i_n}}) - \ell(\alpha^* + \dot{\alpha}^* \delta_{\alpha^*}, \beta^* + \dot{\beta}^* \delta_{\beta^*}) < \gamma.$$

Then for all i_n sufficiently large we have

$$\ell(\alpha_{i_n} + \lambda^* \delta_{\alpha_{i_n}}, \beta_{i_n} + \lambda^* \delta_{\beta_{i_n}}) - \ell(\alpha^*, \beta^*)$$

$$< [q'(0) + \varepsilon] \lambda^* + \gamma = -c^2 \quad \text{say}.$$  

Now for i_n large enough
is interior to $T$ so that $\lambda^* \in [0,1]$ and hence

$$L(\alpha_{i_n}, \beta_{i_n}) - L(\alpha^*, \beta^*) < -c^2.$$ 

This contradicts the fact that

$$L(\alpha_{i_n}, \beta_{i_n}) \rightarrow L(\alpha^*, \beta^*)$$

as $i_n \rightarrow \infty$.

Thus

$$\begin{bmatrix}
\delta^* \\
\delta^* \\
\delta^* \\
\delta^* \\
\end{bmatrix} = 0.$$
Therefore it follows from (3.5.12) that since \( \delta^* = 0 \) then

\[
H' \psi^* Z^* = 0
\]

and hence that

\[
\frac{\partial I(\alpha^*, \beta^*)}{\partial \alpha_s} = 0 \quad s = 1, 2, \ldots, r
\]

and

\[
\frac{\partial I(\alpha^*, \beta^*)}{\partial \beta_u} = 0 \quad u = 1, 2, \ldots, p.
\]

Thus given any subsequence of \( \left\{ \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \right\}_{i=1}^\infty \) in \( T \) by the above

there is a convergent subsequence with limit point \( \begin{bmatrix} \gamma^*_n \\ \beta_n \end{bmatrix} \) in \( T \) such that
\[
\frac{\partial \ell(\alpha'', \beta'\prime)}{\partial \alpha_s} = 0 = \frac{\partial \ell(\alpha^*, \beta^*)}{\partial \alpha_s}
\]

\(s = 1, 2, \ldots, r;\)

\[
\frac{\partial \ell(\alpha'', \beta'\prime)}{\partial \beta_u} = 0 = \frac{\partial \ell(\alpha^*, \beta^*)}{\partial \beta_u}
\]

\(u = 1, 2, \ldots, p;\)

and

\[\ell(\alpha'', \beta'\prime) = \ell^* = \ell(\alpha^*, \beta^*).\]

Therefore by Assumption 3 \[
\begin{bmatrix}
\alpha''

\beta''
\end{bmatrix}
= \begin{bmatrix}
\alpha^*

\beta^*
\end{bmatrix}.\] This completes the proof. \[]

We have shown above that if \(\ell(\alpha, \beta)\) has a unique minimum interior to a convex bounded subset \(T\) of \(\Theta\) and if we have a starting value of

\[
\begin{bmatrix}
\alpha \\
\beta
\end{bmatrix}, \begin{bmatrix}
\alpha_0 \\
\beta_0
\end{bmatrix}, \text{ in } T \text{ such that } \ell(\alpha_0', \beta_0') \text{ is smaller than on any boundary point of } T \text{ then the Modified Gauss Newton Procedure can be used to obtain a sequence of estimators }
\]

\[
\begin{bmatrix}
\alpha_i \\
\beta_i
\end{bmatrix}_{i=1}^n
\]

that converge to the minimum of \(\ell(\alpha, \beta)\) in \(T\).
We have shown that under Assumptions 1-8 given in IIIA the following statements are true.

1. A maximum likelihood estimator of \( \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \) exists and is consistent.

2. If the maximum likelihood estimator is an interior point of \( \Theta \) then a root \( \begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \) of the likelihood equations exists corresponding to the maximum of the likelihood function. If the maximum of the likelihood equations is not an interior point of \( \Theta \) then there may not be a root of the likelihood equation corresponding to the maximum likelihood estimator.

3. The likelihood equations have a root \( \begin{bmatrix} \alpha^w \\ \beta^w \end{bmatrix} \) with probability 1 as \( n \to \infty \) which is consistent. There is only one root that is consistent. The probability that the likelihood function is a maximum at \( \begin{bmatrix} \alpha^w \\ \beta^w \end{bmatrix} \) converges to 1 as \( n \to \infty \).

