SOME ASPECTS OF LINEAR REGRESSION SYSTEMS

by

William S. Mallios

Institute of Statistics
Mimeograph Series No. 298
November, 1961
ERRATA SHEET

Page 5
7.1 should read The Estimation of Stationary Points of Linear Functions of Linear Regressions
7.2 should read ... when η(\hat{\gamma}) is a ...

Page 4
4th line from bottom - Roy's 267 instead of Roy's 267

Page 11
2nd paragraph, line one, 1959 instead of 1958

Page 26
in (5.1),

\[ y_T = \delta_T \quad y_T + \delta_T \]
\[ y_L = \delta_L \quad y_L + \delta_L \]
\[ y_V = \delta_V \quad y_V + \delta_V \]

2nd paragraph, next to last line

\[ E(\varepsilon_T \varepsilon_L) = \sigma_{TL} \quad \text{instead of} \quad E(\varepsilon_G \varepsilon_L) = \sigma_{TL} \]

in (5.2), 2nd equation

\[ y_L = b_0 + b_1 y_T + b_2 x_A + b_3 x_D + b_4 x_A x_D + b_5 x_A^2 + b_6 x_D^2 + \varepsilon_L \quad \text{instead of} \]
\[ y_L = b_0 + b_1 y_T + b_2 x_A + b_3 x_D + b_4 x_A x_D + b_5 x_A + b_6 x_D + \varepsilon_L \]

Page 29
in (5.5)

\[ \tilde{X}' = (x_A, x_D, x_A^2, x_D^2) \quad \text{instead of} \]
\[ \tilde{X}' = (x_A, \ldots, x_A^2, x_D^2) \]

7th line from bottom

\[ y_T - y_L \quad \text{instead of} \quad x_T - y_L \]

Page 31-37, l.h.
Read causal (non-causal) instead of casual (non-causal)

Page 31
2nd paragraph, 4th line

For example, in the model

\[ \gamma = \sum_{i=0}^{p} b_i x_i + \varepsilon \quad (x_0 = 1) \]
First paragraph, last line

\[ \mathbb{E}(\varepsilon_1^2) = -b_2 \sigma_{22} \text{ for } b_2 \neq 0 \text{ and } \sigma_{12} = 0 \text{, instead of} \]

\[ \mathbb{E}(\varepsilon_1^2) = b_2 \hat{\sigma}_{22} \text{ for } b_2 \neq 0 \text{ and } \delta_{12} = 0 \]

* Figure 6.1

\[ \tilde{y}_2 = \tilde{b}_0 + \tilde{b}_1 x \text{ instead of} \]

\[ \tilde{y}_1 = \tilde{b}_0 + \tilde{b}_1 x \]

In the corollary, change "test" to "best".

3rd line from bottom, \( p_{jj} \) instead of \( p_{jj} \)

(6.60), (6.61), (6.62)

\[ (\hat{\phi}/\hat{b} \mid b = \hat{b}) \text{ instead of } (\hat{\phi}/\hat{b} \mid b = \hat{b}) \]

(6.70)

\[ y = A_p(x)/B_q(x) + \varepsilon \text{ instead of } y = A_p(x)/B_q(x) + \varepsilon \]

in theorem, 2nd line

\[ \psi_H(\theta) \text{ instead of } \psi_{\hat{H}}(\hat{\theta}) \]

2nd paragraph, 1st line

Hancock \[\hat{F}\] instead of Hancock \[F\]

3rd paragraph, 3rd line from bottom

method of steepest descent \[\hat{T}\] instead of

method of steepest descent \[T\]

2nd paragraph, 4th line

Frank and Wolf \[\hat{I}\] instead of Frank and Wolf \[I\]

2nd paragraph, 7th line

Kuhn-Tucker \[\hat{Z}\] instead of Kuhn-Tucker \[Z\]

2nd paragraph, last line

Lagrangian form \[\hat{Z}\] instead of Lagrangian form \[Z\]

* Same paragraph, 1st line from bottom

\[ \mathbb{E}(\varepsilon_1^2) = \sigma_{12} = 0 \text{ instead of } \mathbb{E}(\varepsilon_1^2) = \delta_{12} = 0 \]
Page 108 2nd line, top of page

Umland and Smith $\text{38}^7$ instead of Umland and Smith $\text{45}^7$

109 12th line

Bodmer $\text{34}^7$ instead of Bodmer $\text{54}^7$

119 2nd paragraph, 1st line

"in finding not" instead of "in finding out".

3rd line from bottom

$\text{a}$ instead of $\text{a}$.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td></td>
</tr>
<tr>
<td>1.0 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2.0 REVIEW OF LITERATURE</td>
<td>7</td>
</tr>
<tr>
<td>3.0 TOXOPLASMOSIS</td>
<td>12</td>
</tr>
<tr>
<td>3.1 A Discussion of Toxoplasmosis</td>
<td>12</td>
</tr>
<tr>
<td>3.2 A Serologic Test for Detecting the Presence or Absence of Toxoplastic Antibodies</td>
<td>12</td>
</tr>
<tr>
<td>4.0 STATISTICS AND THE DYE TEST</td>
<td>14</td>
</tr>
<tr>
<td>4.1 The Probit Transform and LD 50</td>
<td>14</td>
</tr>
<tr>
<td>4.2 Linear Weighted Regression in Estimating the Line of Descent</td>
<td>16</td>
</tr>
<tr>
<td>4.3 The Role of Compound Distributions in the Dye Test</td>
<td>20</td>
</tr>
<tr>
<td>5.0 THE CYCLE OF INVESTIGATION AS APPLIED TO ONE STAGE OF THE DYE TEST</td>
<td>24</td>
</tr>
<tr>
<td>6.0 SYSTEMS OF LINEAR REGRESSION EQUATIONS</td>
<td>31</td>
</tr>
<tr>
<td>6.1 The Nature of Errors Associated with Linear Regression Systems</td>
<td>31</td>
</tr>
<tr>
<td>6.2 Estimation of the Parametric Coefficients in a Linear Regression System When W Is Known</td>
<td>36</td>
</tr>
<tr>
<td>6.3 Estimation of $\sigma^2$ and $\Sigma$</td>
<td>53</td>
</tr>
<tr>
<td>6.4 Estimation of $b$ When W Is Unknown</td>
<td>55</td>
</tr>
<tr>
<td>6.5 A Sufficient Condition for Convergence and an Approximate Variance of the Iterated Estimate</td>
<td>61</td>
</tr>
<tr>
<td>6.6 General Comments on Estimation in the Predictive Regression System</td>
<td>66</td>
</tr>
<tr>
<td>6.7 A Note on the Variance of a Vector Coefficient Estimate Obtained Through Iteration in a Single Rational Regression Model</td>
<td>73</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS (continued)

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.0</td>
<td>SOME UNCONSTRAINED EXTREME VALUE PROBLEMS</td>
<td>78</td>
</tr>
<tr>
<td>7.1</td>
<td>The Estimation of Stationary Points of Linear Regressions</td>
<td>78</td>
</tr>
<tr>
<td>7.2</td>
<td>A Confidence Region for $E(\hat{Z})$ when $\phi(\hat{y})$ is a Second-order Linear Function of the Z's</td>
<td>79</td>
</tr>
<tr>
<td>7.3</td>
<td>Remarks Concerning the Confidence Region</td>
<td>83</td>
</tr>
<tr>
<td>7.4</td>
<td>Intersection Region Confidence Procedures in Finding an Approximate Confidence Region for $E(\hat{Z})$</td>
<td>90</td>
</tr>
<tr>
<td>8.0</td>
<td>SOME PROBLEMS OF CONSTRAINED EXTREMA</td>
<td>99</td>
</tr>
<tr>
<td>8.1</td>
<td>Uses of Lagrangian Multipliers in a System of Linear Regression Equations</td>
<td>99</td>
</tr>
<tr>
<td>8.2</td>
<td>Errors in the Optimal Solution</td>
<td>108</td>
</tr>
<tr>
<td>8.3</td>
<td>An Approximate Confidence Interval for an Individual Lagrangian Multiplier</td>
<td>112</td>
</tr>
<tr>
<td>9.0</td>
<td>SUMMARY AND GENERAL REMARKS</td>
<td>117</td>
</tr>
<tr>
<td>10.0</td>
<td>LIST OF REFERENCES</td>
<td>120</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Percentage modified plotted on the arithmetic scale of twofold dilutions from 1:256 through 1:4096</td>
<td>17</td>
</tr>
<tr>
<td>4.2</td>
<td>Percentage modified plotted on the probit scale at twofold dilutions from 1:256 through 1:4096</td>
<td>18</td>
</tr>
<tr>
<td>5.1</td>
<td>A possible series of events characterizing the phenomenon of interest</td>
<td>25</td>
</tr>
<tr>
<td>6.1</td>
<td>Unbiased and best unbiased prediction lines</td>
<td>45</td>
</tr>
<tr>
<td>7.1</td>
<td>The intersection confidence region, $R^{(I)}$ for $E[(\hat{z}_1, \hat{z}_2)]$ for specific $k_1, k_2$ vectors</td>
<td>95</td>
</tr>
<tr>
<td>7.2</td>
<td>The $R^{(1)}$ approximation to $R^{(I)}$</td>
<td>95</td>
</tr>
<tr>
<td>7.3</td>
<td>The $R^{(2)}$ approximation to $R^{(I)}$</td>
<td>96</td>
</tr>
<tr>
<td>7.4</td>
<td>The $R^{(3)}$ approximation to $R^{(I)}$</td>
<td>97</td>
</tr>
<tr>
<td>8.1</td>
<td>A geometrical example of the compromise solutions and the optimal solution</td>
<td>102</td>
</tr>
</tbody>
</table>
1.0 INTRODUCTION

Suppose an inquiry were made concerning a certain phenomenon. The deeper inquiry, as opposed to phenomenology, may involve questions such as, "What are the mechanics of the phenomenon?", and "What are the laws of nature governing such an event?"

Though the second question is nebulous and/or perhaps beyond the generation of the scientist questioned, the first may fall within the framework of casual theory. As such, one may first sketch the network of hypothesized occurrences characterizing the phenomenon in a path diagram. From a statistical viewpoint, such diagrams may be described mathematically in terms of linear and/or nonlinear structural regression models; i.e., a system of models deduced possibly from one or more differential equations. Thereafter comes experimentation designed to support or discount the hypothesized models. Then the process is repeated with inferences gained in the first cycle dictating modifications in the hypothesized models used in the second cycle of investigation. The Humean reader may now expect the statement, "then at the termination of some future cycle, the true law of nature is uncovered." But does there exist a true law of nature; e.g., is the functional relationship $f(x)$, $20^1$ exact? More precisely, does the error entering in the functional relationship stem from extraneous effects on the true values of

---

$^1$ Numerals appearing in slanted brackets refer to citations in the List of References.
associated variables, or is there perhaps an error associated with the relationship itself? No attempt is made to answer this due to the uncertainty of the writer. Rather, we will say that the cycle of investigation is continued until a structural regression system is determined which describes adequately the mechanics of the phenomenon and hence, predicts adequately the event.

Almost any causal theory comes sooner or later to deal with structural regression rather than predictive regression. (See Chapter 5 for definitions of structural and predictive regressions.) Though the preceding quote was, most likely, not intended to slight predictive regression, one might wonder whether there has been a tendency to neglect predictive regression in favor of structural regression. The former may yield solutions to certain problems unsolvable by structural regression systems (hypothesized in early cycles and therefore possibly inadequate) and also offer aid in the cycles of investigation, particularly when structural regression coefficients are nonestimable. Further, interest does not always lie in the structural regression, but, even so, the latter may not be totally neglected, since the predictive regression models usually stem from the structural models.

Consider now the experimental stage of the cycle. In general, let us say the phenomenon of interest occurs "within" the experimental unit. Whether it be a human body (a group of individuals), a corn plant (an acre of corn plants), or a missile warhead, the unit can be characterized by
infinitely many variable (from one unit to the next) endogenous responses (or response types), only some of which are of interest. Let these be measurable and, of course, descriptive of the phenomenon of interest. Further, to alleviate the possible time difficulty, we will, throughout this writing, be confined to responses which can be regarded as occurring simultaneously or approximately so; e.g., sufficiently small time differentials between occurring responses may offer possible justification to an assumption of simultaneity.

By either theoretical or primitive means, the structural regression system is established, whereupon each endogenous response type is described in terms of a set of variables, part of which are exogenous (assumed controllable) and the others being endogenous (some or all of the other response types). After the structural system has been reduced to the predictive system; e.g., the endogenous response types are functions of only the exogenous variables; the first directive is to estimate jointly the parametric coefficients in a system of linear regression models; i.e., models which are linear in the unknown coefficients and errors, where errors within, but not between, an experimental unit are correlated, and the independent variables; i.e., the exogenous variables and functions thereof, are not necessarily the same for each response type. If the latter is true and the covariance matrix (or the relative magnitude of its elements) of the
"within" experimental unit errors is unknown, then, under the normality assumption, the maximum likelihood estimates of the covariances and coefficients are usually unattainable (in a practical sense) due to the complexity of the nonlinear equations resulting from equating to zero the partial derivatives of the likelihood function. As a result, an iteration scheme is proposed for finding estimates of the parameters. Moreover, a sufficient condition for convergence when iterating is given along with an approximate covariance matrix of the resulting coefficient estimates.

There is no need to consider the latter regression system as such if interest is centered on the effects of levels of the exogenous variables (with, say, one level of each exogenous variable combined into a so-called treatment as in a factorial experiment) on the pattern of any particular endogenous response type, for then, a series of univariate analyses of variance may be the most fruitful recourse; i.e., the response types are analyzed separately. Of course, joint probability statements across response types (ignoring the "within" experimental unit correlations) are erroneous unless errors within experimental units are uncorrelated or approximately so. In this connection, if it is desired to study the effects of a common set of exogenous variables on the patterns of the endogenous response types jointly, then Roy's multivariate analysis of variance techniques are available. Here, one may test contrasts between treatments (for each response type) and within treatments (for different
response types) simultaneously with associated confidence bounds following therefrom. Unfortunately these methods are, at present, somewhat esoteric.

In any event, the "real world" statistician is conscious of the fact that many pertinent questions regarding relationships between endogenous response types are left unanswered by analysis of variance. Further, it is inconceivable to the writer that certain questions could be answered without at least establishing

(i) that the syndromic response types (or functions thereof) can be predicted to some degree and
(ii) the pattern or shape of such surfaces resulting from the prediction.

One such class of questions leads to problems of extrema or constrained extrema for which the predictive regression system is particularly useful. For problems of extrema, conditions on the exogenous variables are found (point estimates and confidence regions) under which linear functions of the predicted endogenous response types are extremal. Problems of constrained extrema may arise in regression systems when extreme values of linear functions of certain response types (say, those in the same unit of measure) are wanted subject to constraints, the constraining functions being defined by, say, one or more of the other response types (possibly those in different units of measure).

All of the foregoing is discussed with respect to a factorial experiment designed for one stage of a laboratory test
used in the quantitative determination of toxoplasma anti-
bodies. However, prior to the outright consideration of the
aforementioned, a detailed discussion of the statistical
aspects of this laboratory test is presented to give perhaps
a greater insight into the nature of the problem.
2.0 REVIEW OF LITERATURE

One of the earliest and perhaps most fruitful applications of regression systems could have occurred with Wright's original work in path analysis [42] shortly after World War I. However, Wright, undoubtedly influenced by the statistical vogue of that period, chose to construe causal relationships in terms of correlation coefficients, although he repeatedly and properly stressed the distinction between causation and correlation. More recent criticisms of using correlation analyses in terms of causation were given in 1953 by Tukey [35] who, very convincingly, related the priority of regression over correlation. Tukey's approach to path analysis was further developed in extensive studies by Turner and Stevens [36] and Turner, Monroe, and Lucas [37].

In the middle thirties came possibly the first direct application of linear regression systems with the publication of Hotelling's Most Predictable Criterion [16]. In another article [17] in 1935, he considered the problem in a more symmetrical manner. Denote a system of linear regression equations by

\[
\hat{y} = \begin{bmatrix}
\hat{y}_1 \\
\cdots \\
\hat{y}_j \\
\cdots \\
\hat{y}_q 
\end{bmatrix} = \hat{B}x = \begin{bmatrix}
\hat{b}_{11} & \hat{b}_{21} & \cdots & \hat{b}_{p1} \\
\cdots & \cdots & \cdots & \cdots \\
\hat{b}_{1j} & \hat{b}_{2j} & \cdots & \hat{b}_{pj} \\
\cdots & \cdots & \cdots & \cdots \\
\hat{b}_{1q} & \hat{b}_{2q} & \cdots & \hat{b}_{pq}
\end{bmatrix} \begin{bmatrix}
x_1 \\
\cdots \\
x_i \\
\cdots \\
x_p
\end{bmatrix}
\] (2.1)
where, if \( y_j \) does not depend on \( x_i \), the corresponding \( b \) in \( \hat{B} \) is defined to be zero, and the \( x \)'s are deviations from their respective means. In \( I \theta 7 \), a linear function of \( \hat{y}_j \) is taken, say

\[
\mathbf{t}' \hat{y} = \mathbf{t}' \hat{B} \mathbf{x} ,
\]

where \( \mathbf{t} \) is chosen such that \( \mathbf{t}' \hat{y} \) will have the greatest multiple correlation, say \( R^2 \), with \( \mathbf{x} \). Since

\[
R^2 = 1 - \frac{\mathbf{t}' \mathbf{V} \mathbf{t}}{\mathbf{t}' \mathbf{U} \mathbf{t}} ,
\]

where \( \mathbf{V} \) is the matrix of residual sum of squares and cross products; i.e.,

\[
\mathbf{V} = (\text{SSE}_{jj'}) ,
\]

and

\[
\mathbf{U} = \left[ \sum_{k=1}^{n} (y_{jk} - \bar{y}_j)(y_{jk'} - \bar{y}_{j'}) \right] ,
\]

\( j, j' = 1, 2, \ldots, q \); then by setting

\[
\frac{\partial R^2}{\partial \mathbf{t}} = 0 ,
\]

\( \mathbf{t} \) is easily found to be the characteristic vector corresponding to the smallest characteristic root of \( \mathbf{U}^{-1} \mathbf{V} \). Hotelling assumed the endogenous response types (or criteria variables in his terminology) to be either mutually independent or regressed on the same vector \( \mathbf{x} \), both of which allow a separate least squares estimation of each parameter vector \( b_j \). However, in many situations, response types are neither
independent nor dependent on the same exogenous variables (at least not in the same linear functional form). By assuming independence when there exist dependencies between response types, the coefficients estimates $\hat{b}_j$ may be inefficient (though unbiased and consistent) when each $b_j$ is estimated separately by least squares (see Chapter 6).