4. The asymptotic distribution of \( \sqrt{n} \begin{bmatrix} \alpha - \alpha^w \\ \beta - \beta^w \end{bmatrix} \) is normal with mean 0 and covariance matrix \( \lim_{n \to \infty} \left( \frac{H \psi^{-1} H}{n} \right)^{-1} \) where \( H \) is a function of the data.

5. There exists a simple least squares estimator \( \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \) such that
\[
\begin{bmatrix} \alpha^0 - \alpha^w \\ \beta^0 - \beta^w \end{bmatrix} = O_p \left( \frac{1}{n} \right).
\]
\( \begin{bmatrix} \alpha^0 \\ \beta^0 \end{bmatrix} \) is asymptotically normal but less...
efficient than the maximum likelihood estimator.

6. A joint least squares estimator given by

\[
\begin{bmatrix}
\hat{\alpha}_1 \\
\hat{\beta}_1
\end{bmatrix} = \delta^*_1 \begin{bmatrix}
\hat{\alpha}_0 \\
\hat{\beta}_0
\end{bmatrix}
\]

exists where

\[
\delta^*_1 = (H_0' \psi_0^{-1} H_0)^{-1} (H_0' \psi_0^{-1} Z_0).
\]

The asymptotic distribution of

\[
\sqrt{n} \begin{bmatrix}
\hat{\alpha}_1 - \alpha^0 \\
\hat{\beta}_1 - \beta^0
\end{bmatrix}
\]

is the same as the asymptotic distribution of

\[
\sqrt{n} \begin{bmatrix}
\alpha - \alpha^0 \\
\beta - \beta^0
\end{bmatrix}
\]

where \( \begin{bmatrix}
\alpha^2 \\
\beta^2
\end{bmatrix} \) is the consistent root of the likelihood equations.

7. An iterative procedure exists which uses the joint least squares procedure to obtain a sequence of estimators which converges to a local maximum of the likelihood function in the interior of \( \Theta \).
IV. APPLICATION OF THE DERIVED PROPERTIES
OF THE ESTIMATORS TO SOME EXAMPLES

In this section we apply the conclusions of Section III to three special cases of the model discussed in Section II. We assume throughout that Assumptions 1 through 8 are satisfied when required. We discuss a nonparametric unequal variance model, a random coefficient model and a generalization of the Prais and Houthakker (1955) model of consumer behavior.

A nonparametric unequal variance model is given by

\[ Y = X\beta + e, \]

\[ Y \text{ nmxl}; \quad X \text{ nmxl}; \quad \beta \text{ pxl}; \quad e \text{ nmxl}; \]

where

\[ \mathbb{E}[e] = 0, \quad \mathbb{E}[ee'] = G; \]

\[ G = \begin{bmatrix} \sigma^2_{11} I_n & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{nn} I_n \end{bmatrix}_{nm \times nm}. \]

Let

\[ P = \begin{bmatrix} 1_{n} \\ \vdots \\ 1_{n} \end{bmatrix}_{nm \times m}; \quad 1_{n} = \begin{bmatrix} 1 \\ \vdots \end{bmatrix}_{n \times 1}. \]
\[ H = \begin{bmatrix} 0 & X \\ P & 0 \end{bmatrix}; \quad \psi = \begin{bmatrix} G & 0 \\ 0 & 2\hat{g} \end{bmatrix} \]

\[ g' = [\sigma^2_1, \ldots, \sigma^2_i, \ldots, \sigma^2_2, \ldots, \sigma^2_m, \ldots, \sigma^2_m]^{\top} \times m \]

\[ \alpha' = [\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_m]^{\top} \times m. \]

The simple least squares estimator of \( \sigma^2_j, j = 1, 2, \ldots, m \) is

\[ \frac{1}{n} \sum_{i=1}^{n} e_{ij}^2 / n. \]

We note that this estimator is equivalent to (2.2.5). The asymptotic covariance matrix is given by the diagonal matrix of elements

\[ \lim_{n \to \infty} \frac{2\sigma^4_j}{n} \quad j = 1, 2, \ldots, m. \]

Clearly this matrix is equivalent to the asymptotic covariance matrix of the estimators given by (2.1.6), (2.1.7) and (2.1.10). The joint least squares estimator of \( \alpha, \alpha_0 \), is equivalent to \( \hat{\alpha}_0 \).