In the immediate period prior to the development of high-speed computers (lasting roughly through the 1940's), methods were given by Hotelling [187], Turing [337], and others for finding better approximations to the inverse of a non-singular square matrix so as to reduce the error in the solution $\underline{z}$ in the linear simultaneous equations

\[ \underline{u} = \underline{A} \underline{z} \]  

(2.2)

Lonseth [237] considered these errors which may have stemmed from rounding in the inversion of $\underline{A}$ and/or observational errors in $\underline{u}$. Other sources of error in $\underline{z}$ could be given; e.g., if (2.2) is a regression system, $\underline{A}$ is subject to error due to its determination by least squares, but it was not intended in [237] to single out any particular cause. Rather, the aim was to find a series expansion for the error in the $z$'s, assuming that (2.2) is a nonhomogeneous system of linear equations with $\underline{A}$, $\underline{u}$, and $\underline{z}$ having additive error components. By means of a Taylor's expansion, a series was obtained for the solution error as a function of the $q(q+1)$ coefficient errors. Conditions for its convergence were then found, giving not only approximations to the maximum error but also
an upper limit. Earlier writings also considering errors in the z's individually are cited in [237].

In the early fifties and even before, the treatment of multiple equations (with dependent errors) was being carried on in econometrics. Here, main efforts were directed to finding estimates of structural parameters. The estimable or nonestimable character of the latter was investigated in detail with the method of limited information, published in 1953 by Hood and Koopmans [157], enabling the estimation of overidentified structural parameters.

At this same time, closely related works were published by Scheffe [297] in "tests for all possible contrasts" and by Roy and Bose [277] in simultaneous confidence interval estimation. (Since then, both topics have been extensively developed by Roy [267] in many different connections.) It was natural, then, to eventually consider the errors in the z's in (2.2) not individually, as did Lomseth [237], but simultaneously, as did Box and Hunter [47] in 1954. In application of the latter article, Box and Hunter extended an argument by Fieller [107] in finding a confidence region for the functionally independent x's which give rise to the stationary point in a single second-order regression model. However, difficulties may occur not only in establishing the confidence region (laborious algebra), but also in visualizing it for applied purposes. As a result, two subsequent papers serve as possible remedies. First, Box and Hunter [57] have shown that a very considerable simplification in the expression
for the exact simultaneous confidence region occurs when a rotatable design is used to generate the experimental data \[ \text{[57]} \]. Another alternative is to give an approximation to the true confidence region, as did Wallace \[ \text{[40]} \], by using a theorem by Scheffe \[ \text{[29]} \]. Here the approximations take the form of convex polyhedra or parallelopipeds with the approximations being easy to compute. References \[ \text{[41]} \] and \[ \text{[40]} \] are discussed in greater detail in Chapter 7.

Finally, attention is directed to a 1958 publication by Williams \[ \text{[41]} \] who used regression systems for inverse estimation as, say, in a multivariate bioassay situation. Assuming (2.2) to be a regression system with the z's still functionally independent but A of order qxq', it is Williams' intent that the u's be the observed variables and the z's, the predicted; i.e., given a new vector of observation \( \underline{u} = \underline{o} \underline{u} \), what is the estimate of \( \underline{z} \) (attainable if the dimension of \( \underline{z} \) exceeds the row rank of A by one) and the fiducial region for that \( \underline{z} \), say \( \text{E}(\underline{z}) \), which would have been attained had \( \underline{o} \underline{u} \) not been subject to error. Williams presents methods for finding the region \( R \) in the statement given \( \underline{u} = \underline{o} \underline{u} \), \( P[\text{E}(\underline{z}) \underline{e} R] = 1 - \alpha \); i.e., given a new observation vector \( \underline{u} = \underline{o} \underline{u} \), the fiducial probability that \( \text{E}(\underline{z}) \) lies within a q' dimensional hyperellipse is \( 1 - \alpha \).
3.0 TOXOPLASMSOSIS

3.1 A Discussion of Toxoplasmosis

A discussion of toxoplasmosis is given by Sonnenworth \(307\). In brief review, toxoplasmosis is a disease of man and animal stemming from the protozoan parasite, Toxoplasma gondii. In humans, the disease is produced in two forms, congenital (existing at birth) and acquired. In the former case, it is believed that toxoplasmosis occurs when the organism invades a nonimmune pregnant woman, with parasitemia occurring upon infection of the fetus. The infection is unapparent in the mother, but, depending on which stage of pregnancy, the infection was contacted

(i) the fetus may die in utero or after abortion;
(ii) the fetus may be born prematurely with active toxoplasmosis; or
(iii) the fetus may be normal at birth and appear well for days, even months, before symptoms are noted.

In (ii) and (iii), possible consequences include cerebral calcification and retardation. Postnatal infections may be so mild as to be unrecognized, with more severe infections afflicting the central nervous system.

3.2 A Serologic Test for Detecting the Presence or Absence of Toxoplasmic Antibodies

From the many contradictory findings reviewed in \(307\), it is evident that the symptoms of toxoplasmosis are extremely variable necessitating a heavy reliance on laboratory
test findings. The serologic test considered in this writing is the Sabin-Feldman \(^{28}\) cytoplasm-modifying dye test.

Sabin and Feldman, while searching for some in vitro method of detecting the presence of toxoplasma antibodies in human serum, discovered that certain chemical compositions have been found capable of indicating antibody activity. Namely, the cytoplasm of toxoplastic organisms in fresh exudate (extracted from inoculated mice) is stained a deep blue by alkaline methylene blue. However, when the cytoplasm of toxoplasma is acted upon by specific antibodies and accessory factor, it loses its affinity for methylene blue and the cytoplasm remains unstained. An explanation of this is found in a recent theory that the antibodies disrupt the cell wall of the toxoplasma organism, releasing the cytoplasm so that no staining can take place. Consequently, the dye test offers a quantitative measure of the absence or presence of toxoplasma organisms.
4.0 STATISTICS AND THE DYE TEST

4.1 The Probit Transform and LD 50

When determinations of unknown sera are made, one or more positive controls are tested. A positive control is serum with toxoplasma antibodies. The degree of positiveness exhibited by a given serum is measured quantitatively in terms of titers. In explanation of the latter, serum is diluted most easily in dilutions of 1:1, 1:2, 1:4, 1:8, 1:16, 1:32, 1:64, etc. A dilution of, say, 1:256 is one part serum to 255 parts saline. The more antibodies contained in serum, the higher the dilution at which they will be detected. From what has been said previously, the more unstained organisms detected (by means of the dye test), the more toxoplasma antibodies present.

At each dilution, organisms are counted from five fields on a particular slide so that

\[
\text{Percent stained} = \frac{\text{number of organisms stained}}{\text{number of organisms counted}}
\]

That dilution at which 50 percent of the organisms are stained is called the LD 50.

LD 50 is a characterization used in early toxicity tests by Trevan. In applying various doses of a stimulus, e.g., poison, to a group of subjects, it is convenient to characterize the stimulus by its effectiveness at a certain dose. It is often difficult to kill all subjects, except possibly at very high doses, so that the lethal dose for which
all subjects are killed (denoted by LD 100) is a poor measure. Conversely, the dose at which no subjects are killed (LD 0) would be an extremely poor characterization, since some subjects would die when exposed to a placebo. As a characteristic of the stimulus which can be more easily determined and interpreted, Trevan has advocated the median lethal dose (LD 50), or, as a more general term to include responses other than death, the median effective dose (ED 50). This is defined as the dose which will produce a response in half the population, and thus, from another point of view, is the mean tolerance. In the dye test we consider those unstained organisms as destroyed by the action of toxoplasmic antibodies. For this reason LD 50 is used as the characterization of any serum rather than ED 50. Consequently the LD 50 gives a quantitative measure of the degree of positiveness, and in this context we define a titer as the LD 50.

An individual is usually regarded as positive if 50 percent or more of the organisms are unstained at a dilution of 1:16 or above, though the criterion is controversial.

It should be mentioned that, for a high positive control with known LD 50 (known from many previous determinations), countings are made at dilutions ranging from 1:256 to 1:8192 or higher; i.e., the percent of unstained organisms is found at 1:256, 1:512, 1:1024, 1:2048, 1:4096, and 1:8192. Low positive controls, i.e., those with LD 50 \( \leq 1:64 \), are unsatisfactory, since they have been found to be unstable in this dye test.
The dilutions are coded to $\log_2$ on the abscissa with the percent unstained plotted on the ordinate axis as in Figure 4.1. But a straight line relationship is desirable for ease in estimating the LD 50. Since the resultant curve resembles a cumulative normal distribution, the percent unstained, $\hat{p}$, is transformed to $t$ where

$$\hat{p} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx,$$

and $t = \text{probit} - 5$. This transformation, called the probit transform, produces approximately a straight line in the range of interest. For a complete discussion of the probit transform see Finney [17]. Figure 4.2 depicts a typical high positive control plotted on probit graph paper. The resultant curve is three intersecting straight lines which we designate as the upper plateau, the line of descent, and the lower plateau. By inverse estimation, the LD 50 is read from the line of descent. In Figure 4.2 the LD 50 = 1:1024.

At first glance two statistical problems are posed:

(i) What are point and interval estimates of the LD 50 for the control?

(ii) What are the abscissal point and interval estimates of intersection of the straight lines?

4.2 Linear Weighted Regression in Estimating the Line of Descent

In answer to the first question, it is evident from the previous discussion that one would initially assume the scatter around the curve prior to transformation to be
Figure 4.1 Percentage modified plotted on the arithmetic scale at twofold dilutions from 1:256 through 1:4096
Figure 4.2 Percentage modified plotted on the probit scale at twofold dilutions from 1:256 through 1:4096
binomial. As such, the variability differs from dilution to dilution and is maximum around the true response corresponding to the LD 50. With n large, one might not choose to use probit analysis \[T1\], i.e., nonlinear regression, but rather, linear weighted regression. (Besides, from what is to be said in Section 4.3, there is reason to question the probit model in the sense that there may be "interaction in the design" in which case inferences may go no further than the data on hand.) If we were to fit a cubic regression line to the untransformed data, as in Figure 4.1, weights could be assigned from the value of \(\hat{p}q/n\) at each dilution. But with a straight line relationship being desirable in the range of interest, a linear weighted regression of the transformed \(\hat{p}\)'s on the coded dilutions is in order. In this case, weights are found approximately by the propagation of error technique \[8\]. From (4.1),

\[
\hat{p} = F(t) \\
d\hat{p} = f(t)dt \\
(d\hat{p})^2 = [f(t)]^2(dt)^2 \\
\approx [f(t)]^2 \text{var } t \\
\approx \text{var } \hat{p}
\]

where F(t) denotes the cumulative normal distribution. Hence

\[
\text{var } (t) \approx [\text{var } \hat{p}] [f(t)]^{-2} = pq/n [f(t)]^{-2}
\]
designates weights to the variances in the line of descent regression where \( f(t) \) is found in the tabulated normal tables.

The method of obtaining a confidence interval for the true LD50 is well known; e.g., see exercise 13.1, Anderson and Bancroft [1].

With regard to question (ii) of the last section, the abscissal values of the estimated intersection points are easily found by estimating the upper and lower plateaus and the line of descent, then finding their intersections. The question of corresponding confidence intervals can be resolved by applying the propagation of error technique or Fieller's theorem.

4.3 The Role of Compound Distributions in the Dye Test

Experimentation was designed to determine whether the slope of the line of descent (for a given control) remained roughly constant from one day test to the next; i.e., the line of descent contrast by dye test interaction was tested for significance. With a nonsignificant interaction, one would merely have to adjust or shift the control line of descent by a certain amount to account for the dye test effect. Then sera determinations would be adjusted by the same amount. Unfortunately, the interaction was highly significant.

It was found that the \( p_i \) (the probability of an organism surviving at dilution \( i, i=1,2,\ldots,N \)) appeared to vary from one field to the next for all \( i \). It was then reasoned that there existed some type of "contagion" [24] between organisms
on reasonably distinct fields on a given slide. As such, \( p \) was hypothesized to follow a beta distribution.

\[
f_i(p)dp = \frac{1}{\beta(\alpha_i, \gamma_i)} p^{\alpha_i-1} q^{\gamma_i-1}dp,
\]

\( i=1,2,\ldots,N, \) \( N \) being the number of dilutions. By fixing the number of organisms counted on any field of a particular slide at \( n' \), compounding the beta and binomial distributions, and integrating out \( p \), the resultant distribution is

\[
\int_0^{n'} \left( \begin{array}{c}
\frac{n'}{x} \\
\alpha_i + x, \gamma_i + n' - x
\end{array} \right)_p \beta(\alpha_i, \gamma_i) p^{\alpha_i-1} q^{\gamma_i-1}dp
\]

\[
= \left( \begin{array}{c}
\frac{n'}{x} \\
\alpha_i, \gamma_i
\end{array} \right)_p \frac{\beta(\alpha_i+n, \gamma_i+n')}{\beta(\alpha_i, \gamma_i)},
\]

where \( \alpha_i \) and \( \gamma_i \) are easily, though not most efficiently, estimated by equating ratios of moments. This distribution was first published by Skellam [31] in 1948, but it is likely that many were aware of its existence since 1920, at which time Yule and Greenwood [44] derived the Poisson-gamma (the compound Poisson distribution) analogy of the beta-binomial distribution.

The beta-binomial distribution gave a close fit at each dilution so that the assumption of "contagion" between organisms on a field appears warranted.

Since \( n' \) is by no means constant, another refinement can be made on the explanatory distribution [24] around the untransformed line, e.g., the curve or untransformed line of Figure 3.1. The nonconstant \( n' \) may stem from the fact that
each of the original organisms inoculated into the system may produce a different number of organisms. With \( n' \) varying according to the probability law, say, \( g(n') \), a further compounding results in

\[
\sum_{n' \geq 0} \binom{n'}{x} p^n q^{n'-x} \frac{p^{\alpha} q^\beta - 1}{\beta(\alpha', \beta')} g(n') dp. \tag{4.2}
\]

Letting \( g(n') \) be the Poisson distribution with parameter \( \eta' \), the moment generating function of the distribution given by (4.2) is

\[
M_x(\theta) = 1_F \left[ \alpha', \alpha + \eta', \eta'(\theta - 1) \right], \tag{4.3}
\]

a confluent hypergeometric function \( \mathcal{F}_1 \). The distribution defined by (4.2) is similar to Neyman's Type A contagious distribution \( \mathcal{F}_{24} \) and was derived during the course of Gur- land's extensions \( \mathcal{F}_{12} \) of Neyman's original work.

The explanatory distribution approach plays a vital role in many stable biological processes, as in Neyman's first application \( \mathcal{F}_{24} \). We have attempted to use this approach in an unstable process; i.e., this particular dye test procedure leads to a wide variety of contradictory conclusions. In doing so, (4.2) was obtained under a set of conjectured occurrences, whereas the same moment generating function given by (4.3) may well have been derived under a totally different set of occurrences; e.g., one may arrive at the negative binomial distribution under at least three different sets of assumptions. On intuitive grounds, it is felt that if the
explanatory distribution approach is used in an unstable process rather than in a stable process, then there is a higher risk of ascribing the end result to one chain of events when, in fact, something different is occurring with the same end or distribution.

In statistical terminology, the instability of the dye test is characterized by the slopes of the line of descent and the plateaus, and the LD 50 of the control which vary widely from one dye test to the next. Further, the slopes for unknown sera, at times, differ drastically from the control slopes in a given test.

Consequently the intent is to isolate the source or sources of trouble in the dye test. As a result, experimentation has progressed "backwards" through the various stages of the dye test.
5.0 THE CYCLE OF INVESTIGATION AS APPLIED
TO ONE STAGE OF THE DYE TEST

As was mentioned previously, exudate, extracted from inoculated mice, is used in the dye test. The question arises as to what amount and dilution of inoculum to use in order that the exudate extracted at a particular time, \( t_0 \) (\( t_0 = 72 \) hours), have a maximum number of toxoplasma organisms, a minimum number of leucocytes (white blood cells which destroy the toxoplasma organisms), and maximum volume. Time from inoculation to extraction is not considered, since previous experiments have determined the optimal and most economical time differential.

A factorial experiment was designed to gain perspective into the three measured response types: number of toxoplasma organisms (\( y_T \)), number of leucocytes (\( y_L \)), and volume of exudate (\( y_V \)). The experiment included various levels of dilution of inoculum (\( x_D \)) and amount of inoculum (\( x_A \)), with exudate from three mice being pooled at each treatment combination.

Data are not presented as the residual variability was excessive for all three response types. Further, there appeared to be a substantial block = day x treatment interaction. However, the majority of the mice were dry or contaminated so that many observations were missing or based on exudate from one mouse. Consequently, it is questionable that the block x treatment interaction is real and that the additive models are completely inadequate.
Suppose we now consider the problem with regard to the cycle of investigation. With the experimental unit being those mice inoculated with the same amount and dilution of inoculum and the exudate being the phenomenon of interest, then the three endogenous variables characterizing the phenomenon are \( y_T \), \( y_L \), and \( y_V \); the two exogenous variables are \( x_A \) and \( x_D \).

![Diagram](image)

**Figure 5.1** A possible series of events characterizing the phenomenon of interest

Figure 5.1 illustrates a hypothetical chain of occurrences at a fixed time, \( t_0 \), after inoculation. During \( t_0 \), the L (leucocyte organisms) are engulfing the T (toxoplasma organisms). The more T produced, the more L the body produces \( \rightarrow \) (1), Figure 5.1\textsuperscript{7}. On the other hand, the more L the body produces, the fewer the number of T \( \rightarrow \) (2)\textsuperscript{7}, since the former supposedly destroy the latter. Finally, volume results from the interchange between the T and L \( \rightarrow \) arrows (3) and (4)\textsuperscript{7}. It should be stressed that Figure 5.1 vastly oversimplifies the true situation, and the preceding discussion thereof could be wrong, especially if the T thrive on the L. Further, to be realistic with regard to the structural system, the element of time should be considered. However,
the writer is at this time more concerned with the method than
the problem.