The random coefficient model of Hildreth and Houck (1968) was given in (2.1.16) of Section II. In matrix notation we have

\[ Y = X\beta + e \]

\[ Y_{nx1}; \quad X_{nxp}; \quad \beta_{px1}; \quad e_{nx1}; \]

\[ E[e] = 0; \quad E[ee'] = G; \]

where \( G \) is a diagonal matrix with elements
The single least squares estimator \( \hat{\alpha}_0 \) and its asymptotic covariance matrix are given by

\[
\hat{\alpha}_0 = (\hat{\mathbf{x}}'\hat{\mathbf{x}})^{-1}(\hat{\mathbf{x}}'\hat{\mathbf{e}}),
\]

and

\[
\lim_{n \to \infty} \left( \frac{\hat{\mathbf{x}}'\hat{\mathbf{x}}}{n} \right)^{-1} \left( \frac{\hat{\mathbf{x}}'\hat{\mathbf{e}}}{n} \right) = \cdots.
\]

The estimators given by (2.1.21) and (2.1.24) are asymptotically equivalent to \( \hat{\alpha}_0 \). The asymptotic covariance matrix of the joint least squares estimator

\[
\begin{bmatrix}
\hat{\alpha}_1 \\
\hat{\beta}_1
\end{bmatrix}
\]

is given by

\[
\lim_{n \to \infty} \left( \frac{\hat{\mathbf{x}}' [\hat{\mathbf{G}}]^{-1} \hat{\mathbf{x}}}{n} \right)^{-1} = \cdots.
\]

which is also the asymptotic distribution of the estimators given by (2.1.22) and (2.1.23).
In Section II we discussed the linear model used in a study of consumer behavior by Prais and Houthakker (1955). In this study it was found that \( \frac{V(Y|X)}{[E(Y|X)]^2} = k \), where \( Y = X \beta + e \) and \( E[Y|X] = X \beta \).

A more general model is obtained by adding the component of error due to measurement error in \( Y \). We then obtain the covariance matrix \( E[ee'] = G \) where

\[
G = 
\begin{bmatrix}
\alpha_1 + \alpha_2 [X_i^\beta]^2 & 0 \\
0 & \alpha_1 + \alpha_2 [X_n^\beta]^2
\end{bmatrix}
\]

The term \( \alpha_1 \) is the component due to measurement error in \( Y \) while \( \alpha_2 [X_i^\beta]^2 \) is the component due to the variance of \( Y \) given \( X_i \) for \( i = 1, 2, \ldots, n \).

We let

\[
H = \begin{bmatrix} 0 & X \\ H_1 & H_2 \end{bmatrix} \quad ; \quad \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \quad ; \quad F = [H_1 H_2]_{nx(p+2)} \quad ;
\]

\[
H_1 = \begin{bmatrix} 1 & (X_1^\beta)\\ 
\vdots & \vdots \\
\vdots & \vdots \\
1 & (X_n^\beta) \end{bmatrix} \quad ; \quad H_2 = \begin{bmatrix}
2\alpha_2 x_{11}(X_1^\beta)^2 & \ldots & 2\alpha_2 x_{1p}(X_1^\beta)^3 \\
\vdots & \vdots & \vdots \\
2\alpha_2 x_{n1}(X_n^\beta)^2 & \ldots & 2\alpha_2 x_{np}(X_n^\beta)^2
\end{bmatrix}.
\]
A preliminary estimator of $\alpha$ may be obtained by replacing $\beta$ by $\hat{\beta}_0$ in $H_1$ to obtain $\hat{H}_{10}$ and then computing the simple least squares estimator

$$\hat{\alpha}_0 = (\hat{H}_{10}' \hat{H}_{10})^{-1} \hat{H}_{10}' \hat{e}.$$ 

The asymptotic covariance matrix of $\hat{\alpha}_0$ is given by

$$\lim_{n \to \infty} \left[ \left( \frac{H_1' H_1}{n} \right)^{-1} \left( \frac{H_1' 2\hat{G} H_1}{n} \right) \left( \frac{H_1' H_1}{n} \right)^{-1} \right].$$

The joint least squares estimator of $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ is given by (3.4.5) and its asymptotic covariance matrix by Theorem 3.4.4.
V. A COMPARISON OF THE PERFORMANCES OF THE SIMPLE AND JOINT LEAST SQUARES ESTIMATORS FOR A RANDOM COEFFICIENT MODEL

In this section we use Monte Carlo procedures to compare the performances of the simple and joint least squares estimators for an example of a random coefficient model.

A. The Procedure

The random coefficient model as given in Section II and IV is summarized as follows:

\[ Y = X\beta + e \]

\( Y_{n \times 1} \); \( X_{n \times p} \); \( \beta_{p \times 1} \); \( e_{n \times 1} \); \( \alpha_{p \times 1} \); \hspace{1cm} (5.1.1)

\[ \mathbb{E}[e] = 0 \; ; \quad \mathbb{E}[ee'] = G \; ; \]

where \( G \) is a diagonal covariance matrix with elements

\[ g_{tt} = \sum_{j=1}^{p} x_{tj}^2 \alpha_j \]

\( t = 1, 2, \ldots, n \).

For the 40x3 matrix \( X \) given in Table 5.1 and the parameter vector
a set of 50, 40x1 vectors \( \mathbf{Y} \) were generated. First a set of 50, 40x1 vectors \( \mathbf{e} \) were generated and then \( \mathbf{Y} \) computed according to \( \mathbf{Y} = \mathbf{X}\beta^0 + \mathbf{e} \). Each \( \mathbf{e} \) vector was multivariate normal with mean 0 and diagonal covariance matrix \( \mathbf{G} \). The generation of \( \mathbf{e} \) began with the generation of uniform random numbers using a technique given by Marsaglia and Bray (1968). Standard normal random numbers were obtained from the uniform random numbers using the Box-Mueller (1958) transformation. The \( \mathbf{e} \) vector was then obtained by multiplying the elements of the 40x1 standard normal vector by \( \sqrt{\mathbf{g}_t} \), \( t = 1, 2, \ldots, 40 \), where \( \mathbf{g}_t \) is given by (5.1.1).

The simple least squares estimators of \( \beta \) and \( \alpha \) were obtained using

\[
\hat{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}
\]

and

\[
\hat{\alpha}_0 = [(\tilde{\mathbf{M}}\tilde{\mathbf{X}})'(\tilde{\mathbf{M}}\tilde{\mathbf{X}})]^{-1} (\tilde{\mathbf{M}}\tilde{\mathbf{X}})' \hat{\mathbf{e}}
\]

The asymptotic covariance matrices are given by

\[
\text{Var}(\hat{\beta}_0) = (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{G}\mathbf{X}) (\mathbf{X}'\mathbf{X})^{-1} \quad (5.1.1)
\]

and

\[
\text{Var}(\hat{\alpha}_0) = [(\tilde{\mathbf{M}}\tilde{\mathbf{X}})'(\tilde{\mathbf{M}}\tilde{\mathbf{X}})]^{-1} [(\tilde{\mathbf{M}}\tilde{\mathbf{X}})'[2\mathbf{G}](\tilde{\mathbf{M}}\tilde{\mathbf{X}})] [(\tilde{\mathbf{M}}\tilde{\mathbf{X}})'(\tilde{\mathbf{M}}\tilde{\mathbf{X}})]^{-1} = \mathbf{C}
\]
Estimates of the asymptotic covariance matrices are obtained by replacing \( \alpha \) by \( \hat{\alpha}_0 \) in \( G \) and \( \hat{G} \) above.

The joint least squares estimators of \( \beta^0 \) and \( \alpha^0 \) require preliminary estimates of the elements of \( G \) and \( \hat{G} \). The diagonal elements of \( G, g_t, t = 1, 2, \ldots, 40 \) were estimated using

\[
\hat{g}_t = g_t, \quad g_t > \varepsilon_t \quad t = 1, 2, \ldots, 40.
\]

\[
= \varepsilon_t, \quad g_t \leq \varepsilon_t
\]

where \( \hat{g}_t = \hat{x}_t' \hat{\alpha}_0 \), \( \varepsilon_t = \sqrt{\bar{V}_t} \), \( \bar{V}_t \) is the asymptotic variance of \( \hat{g}_t \), and \( \bar{V}_t \) is the estimate of \( V_t \) obtained by using \( \hat{\alpha}_0 \) in place of \( \alpha \). \( V_t \) is given by

\[
\bar{V}_t = \hat{x}_t' \bar{\Sigma} \hat{x}_t
\]

where \( \bar{\Sigma} \) is the estimator of \( \Sigma \) and \( \Sigma \) is defined in (5.1.1).