From Figure 5.1, the following relationships result:

\[ y_T + \delta_T = f_T(y_L + \delta_L, x_A, x_D) + \epsilon_T \]
\[ y_L + \delta_L = f_L(y_T + \delta_T, x_A, x_D) + \epsilon_L \]
\[ y_V + \delta_V = f_V(y_T + \delta_T, y_L + \delta_L) + \epsilon_V \]  
(5.1)

where the \( \delta \)'s represent measurement error, the \( y \)'s measured
values, and the \( \epsilon \)'s are inherent errors; i.e., the \( \epsilon \)'s are
errors associated with the models. The \( \delta \)'s can be made ar-
bitrarily small, or even negligible, depending on the refine-
ment of the measurement technique. Usually the measurement
error is pooled with the inherent error through no other re-
course. Further, it is assumed that the \( \delta \)'s and \( \epsilon \)'s are un-
correlated with each other (within a particular response type)
and the \( x \)'s, \( E(\epsilon) = 0, E(\epsilon_T \epsilon_L) = \sigma_{TL}, E(\epsilon_T \epsilon_V) = \sigma_{TV}, E(\epsilon_L \epsilon_V) = \sigma_{LV}, \)
\( E(\epsilon_T^2) = \sigma_{TT}, E(\epsilon_L^2) = \sigma_{LL}, \) and \( E(\epsilon_V^2) = \sigma_{VV}. \)

As a very crude first attempt, one might choose

\[ y_T = a_0 + a_1 y_L + a_2 x_A + a_3 x_D + a_4 x_A x_D + a_5 x_A^2 + a_6 x_D^2 + \epsilon_T' \]
\[ y_L = b_0 + b_1 y_T + b_2 x_A + b_3 x_D + b_4 x_A x_D + b_5 x_A^2 + b_6 x_D^2 + \epsilon_L' \]
\[ y_V = c_0 + c_1 y_T + c_2 y_L + \epsilon_V' \]  
(5.2)

as the structural system of models at time, \( t_0, \) where

\[ \epsilon_T' = \epsilon_T + a_1 \delta_L + \delta_T; \quad \epsilon_L' = \epsilon_L + b_1 \delta_T + \delta_L; \quad \text{and} \quad \epsilon_V' = \epsilon_V + c_1 \delta_T + c_2 \delta_L + \delta_V. \]
The equations of (5.2) are called the structural regression models. Ordinary regression analyses are not appropriate, since the "independent" variables are correlated with the error $\varepsilon_T$; e.g., in the first equation, $y_L$ is correlated with $\varepsilon_T'$, since $y_L = \gamma_L + \delta_L$ where $\gamma_L$ is the true measurement value. However, the structural regression models of (5.2) can be reduced to

$$
y_T = a_0' + a_1'x_A + a_2'x_D + a_3'x_Ax_D + a_4'x_A^2 + a_5'x_D^2 + \varepsilon_T''
y_L = b_0' + b_1'x_A + b_2'x_D + b_3'x_Ax_D + b_4'x_A^2 + b_5'x_D^2 + \varepsilon_L''
y_V = c_0' + c_1'x_A + c_2'x_D + c_3'x_Ax_D + c_4'x_A^2 + c_5'x_D^2 + \varepsilon_V''\tag{5.3}
$$

where, by substituting for the $y$'s occurring as "independent" variables, it is seen that

$$
a_0' = \frac{a_0 + a_1b_0}{1 - a_1b_1}; \quad a_1' = \frac{a_2 + a_1b_2}{1 - a_1b_1}; \quad a_2' = \frac{a_3 + a_1b_3}{1 - a_1b_1};
$$

$$
a_3' = \frac{a_4 + a_1b_4}{1 - a_1b_1}; \quad a_4' = \frac{a_5 + a_1b_5}{1 - a_1b_1}; \quad a_5' = \frac{a_6 + a_1b_6}{1 - a_1b_1} \tag{5.4}
$$

and similarly for the $b$'s and $c$'s. The equations of (5.3) are called the reduced equations or the predictive regression models, and for the system of (5.3), ordinary regression analyses are appropriate. If, however, the $y$'s should not be dependent, functionally, on the same $x$'s, and the errors "between models," i.e., within experimental units, are correlated, then the joint estimation of the parametric coefficients may present difficulties. This problem is considered in Chapter 6.
After estimating the coefficients in the predictive regression system, the latter are used to estimate the parameters of the structural regression system. But in this problem, we have the difficulty of overidentification; i.e., there exist 17 unknowns, $a_0, \ldots, a_6; b_0, \ldots, b_6; c_0, c_1, c_2$, in 18 equations, 6 of which are given by (5.4). In this case, the Hood-Koopmans method of limited information $I5$ is available for estimating the structural parameters, with the variances of the resulting estimates being obtained by the propagation of error technique $I8$. With more unknowns than equations, there is the problem of underidentification with no known solution. (This may possibly be remedied by imposing further relationships derived from physical considerations.)

It would be surprising if the structural regression system (5.2) proves to be anything but a rough and possibly invalid approximation. It is thought that eventually (in some future cycle of investigation), models of a nonlinear nature, and involving the element of time after inoculation, will be deduced (from one or more differential equations) to replace the system (5.2). This should, however, not dampen one's hopes of answering the question posed at the beginning of this section due to the fact that any one of many structural systems (possibly the "correct one") could have given rise to (5.3). Thus, even though the structural system (5.2) may be totally unrealistic, the resulting predictive system (5.3) may offer valid answers to what amount and dilution of
inoculum to choose in order that exudate (extracted at time, \( t_0 \)) have maximum volume and a maximum difference in opposing organisms.

There were scant and possibly misleading indications that within the range of the factor space considered, each of the response types achieved a maximum. Further, it appeared that these maximums occurred for different \( x_A, x_D \) values with \( \max y_T \) and \( \max y_L \) being roughly in the same locality of the \( x_A, x_D \) dimensions. In this connection, one function of interest is \( \phi(y_T, y_L) = y_T - y_L \), which is estimated by the difference in predicted values,

\[
\phi(\hat{y}_T, \hat{y}_L) = \hat{y}_T - \hat{y}_L = \hat{a}'x - \hat{b}'x,
\]

\[
x' = (x_A, x_D, x_A^2, x_D^2),
\]

where the \( x \)'s are deviations from their respective means and within each response type, the variance is heterogeneous necessitating, possibly, a transformation. By taking partial derivatives in (5.5), we are able to estimate by \( (\hat{x}_A, \hat{x}_D) \) any \( (x_A, x_D) \) value, say \( E[\hat{x}_A, \hat{x}_D] \), for which \( x_T - y_L \) is maximum. Further, if \( \hat{y}_T \) and \( \hat{y}_L \) are second-order regressions, a confidence region is available for \( E[\hat{x}_A, \hat{x}_D] \) which, among other things, provides the laboratory worker with a range of amounts and dilutions of inoculum to aim at rather than one value which may be difficult to make up. The problem of bringing \( y_V \) into the picture is dealt with in detail in Chapter 8.
Note that we have considered the difference in regressions \( \hat{y}_T - \hat{y}_L \) rather than regressing \( y_T - y_L \) on the \( x \)'s. It is easily seen that \( \hat{y}_T - \hat{y}_L = \hat{y}_T - \hat{y}_L \) if \( y_T, y_L, \) and \( y_T - y_L \) are regressed, linearly, on the same \( x \)'s. However, the error mean square estimating \( \sigma^2_{TL} \) pools the entities \( \sigma_{TT}, \sigma_{LL}, \) and \( \sigma_{TL} \) which can be calculated with little effort and should be for a better conception of the situation. Further, if one can adequately predict the response patterns for \( y_T \) and \( y_L \), then the difference in responses is automatically known. Obviously, the converse is not true, and by regressing \( y_T - y_L \) on the \( x \)'s without investigating the response patterns of \( y_T \) and \( y_L \) separately, one may run the risk of fitting, say, a cubic regression when, in fact, \( y_T \) and \( y_L \) follow quadratic patterns.

As the problem now stands (in the first cycle of investigation), the methods of solutions are relatively simple partly due to the fact that the correlated response types are (in the predictive system) related (linearly) to precisely the same \( x \)'s. We now pass on to other aspects of estimation in more difficult predictive systems.
6.0 SYSTEMS OF LINEAR REGRESSION EQUATIONS

6.1 The Nature of Errors Associated
with Linear Regression Systems

In analyzing relationships between different types of
dependent responses where each response type is to be re-
gressed on a set of fixed or controlled variables, one should
first give careful scrutiny to the nature of the errors as-
associated with the models. (Since we are concerned with linear
regression, the discussion will pertain to models which are
linear in the unknown coefficients and errors, unless other-
wise stated.)

It should be realized that in any regression relationship
the inherent error, \( \epsilon \) (the error associated with the model),
is never truly noncasual, and, in fact, is a function of other
extraneous, uncontrolled variables or disturbances. For ex-
ample, the model

\[
\eta = \sum_{i=1}^{p} b_1 x_1 + \epsilon \quad (x_0 = j),
\]

\[
\epsilon = g(\xi_1, \xi_2, \xi_3, \ldots),
\]

where the \( \xi \)'s are disturbances. In applying the above model,
it is hoped that the \( \xi \)'s contribute little to \( \eta \) in terms of
casuality; i.e., that the major factors causing \( \eta \) are included
in the \( x \)'s, and that the \( \xi \)'s tend to offset each other. Let
\( x_k = (x_{1k}, \ldots, x_{ik}, \ldots, x_{pk}) \) denote a treatment applied at random
to an experimental unit. Randomization not only tends to make
the correlation between \(\varepsilon_k\) and \(\varepsilon_{k'}\) (\(k \neq k'\)) cancel as the treatments \(x_k\) and \(x_{k'}\) are applied to an increasing number of experimental units, but also neutralizes many of the disturbances (or biases). However, as the treatment, \(x_k\), is applied to an increasing number of units (resulting in \(\gamma_{k1}, \gamma_{k2}, \ldots\)), it does not follow that \(\varepsilon_{k1}, \varepsilon_{k2}, \ldots\) neutralize each other to the extent that \(E(\varepsilon|x_k)=0\), since it is inconceivable that all the "minor" casual agents (the \(\xi\)'s) will be offset (either within or between experimental units) in the repeated application of \(x_k\). As such, the most that should be said with regard to the inherent error associated with a particular regression model is that the average value of \(\varepsilon\) is approximately zero; i.e., \(E(\varepsilon)=0\). Consequently, for the class of all linear models, the inherent error will be called weakly casual when we assume \(E(\varepsilon)=0\), when actually \(E(\varepsilon)\neq 0\), and strongly casual when we correctly assume that \(E(\varepsilon)\neq 0\). Note that weakly casual is not a necessary requirement for an adequate model; e.g., a simple, though biased model may be a better predictor than a complex though unbiased model. It is, however, convenient to use terms such as weakly and strongly casual to describe the nature of errors.