The joint least squares estimators are given by

\[
\hat{\beta} = (X' \bar{G}^{-1} X)^{-1} (X' \bar{G}^{-1} Y)
\]

\[
\hat{\alpha} = [(\hat{M} \hat{\alpha})' [\hat{2} \hat{G}]^{-1} (\hat{M} \hat{\alpha})^{-1} [(\hat{M} \hat{\alpha})' [\hat{2} \hat{G}]^{-1} Y]
\]

where \( \bar{G} \) and \( \hat{G} \) denote the diagonal matrices of elements \( \tilde{g}_t \) and \( \tilde{g}_t^2 \) respectively, \( t = 1, 2, \ldots, 40 \). Estimates of the asymptotic covariance matrices were obtained from
\[ \hat{\text{Var}}(\hat{\beta}) = (X'\hat{\Gamma}^{-1}X)^{-1} \]
\[ \hat{\text{Var}}(\hat{\alpha}) = [(\hat{\Phi})' \hat{2\hat{\Phi}}^{-1} (\hat{\Phi})]^{-1} . \] (5.1.3)

In addition to estimates \( \hat{\omega} \) of \( \omega^0 \), the following sample quantities were obtained for the two estimators for each 50 samples.

\[ t = (\hat{\kappa}_{i}) = \frac{\hat{\omega}_{i} - \omega_{i}^{0}}{\sqrt{\hat{\text{Var}}(\hat{\omega}_{i})}} \quad i = 1, 2, \ldots, 2p ; \]
\[ z = (\hat{\zeta}_{i}) = \frac{\bar{\theta}_{i} - \omega_{i}^{0}}{\sqrt{\text{Var}(\bar{\theta}_{i})}} \quad i = 1, 2, \ldots, 2p . \]

The Monte Carlo mean and variance of each of the sample statistics was computed for each estimator over the 50 samples.

B. Summary of Results

The performances of the two estimators are compared in Table 5.2. For both estimators of \( \beta \) the estimated bias is small. This is consistent with the theory since both estimators are unbiased. In all cases the asymptotic variances and Monte Carlo variances of the joint least squares estimators were smaller than the respective variances for the simple least squares estimators. Also, on the basis of the first two moments, the t
and \( z \) statistics for \( \hat{\beta} \) are well approximated by their asymptotic distributions.

In Table 2 we note that the Monte Carlo variances of the estimators of \( \beta^0 \) were, in general, less than the asymptotic variances for both estimators. Also, the second moments of \( t \) and \( z \) were in general less than 1. A partial explanation of these results was obtained by examination of the performance of the random number generator. From statistical theory we know that if \( x \) has a standard normal distribution then

\[
E[x] = 0; \quad E[x^2 - (E[x])^2] = 1.0; \quad E[x^4 - (E[x^2])^2] = 2.0.
\]

Monte Carlo estimates of each of the above quantities were obtained from the 50 samples for each of the 40 elements of the standard normal random vector. The means for each of these quantities over the 40 elements were 0.006, 0.985 and 1.849 respectively. We note that the mean values of the Monte Carlo estimates of \( E[x^2 - E[x]^2] \) and \( E[x^4 - (E[x^2])^2] \) were slightly lower than their theoretical values. We would therefore expect the Monte Carlo variances of the statistics to be slightly less than the corresponding theoretical variances.

In the case of the estimation of \( \alpha^0 \) we note that the estimated bias for the simple least squares estimator is relatively close to zero but the estimated bias of the joint least squares estimator is relatively large. These results are consistent with theory since the simple least squares estimator employing \( \hat{M} \) is unbiased but the joint least squares estimator is in general biased. The asymptotic variances and Monte Carlo variances for
the joint least squares estimator are similar. As observed in the estimation of $\beta^0$, however, the Monte Carlo variances of $\hat{\beta}$ tend to be less than the asymptotic variances.

In Table 5.2 we note the large difference between the Monte Carlo variances and asymptotic variances for the simple least squares estimator of $\alpha^0$. In fact, the Monte Carlo variances of the simple least squares estimator are quite similar to the Monte Carlo variances of the joint least squares estimator. This is largely explained by a comparison of the true variance of the simple least squares estimator with the asymptotic variance. The true covariance matrix is given by

$$[(\dot{M})'(\dot{M})]^{-1} (\dot{M})' T (\dot{M}) [(\dot{M})'(\dot{M})]^{-1}$$

where $T$ is the covariance matrix of

$$W = \hat{e} - E[\hat{e}] .$$

The asymptotic covariance is given by (5.11) as

$$[(\ddot{M})'(\ddot{M})]^{-1} (\ddot{M})' [\dot{G}] (\ddot{M}) [(\ddot{M})'(\ddot{M})]^{-1}$$