For purposes of illustrating the types of correlation between response types (arising from dependent errors), consider the following example. Among the characteristic responses of an experimental unit, let, say, two of interest be measured with error by \(y_1=\gamma_1+\delta_1\) and \(y_2=\gamma_2+\delta_2\) where the \(\gamma\)'s are true measurement values and the \(\delta\)'s are measurement errors. If
\( \gamma_1 \) and \( \gamma_2 \) can be expressed adequately and linearly in terms of \( x_1 \), and \( x_1 \) and \( x_2 \), respectively, then

\[
\eta_1 = b_0 + b_1 x_1 + \varepsilon_1 \quad (6.1)
\]
and

\[
\eta_2 = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon_2, \quad (6.2)
\]

where it is assumed the \( x \)'s are controlled or fixed. Now, \( E(\varepsilon_1 \varepsilon_2) \neq 0 \), since the experimental unit forces a relation between \( \varepsilon_1 \) and \( \varepsilon_2 \) or between \( \gamma_1 \) and \( \gamma_2 \); i.e., the extraneous variables influencing the experimental unit induce the non-independence between \( \varepsilon_1 \) and \( \varepsilon_2 \). With there existing no casual exchanges between the two response types, and with \( \varepsilon_1 \) and \( \varepsilon_2 \) being weakly casual, then the correlation between \( \gamma_1 \) and \( \gamma_2 \) stems from a linear type association between the random variables \( \varepsilon_1 \) and \( \varepsilon_2 \); i.e., the correlation between \( \gamma_1 \) and \( \gamma_2 \) originates from the nonindependence of the inherent errors, \( \varepsilon_1 \) and \( \varepsilon_2 \), which are weakly casual for \( \gamma_1 \) and \( \gamma_2 \), or, in a practical sense, are noncasual for \( \gamma_1 \) and \( \gamma_2 \). As such, a "regression type" correlation does not exist between \( \gamma_1 \) and \( \gamma_2 \).

By introducing the measurement errors, (6.1) and (6.2) become

\[
y_1 = b_0 + b_1 x_1 + \varepsilon_1 + \delta_1, \quad (6.3)
\]
and

\[
y_2 = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon_2 + \delta_2. \quad (6.4)
\]

Usually the measurement errors are assumed almost negligible so that \( \varepsilon \) and \( \delta \) are pooled. Note that if \( E(\varepsilon_1 \varepsilon_2) = 0 \) and
\( \text{E}(\hat{\xi}) = 0 \), while \( \text{E}(\hat{\xi}_1 \hat{\xi}_2) \neq 0 \), then there results a linear association type correlation between \( y_1 \) and \( y_2 \). In either case, we still have correlated errors and the parameters of the models (6.3) and (6.4) may be estimated jointly by methods given in the next sections.

Suppose, after one cycle of investigation, it is found that there are casual exchanges between the two characteristic responses, with, say, \( \gamma_2 \) influencing \( \gamma_1 \); e.g., the casual exchanges may be in terms of competition. Then, assuming additivity, the error \( \varepsilon_1 \) in (6.1) takes the form

\[
\varepsilon_1 = b_2 \gamma_2 + \varepsilon'_1 \quad (b_2 \neq 0)
\]

so that (6.1) becomes

\[
\gamma_1 = b_0 + b_1 x_1 + b_2 \gamma_2 + \varepsilon'_1.
\]

Under (6.5) the error, \( \varepsilon_1 \), in (6.1) is strongly casual for \( \gamma_1 \) with \( \varepsilon'_1 \) now taken as weakly casual. Then not only do the errors, \( \varepsilon_1 \) and \( \varepsilon_2 \), give rise to the linear association type correlation between \( \gamma_1 \) and \( \gamma_2 \), \( \text{E}(\varepsilon'_1 \varepsilon_2) \neq 0 \), but further, \( \varepsilon'_1 \), in the role of a strongly casual agent for \( \gamma_1 \), describes an additional correlation between \( \gamma_1 \) and \( \gamma_2 \), a "regression type" correlation.

Under (6.5), the structural models, (6.6) and (6.2), are referred to the reduced equations or predictive models, (6.2), and

\[
\gamma_1 = b'_0 + b'_1 x_1 + b'_2 x_2 + \varepsilon''_1,
\]
where \( b_0' = b_0 + b_2 a_0; \) \( b_1' = b_1 + b_2 a_1; \) \( b_2' = b_2 a_2; \) and \( \xi_1'' = \xi_1' + b_2 \xi_2. \) Then, assuming that

\[
E(\xi_1'') = E(\xi_2) = E(\xi_1') = 0,
\]

(6.8)

and \( E(\xi_1' \xi_2) = \sigma_{12} \neq 0 \) (the latter stemming from the effect of extraneous variables on the experimental unit), it follows that

\[
E(\xi_1''^2) = E[(\xi_1' + b_2 \xi_2)^2] = \sigma_{11} + 2b_2 \sigma_{12} + b_2^2 \sigma_{22},
\]

\[
E(\xi_2^2) = \sigma_{22},
\]

and

\[
E(\xi_1'' \xi_2) = E[(\xi_1' + b_2 \xi_2) \xi_2] = \sigma_{12} + b_2 \sigma_{22}.
\]

Although from (6.8), \( \xi_2 \) and \( \xi_1'' \) are weakly casual, a regression relationship exists between \( \xi_1'' \) and \( \xi_2 \), since \( \xi_1'' = \xi_1' + b_2 \xi_2. \) The latter will be true whenever it becomes necessary to refer a system of linear models to the reduced equations. Further, even if one were to assume that \( E(\xi_1' \xi_2) = \delta_{12} = 0 \) (which may be a precarious assumption when \( \gamma_1 \) and \( \gamma_2 \) stem from the same experimental unit), we still have correlated errors, since \( E(\xi_1'' \xi_2) = b_2 \delta_{22} \) for \( b_2 \neq 0 \) and \( \delta_{12} \neq 0. \)

By introducing the measurement error, (6.7) becomes

\[
y_1 = b_0 + b_1' x_1 + b_2' x_2 + \xi_1'' + \delta_1.
\]

Assuming that \( E(\delta) = 0; \) \( E(\delta_1^2) = \sigma_{12}^2; \) \( E(\delta_2^2) = \sigma_{22}^2; \) \( E(\delta_1 \delta_2) = \sigma_{12} \delta_2; \) and \( E(\delta \xi) = 0, \) then
\[
\begin{align*}
[E_\epsilon (\epsilon_1'' + \delta_1)^2] &= E_\epsilon [ (\epsilon_1' + b_2 \epsilon_2' + \delta_1)^2 ] \\
&= \sigma_{11} + 2b_2 \sigma_{12} + b_2^2 \sigma_{22} + \sigma_{\delta_1}^2; \\
E_\epsilon [(\epsilon_2' + \delta_2)^2] &= \sigma_{22} + \sigma_{\delta_2}^2; \\
\text{and} \\
E_\epsilon [(\epsilon_1'' + \delta_1)(\epsilon_2' + \delta_2)] &= E_\epsilon [(\epsilon_1' + b_2 \epsilon_2' + \delta_1)(\epsilon_2' + \delta_2)] \\
&= \sigma_{12} + b_2 \sigma_{22} + \sigma_{\delta_1} \sigma_{\delta_2}. 
\end{align*}
\]

Consequently, it is very likely that the pooled errors (for each response type taken from the same experimental unit) will be correlated since nonindependence will arise if at least one of the following conditions hold: (i) \(E(\epsilon_1' \epsilon_2') \neq 0\); (ii) \(b_2 \neq 0\); and/or \(E(\delta_1 \delta_2) \neq 0\).

Main emphasis in this section has been given to the nature of the errors. Whether the latter are strongly or weakly casual may, of course, be determined through one or more cycles of investigation. In any event, we have presented a strong argument for the existence of correlated errors between models, and this forms a premise for developments in the following sections.

6.2 Estimation of the Parametric Coefficients in a Linear Regression System When \(W\) Is Known

Let there exist \(n\) randomly selected experimental units with \(o_{yk} = (o_{y_{1k}}, \ldots, o_{y_{jk}}, \ldots, o_{y_{qk}})\) being a vector of \(q\) dependent measurements from the \(k^{th}\) unit which characterize a phenomenon of interest from the \(k^{th}\) unit, \(k=1, 2, \ldots, n\).
Assume that \( o^y_{jk} \) conforms to the general linear model,

\[
o^y_{jk} = a_j + b'_j o^x_{jk} + o^\epsilon_{jk},
\]

where \( o^x_{jk} = (o^{x_{1jk}}, \ldots, o^{x_{ijk}}, \ldots, o^{x_{pjk}}) \); \( i = 1, 2, \ldots, p_j \) is a vector of known constants (or at least controlled variables) or independent variables; \( a_j \) is an unknown parameter;
\( b'_j = (b_{1j}, \ldots, b_{ij}, \ldots, b_{pj}) \) is a vector of unknown parameters; and \( o^\epsilon_{jk} \) is a weakly casual error associated with the model. (For the remainder of this writing, we will, for convenience, assume the \( \epsilon \)'s to include both the measurement and inherent errors.) Further assume that

\[
o^\epsilon_k = 
\begin{bmatrix}
o^\epsilon_{1k} \\
\vdots \\
o^\epsilon_{jk} \\
\vdots \\
o^\epsilon_{qk}
\end{bmatrix} \sim N(0, \Sigma)
\]

and

\[
E\left(\frac{o^\epsilon_k}{o^\epsilon_{k'}}\right) = 0, \ k \neq k',
\]

where the symmetric matrix

\[
\Sigma = 
\begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1q} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2q} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{q1} & \sigma_{q2} & \cdots & \sigma_{qq}
\end{bmatrix}
\]
Set

\[ \mathbf{y}_j = (o_{v_1j}, \ldots, o_{v_kj}, \ldots, o_{v_nj}); \quad \mathbf{\xi}_j = (o_{\varepsilon_1j}, \ldots, o_{\varepsilon_kj}, \ldots, o_{\varepsilon_nj}); \]

\[ x_{ijk} = o_{x_ijk} - \bar{x}_{ij}; \quad \bar{x}_{ij} = \frac{1}{n} \sum_{k=1}^{n} (o_{x_ijk}/n); \]

and

\[ x_{ij} = (x_{i1}, \ldots, x_{i1}, \ldots, x_{ip}), \quad \text{nxp} \]

where if \( \mathbf{y}_j \) depends on \( p_{j\leq p} \) of the vectors of \( X \) in (6.13), then \( x_{ij} \) implies that \( \mathbf{y}_j \) depends on \( x_i \) and the latter is included once and only once in \( X \). Then an alternative population form of the model is

\[ \mathbf{y}_j = b_{j01} + x_{ij} \frac{b_j}{p_{j\leq p}} + \mathbf{\xi}_j \]

for \( j = 1, 2, \ldots, q \), where the subscript \( j \) on \( X_j \) implies that \( X_j \) includes the \( p_{j\leq p} \) column vectors of \( X \) upon which \( y_j \) depends.

Further,

\[ b_{j0} = a_j - \sum_{i=1}^{p_j} b_{ij} \bar{x}_{ij}. \]

Correspondingly, let the sample form of the model be written as

\[ \mathbf{y}_j = \hat{b}_{j01} + x_{ij} \hat{b}_j + \mathbf{e}_j \]

Consider now the determination of \( \hat{b}_{j0} \) and \( \hat{b}_j \) in (6.15). From (6.10), (6.11), and (6.14), it follows that
\[
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_j \\
  \vdots \\
  y_q 
\end{bmatrix}
\begin{pmatrix}
  b_{101} \\
  \vdots \\
  b_{j01} \\
  \vdots \\
  b_{q01}
\end{pmatrix}
\begin{bmatrix}
  x_1 \\
  \vdots \\
  x_j \\
  \vdots \\
  x_q 
\end{bmatrix}
\begin{pmatrix}
  b_1 \\
  \vdots \\
  b_j \\
  \vdots \\
  b_q
\end{pmatrix},
\]
\[
\text{where } \text{var}(y) = I \otimes \Sigma \text{ is a symmetric matrix written as the direct product or Kronecker product of } I \text{ and } \Sigma; \ i.e.,
\]
\[
\text{var } y = \begin{bmatrix}
  I \sigma_{11} & I \sigma_{12} & \ldots & I \sigma_{1q} \\
  I \sigma_{21} & I \sigma_{22} & \ldots & I \sigma_{2q} \\
  \vdots & \vdots & \ddots & \vdots \\
  I \sigma_{q1} & I \sigma_{q2} & \ldots & I \sigma_{qq}
\end{bmatrix} = I \otimes \Sigma.
\]

Suppose now \(X_j \neq X\) for some \(j\) and no \textit{a priori} information of \(\Sigma\) is available. If an attempt should be made to find estimates of \(b_{j0}, b_j,\) and \(\Sigma\) such that the likelihood function,
\[
L = \begin{pmatrix}
  2 \\
  -(n/2)
\end{pmatrix}^{-\left(nq/2\right)} \left|\Sigma\right|^{-\left(n/2\right)} \text{ times}
\]
\[
\exp \left\{ -(1/2) \sum_{k=1}^{n} \left[ o_{yk} - \text{E}(o_{yk}) \right] (\Sigma^{-1}) \left[ o_{yk} - \text{E}(o_{yk}) \right] \right\},
\]
is maximized, then the nonlinear system of equations (in the parameters), resulting from equating to zero partial derivatives of \(L\), is unsolvable in a practical sense, except possibly
in very simple cases. A discussion of estimation in this case is deferred until Section 6.4 in order that supplementary material can be presented.

If \( I*\Sigma \) can be rewritten as \( (I*W)\sigma^2 \), where \( \sigma^2 \) is unknown and \( W \) denotes a known weight matrix such that \( \Sigma = W\sigma^2 \), then, under (6.16), it is easily verified that the maximum likelihood estimates of the \( b_{j0} \) and \( b_j \) are the same as those attained through the method of weighted (or joint) least squares. This fact should be obvious since maximum likelihood, under (6.16) and known \( W \), and weighted least squares, under \( y : [E(y), (I*W)\sigma^2] \) and known \( W \), both provide minimum variance, linear, unbiased estimates of the \( b_{j0} \) and \( b_j \).

We will now derive these estimates by means of weighted least squares and refer the reader to Anderson for their derivation by maximizing the likelihood function when \( X_j = X \) for all \( j \).

Denoting \( K^{1/2} \) as the square root of a nonsingular \( k \times k \) matrix \( K \), where \( K^{1/2'}K^{1/2} = K \) and \( K^{-1/2'}KK^{-1/2} = I \), then the transformed vector, \( (I*W)^{-1/2'}y \), is distributed as

\[
(I*W)^{-1/2'}y \overset{\text{nqxnxq}}{\sim} \mathcal{N}(E(y), I\sigma^2).
\]

Now let \( E(y) \) in (6.16) be rewritten as \( E(y) = \hat{b}_0^* + \check{X}b \), where

\[
\hat{b}_0 = \begin{bmatrix}
\hat{b}_{10} & \vdots & \hat{b}_{q0}
\end{bmatrix}, \quad \check{X} = \begin{bmatrix}
X_1 & 0 \\
\vdots & \ddots & \vdots \\
0 & \ddots & X_j \\
& & 0 & X_q
\end{bmatrix}
\]
and $b' = (b'_1, \ldots, b'_j, \ldots, b'_q)$. From (6.15) we have

$$e = \begin{bmatrix} e_1 \\ \vdots \\ e_j \\ \vdots \\ e_q \end{bmatrix} = y - \hat{b}_0 - \hat{X}b.$$ 

Since $e_{jk}$ has variance directly proportional to $\sigma_{jj}$ and, say, $\sigma_{jj} < \sigma_{j'j''}$, it is intuitively evident that $e_j$ should have a "greater influence" than $e_{j'}$, in the determination of the estimates and as such, $e_j$ should be weighted more heavily; i.e., the weights imply that greater information is contained in those $e_j$ with smaller variances than those with larger variances. Hence, if the squares and cross products of the residuals are weighted inversely proportional to their weights in $W$ (implying that the squares and cross products of the residuals are to be weighted by their counterpart with $W^{-1}$) with the resulting quadratic form being minimized with respect to $\hat{b}_0$ and $\hat{b}$, then we are utilizing that quadratic in the $e'$s with the greatest information to obtain the estimates.

The latter argument with the mathematical justification following from (6.18) directs one to minimize

$$e'(I*W)^{-1}e = (y - \hat{b}_0 - \hat{X}b)'(I*W)^{-1}(y - \hat{b}_0 - \hat{X}b)$$

(6.19)

with respect to $\hat{b}_0$ and $\hat{b}$. Taking partials and equating to zero, we have
\[
\frac{\partial e'(I^*W)^{-1}e}{\partial b_0} = (y - \hat{b}_0 - \hat{x})'(I^*W)^{-1}1 = 0
\]
and
\[
\frac{\partial e'(I^*W)^{-1}e}{\partial \sigma} = (y - \hat{b}_0 - \hat{x})'(I^*W)^{-1}(-\hat{x}) = 0,
\]
so that
\[
\hat{b}_{j0} = \overline{y}_j,
\]
\[
\hat{a}_{j0} = \overline{y}_j - \sum_{i=1}^{p_j} \hat{b}_{ij} \overline{x}_{ij}
\]
in (6.9), and
\[
\hat{b} = [\hat{x}'(I^*W)^{-1}\overline{x}]^{-1}\overline{x}'(I^*W)^{-1}y.
\]  
(6.20)
For example, let \( q=2 \) and
\[
\Sigma = \Sigma \theta^2 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \theta^2.
\]
Then (6.20) becomes
\[
\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} x_1'x_1 & -\rho x_1'x_2 \\ -\rho x_2'x_1 & x_2'x_2 \end{bmatrix}^{-1}x_1(y_1 - \rho y_2),
\]
(6.21)
Clearly, when the response types are independent, i.e., \( \rho=0 \), (6.21) reduces to the usual separate least squares estimation of each \( b_j \).
Consider next the case when \( X_j=X \) for all \( j \), where \( X \) is given by (6.13). Then (6.20) reduces to
\[ \hat{b} = [(X'X)W^{-1}]^{-1}(X'Wy) \]
\[ = [(X'X)^{-1}W](X'Wy) \]
\[ = [(X'X)^{-1}X'y] * (W^{-1}y) \]
\[ \hat{b} = [(X'X)^{-1}X'y] * (I)y \]

or

\[ \hat{b}_j = (X'X)^{-1}X'y_j. \] (6.22)

Hence, when \( X_j = x \) for all \( j \), the weight matrix offers no information regarding \( \hat{b} \), and the separate least squares estimation of each \( b_j \) is appropriate.