where $\dot{G}$ is the covariance of $\hat{e} - E[e]$. Since by Lemma 3.3.9 $T = 2\hat{G} + O \left( \frac{1}{n} \right)$ for large samples the two covariance matrices should be similar. In our case the true variances of $\alpha_1$, $\alpha_2$, $\alpha_3$ were 1.83, 0.012 and 0.015 respectively which are less than the corresponding asymptotic variances 2.867, 0.018 and 0.024 given in Table 5.2. Also, from this table the Monte Carlo variances of $\alpha_1$, $\alpha_2$ and $\alpha_3$ are 1.190, 0.011 and 0.011. As we observed above the Monte Carlo variances may also be less than the asymptotic variances because of the random numbers generated. It would
appear that for a sample of size $n = 10$ the distribution of the $t$ and $z$
statistics differ considerably from normal $(0, 1)$.

In Subsection A of this Chapter the procedure for establishing a
positive lower bound on variance estimates was outlined. The Monte Carlo
average frequency of the use of the lower bound for the $40 \times 1$ vector of
estimates $\{\hat{\sigma}_t\}_{t=1}^n$ was obtained for both estimators. In the case of
simple least squares an average of 5.5 of the 40 variance estimates fell
below the lower bound and for the joint least squares estimator this average
was 3.3. This result is consistent with theory since as the variance of
the estimator decreases the lower bound decreases and hence fewer estimates
should be below the lower bound.

We have seen that the asymptotic distributions of the simple least
squares and joint least squares estimators of $\beta$ furnished an adequate
approximation to the observed distributions. The simple least squares
estimator of $\alpha$ was as efficient as the joint least squares estimator
for our example. The joint least squares estimator of $\alpha$ displayed a large
bias while the simple least squares estimator was unbiased. Neither estima-
tor of $\alpha$ was well approximated by the asymptotic distribution.
Table 5.1. X MATRIX FOR RANDOM COEFFICIENT MODEL

<table>
<thead>
<tr>
<th>n</th>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
<th>n</th>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>9.040</td>
<td>-7.820</td>
<td>21</td>
<td>1.000</td>
<td>18.050</td>
<td>0.340</td>
</tr>
<tr>
<td>2</td>
<td>1.000</td>
<td>13.160</td>
<td>-29.350</td>
<td>22</td>
<td>1.000</td>
<td>21.420</td>
<td>19.170</td>
</tr>
<tr>
<td>3</td>
<td>1.000</td>
<td>12.500</td>
<td>-0.730</td>
<td>23</td>
<td>1.000</td>
<td>22.910</td>
<td>0.270</td>
</tr>
<tr>
<td>4</td>
<td>1.000</td>
<td>14.200</td>
<td>-13.290</td>
<td>24</td>
<td>1.000</td>
<td>22.530</td>
<td>-10.270</td>
</tr>
<tr>
<td>5</td>
<td>1.000</td>
<td>16.930</td>
<td>11.610</td>
<td>25</td>
<td>1.000</td>
<td>10.270</td>
<td>12.260</td>
</tr>
<tr>
<td>6</td>
<td>1.000</td>
<td>19.270</td>
<td>13.510</td>
<td>26</td>
<td>1.000</td>
<td>14.410</td>
<td>8.070</td>
</tr>
<tr>
<td>7</td>
<td>1.000</td>
<td>9.560</td>
<td>-12.480</td>
<td>27</td>
<td>1.000</td>
<td>20.310</td>
<td>7.190</td>
</tr>
<tr>
<td>8</td>
<td>1.000</td>
<td>14.550</td>
<td>-27.310</td>
<td>28</td>
<td>1.000</td>
<td>22.050</td>
<td>15.610</td>
</tr>
<tr>
<td>9</td>
<td>1.000</td>
<td>15.640</td>
<td>-5.680</td>
<td>29</td>
<td>1.000</td>
<td>25.960</td>
<td>8.690</td>
</tr>
<tr>
<td>10</td>
<td>1.000</td>
<td>16.740</td>
<td>14.340</td>
<td>30</td>
<td>1.000</td>
<td>26.850</td>
<td>-1.910</td>
</tr>
<tr>
<td>11</td>
<td>1.000</td>
<td>17.330</td>
<td>-14.060</td>
<td>31</td>
<td>1.000</td>
<td>14.840</td>
<td>10.540</td>
</tr>
<tr>
<td>12</td>
<td>1.000</td>
<td>17.800</td>
<td>-14.190</td>
<td>32</td>
<td>1.000</td>
<td>19.310</td>
<td>5.610</td>
</tr>
<tr>
<td>13</td>
<td>1.000</td>
<td>11.040</td>
<td>-0.910</td>
<td>33</td>
<td>1.000</td>
<td>23.720</td>
<td>-3.150</td>
</tr>
<tr>
<td>14</td>
<td>1.000</td>
<td>15.480</td>
<td>-4.780</td>
<td>34</td>
<td>1.000</td>
<td>22.700</td>
<td>9.860</td>
</tr>
<tr>
<td>15</td>
<td>1.000</td>
<td>16.940</td>
<td>-6.040</td>
<td>35</td>
<td>1.000</td>
<td>32.090</td>
<td>34.710</td>
</tr>
<tr>
<td>16</td>
<td>1.000</td>
<td>19.380</td>
<td>7.580</td>
<td>36</td>
<td>1.000</td>
<td>17.050</td>
<td>19.780</td>
</tr>
<tr>
<td>17</td>
<td>1.000</td>
<td>27.540</td>
<td>19.820</td>
<td>37</td>
<td>1.000</td>
<td>20.250</td>
<td>1.780</td>
</tr>
<tr>
<td>18</td>
<td>1.000</td>
<td>23.030</td>
<td>17.350</td>
<td>38</td>
<td>1.000</td>
<td>20.160</td>
<td>12.180</td>
</tr>
<tr>
<td>19</td>
<td>1.000</td>
<td>23.700</td>
<td>-0.440</td>
<td>39</td>
<td>1.000</td>
<td>28.840</td>
<td>16.360</td>
</tr>
<tr>
<td>20</td>
<td>1.000</td>
<td>15.330</td>
<td>-9.440</td>
<td>40</td>
<td>1.000</td>
<td>36.190</td>
<td>-3.770</td>
</tr>
</tbody>
</table>
### Table 5.2. A Comparison of the Performances of the Estimators

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True value</strong></td>
<td>10.00</td>
<td>-4.00</td>
<td>3.00</td>
<td>30.00</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td><strong>Estimated bias</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLS a</td>
<td>-0.312</td>
<td>0.023</td>
<td>-0.024</td>
<td>0.834</td>
<td>0.001</td>
<td>-0.012</td>
</tr>
<tr>
<td>JLS b</td>
<td>0.204</td>
<td>-0.010</td>
<td>-0.008</td>
<td>5.08</td>
<td>-0.013</td>
<td>-0.014</td>
</tr>
<tr>
<td><strong>Asymptotic variance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>28.6</td>
<td>0.087</td>
<td>0.017</td>
<td>2867</td>
<td>0.018</td>
<td>0.024</td>
</tr>
<tr>
<td>JLS</td>
<td>23.4</td>
<td>0.072</td>
<td>0.012</td>
<td>1306</td>
<td>0.011</td>
<td>0.014</td>
</tr>
<tr>
<td><strong>Monte Carlo variance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>25.3</td>
<td>0.083</td>
<td>0.014</td>
<td>1190</td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td>JLS</td>
<td>20.0</td>
<td>0.064</td>
<td>0.012</td>
<td>1220</td>
<td>0.011</td>
<td>0.009</td>
</tr>
<tr>
<td><strong>$T$ average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>-0.057</td>
<td>0.082</td>
<td>-0.227</td>
<td>0.139</td>
<td>-0.238</td>
<td>-0.373</td>
</tr>
<tr>
<td>JLS</td>
<td>0.053</td>
<td>-0.050</td>
<td>-0.222</td>
<td>0.104</td>
<td>-0.316</td>
<td>-0.270</td>
</tr>
<tr>
<td><strong>$T$ variance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>0.921</td>
<td>1.044</td>
<td>0.983</td>
<td>0.384</td>
<td>0.692</td>
<td>0.741</td>
</tr>
<tr>
<td>JLS</td>
<td>0.827</td>
<td>0.875</td>
<td>0.893</td>
<td>0.820</td>
<td>1.233</td>
<td>0.799</td>
</tr>
<tr>
<td><strong>$Z$ average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>-0.058</td>
<td>0.079</td>
<td>-0.186</td>
<td>0.016</td>
<td>0.009</td>
<td>-0.078</td>
</tr>
<tr>
<td>JLS</td>
<td>0.042</td>
<td>-0.036</td>
<td>-0.193</td>
<td>0.141</td>
<td>-0.121</td>
<td>-0.115</td>
</tr>
<tr>
<td><strong>$Z$ variance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLS</td>
<td>0.882</td>
<td>0.949</td>
<td>0.852</td>
<td>0.414</td>
<td>0.595</td>
<td>0.438</td>
</tr>
<tr>
<td>JLS</td>
<td>0.853</td>
<td>0.885</td>
<td>0.830</td>
<td>0.934</td>
<td>1.012</td>
<td>0.638</td>
</tr>
</tbody>
</table>

*a* SLS stands for simple least squares.

*b* JLS stands for joint least squares.
VI. SUMMARY

The problem of estimation in the linear model $Y = X\beta + e$ with unknown diagonal covariance matrix $G$ was considered. It was assumed that the elements of $G$ are known functions of the explanatory variables $X$ and an unknown parameter $\Theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ where the elements of $\beta$ are a subset of the elements of $\beta$.

The least squares procedure and the maximum likelihood procedure are commonly used to estimate $\Theta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$. The maximum likelihood procedure is employed under the assumption that the vector $e$ is normally distributed. The least squares procedure uses residuals from a simple least squares fit to obtain an estimate $\hat{\Theta}$ of $\Theta$. An estimated covariance matrix $\hat{G}$ and the weighted least squares estimator $\bar{\beta}$ are then constructed where

$$\bar{\beta} = (X'G^{-1}X)^{-1}(X'G^{-1}Y).$$

It was demonstrated that the maximum likelihood estimator of $\Theta$ exists and is weakly consistent. If the maximum likelihood estimator is an interior point of the parameter space $\Theta$ then the likelihood equations have a unique consistent root at this point. As $n \to \infty$ the likelihood equations have a unique consistent root $\tilde{\Theta}$ and the probability that the likelihood function is a maximum at $\tilde{\Theta}$ converges to 1 as $n \to \infty$. The estimator $\tilde{\Theta}$ is asymptotically normal.

A joint least squares procedure was developed which uses a preliminary estimator $\hat{\Theta}_0$ assumed to satisfy $\hat{\Theta}_0 - \Theta = O_p \left( \frac{1}{\sqrt{n}} \right)$. The joint least
squares procedure obtains estimators that are in general asymptotically more efficient than simple least squares estimators. The joint least squares estimator of $\Theta, \hat{\Theta}_1$, is consistent and has the same asymptotic distribution as the consistent root, $\tilde{\Theta}$, of the likelihood equations. If $\beta$ is not an unknown parameter of $G$ then the joint least squares estimator of $\beta$ is equivalent to $\tilde{\beta}$. If the covariance matrix is a function of $\beta$ it was shown that the joint least squares estimator of $\beta, \hat{\beta}_1$, is at least as efficient as the estimator $\tilde{\beta}$.

The joint least squares procedure requires a preliminary estimator $\hat{\Theta}_0$ of $\Theta$ such that $\hat{\Theta}_0 - \Theta = O_p\left(\frac{1}{\sqrt{n}}\right)$. It has been shown that a simple least squares estimator exists and has this property.

In general an analytical expression for the maximum likelihood estimator does not exist. We have shown that an iterative procedure which uses the joint least squares procedure can be used to obtain a sequence of estimators that converges to a local maximum of the likelihood function. This procedure is similar to the Modified Gauss Newton procedure for nonlinear least squares estimation.

A Monte Carlo procedure was used to study the performance of the simple and joint least squares estimators of $\Theta$ for an example of a random coefficient model. For a sample size $n = 40$ the behavior of the estimators of $\beta$ were well approximated by the asymptotic theory but the behavior of the estimators of $\alpha$ were not. For the example considered the simple least squares estimator of $\alpha$ was as efficient as the joint least squares estimator.
VII. LITERATURE CITED


VIII. ACKNOWLEDGEMENTS

The author wishes to express his appreciation to Professor Wayne Fuller for his patient guidance, encouragement and many helpful suggestions during the course of this investigation and throughout the graduate program. Also, the influence of Professors Barry Arnold, Chien-Pai Han, Dean Isaacson, and Glen Meeden through their enthusiastic teaching and consultation is gratefully acknowledged.

Special thanks is given to Professor George Ladd for serving on my committee, to Professor T. A. Bancroft for providing financial assistance and to Mrs. Norma Elwick for the accurate typing of this manuscript.

A special debt is owed my wife, Leone, for her constant enthusiasm and encouragement and for maintaining a happy and cheerful home despite the author's frequent absence.