The intuitive arguments underlying the foregoing developments become apparent in the following examples.

**Example 6.1** Consider the models

\[ y_1 = a_0 + a_1 x + \varepsilon_1 \]
\[ y_2 = b_0 + \varepsilon_2, \]

where \( x \) is a deviation; i.e., \( x \) takes on the potentially observable values,

\[ \hat{o}x_k = \frac{\sum_{k=1}^{n} (o_k^x/n)}{n}, \]

or the "already observed" values,

\[ o^x_k = \frac{\sum_{k=1}^{n} (o_k^x/n)}{n}, \]

depending on which came first, the model or the data.
Assume the errors to be weakly casual for any error, \( E(\varepsilon) = 0 \) and that
\[
E(\varepsilon^2) = \sigma^2
\]
\[
E(\varepsilon_{1k}\varepsilon_{2k'}) = \rho \sigma^2 \quad \text{if} \ k = k' = 1, \ldots, n
\]
\[
= 0 \quad \text{if} \ k \neq k'.
\]

With \( \rho \) known, the joint or weighted least squares estimates of \( a_0, a_1, \) and \( b_0 \) given by (6.20) are
\[
\hat{a}_0 = \bar{y}_1,
\]
\[
\hat{a}_1 = \frac{\sum xy_1}{\sum x^2} - \rho \left( \frac{\sum xy_2}{\sum x^2} \right),
\]
\[
\hat{b}_0 = \bar{y}_2
\]
and
\[
\text{var} \begin{bmatrix}
\hat{b}_0 \\
\hat{a}_0 \\
\hat{a}_1
\end{bmatrix} = \begin{bmatrix}
1/n & \rho/n & 0 \\
\rho/n & 1/n & 0 \\
0 & 0 & (1-\rho^2)/\sum x^2
\end{bmatrix} \sigma^2.
\]

Consider Figure 6.1, where \( \hat{y}_2 = \hat{b}_0 \) and \( \hat{y}_1 = \hat{a}_0 + \hat{a}_1 x \) are the best unbiased prediction lines for \( y_2 \) and \( y_1 \), \( \hat{y}_2 = \hat{b}_0 + \hat{b}_1 x \) is an unbiased prediction line for \( y_2 \) which is obtained by regressing \( y_2 \) on \( x \), and \( \hat{y}_1 = \hat{a}_0 + \hat{a}_1 x \) is the unbiased prediction line for \( y_1 \) determined by least squares independently of \( y_2 \).

Clearly the inherent (and measurement) error will induce a positive or negative slope, say, \( \hat{b}_1 = \frac{\sum xy_2}{\sum x^2} \), if \( y_2 \) is regressed on \( x \). But if, as in Figure 6.1, \( \hat{b}_1 > 0 \), then it necessarily follows that \( \hat{a}_1 \) is too large (or that \( \hat{y}_1 = \hat{a}_0 + \hat{a}_1 x \)

is overly steep) if \( \rho > 0 \). Consequently, estimation through weighted least squares accounts for the positive correlation and corrects \( \hat{a}_1 \) by an amount minus \( \rho \tilde{b}_1 \), making \( \hat{a}_1 = \hat{a}_1 - \rho \tilde{b}_1 \) the best unbiased estimate of \( b_1 \) under the given system of models.

\[
\hat{y}_1 = \hat{a}_0 + \hat{a}_1 x
\]

\[
\tilde{y}_1 = \tilde{b}_0 + \tilde{b}_1 x
\]

\[
\hat{y}_2 = \hat{b}_0
\]

Figure 6.1 Unbiased and best unbiased prediction lines

**Example 6.2** Likewise for the models,

\[
y_1 = a_0 + \sum_{i=1}^{P} a_i x_i + \xi_1
\]

\[
y_2 = b_0 + \xi_2
\]

under the same assumptions following the models of example 6.1, it is seen that the joint least squares estimates of the \( a_i \) are
\[
\begin{bmatrix}
\hat{a}_1 \\
\vdots \\
\hat{a}_p
\end{bmatrix}
= (X'X)^{-1}X'y_1 - \rho(X'X)^{-1}y_2 = (X'X)^{-1}X'(y_1 - \rho y_2).
\]

Clearly, the separate or unweighted least squares estimates of the \(a_i\), given by \(\hat{a}_i = (X'X)^{-1}X'y_1\), are corrected by an amount minus \(\rho(X'X)^{-1}y_2\) to account for the inherent (and measurement) error which makes \(\hat{b}_i = (X'X)^{-1}X'y_2\) a nonzero vector.

**Example 6.3** The models,

\[
y_1 = a_0 + b_1 x_1 + a_2 x_2 + \xi_1 \\
y_2 = b_0 + b_1 x_1 + \xi_2,
\]

under the same assumptions as the previous examples, also have a direct analogy to the other arguments.

Suppose that \(x_1\) and \(x_2\) are functionally independent and that \(\sum x_1 x_2 = 0\). Then under the given system of models, at every point on the \(x_2\) axis, \(E(y_2)\) retains the same straight line relationship in the direction of the \(x_1\) axis. But due to sampling error, if \(y_2\) is regressed on both \(x_1\) and \(x_2\), \(\hat{b}_2 = \sum x_2 y_2 / \sum x_2^2\) will be nonzero.

Since \(\rho \neq 0\), the weighted least squares estimate, \(\hat{a}_2\) of \(a_2\), takes into account the nonzero \(\hat{b}_2\) and corrects \(\hat{a}_2 = \sum x_2 y_1 / \sum x_2^2\) by an amount minus \(\rho \hat{b}_2\) so that \(\hat{a}_2 = \hat{a}_2 - \rho \hat{b}_2\).

**Example 6.4** Consider now the models,

\[
y_1 = a_0 + \sum_{i=1}^{p} a_i x_i + \xi_1
\]
\[ y_2 = b_0 + \sum_{i=1}^{p} b_i x_i + \epsilon_2, \]

under the same assumptions as previously. In this case, it has been seen by (6.22) that the correlations do not enter in the estimation of the coefficients (a known correlation offers no information in the estimation of coefficients corresponding to response types based on precisely the same set of independent variables). From the given examples, it was seen that the correlation was used in adjusting one set of coefficient estimates when a type of constraint (viz., that certain coefficients be zero) was placed on the other set of coefficients. That is, under previous models, the expectation of the second prediction line was to remain constant along the added dimensions of the first prediction line. Clearly, sampling variation prevents this from being true if we pay no attention to the restriction, so that weighted least squares correct the coefficients corresponding to the added dimensionality of the first prediction line. However, in the above models both lines have the same dimensionality, being based on precisely the same x's. Clearly, there is no possible way for a known correlation to enter in the estimation of coefficients, since adjustments are impossible. Hence a knowledge of the correlations in this case offers irrelevant information in the estimation of the coefficients.

Now, if \( X_1 \cap X_2 \), let the notation, \( X_1 \subset X_2 \), imply that the matrix \( X_1 \) is contained in the matrix \( X_2 \) so that \( X_2 \) can be augmented into parts, \( X_2 = (X_1 | \tilde{X}_2) \), where \( \tilde{X}_2 \) is that part of \( X_1 \)
not contained in \(X_2\). For example, if \(X_1 = (\hat{x}_1, \hat{x}_2)\) and \\
\(X_2 = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)\), then \(X_1 \subset X_2\), since the columns vectors of \(X_1\) are the first two column vectors of \(X_2\); i.e., \(X_2 = (X_1 | \hat{x}_2)\), where \\
\(\hat{x}_2 = (\hat{x}_3, \hat{x}_4)\). Considering the examples 6.1 through 6.4, the following theorem becomes apparent.

**Theorem:** If \(y : \left[ b_0 + \overset{o}{X_1} b, (I \cdot W) \sigma^2 \right] \), \(W\) is known, \\
\(X_1 \neq X_2 \neq \cdots \neq X_j \neq \cdots \neq X_q\), and \\
\[
X_1 \subset X_2 \subset \cdots \subset X_j \subset \cdots \subset X_q,
\]
(6.23)

then \(y_j\) enters in the joint least squares estimation of some or all of the elements of the vectors, \(b_j, b_{j+1}, \ldots, b_q\), but does not enter in the joint least squares estimation of \\
b_{j-1}, b_{j-2}, \ldots, b_1.

**Proof:** Let \(X_q = X\) in (6.13). Transform the vector \(x_i^*\) of \(X\) to \\
x_i^* such that \\
\[
\sum_{k=1}^{n} x_{ik} = \sum_{k=1}^{n} x_{ik} x_{i^* k} = 0 \quad \text{for } i \neq i^* = 1, 2, \ldots, p. \quad (6.24)
\]

Note that this is directly analogous to determination of orthogonal polynomials. Then the model (6.14) is rewritten as \\
\(y_j = b_{j0} + x_j^* b_j^* + \varepsilon_j\). For example, if \(X\) is of order \(nx5\) and a typical row of \(X\) is \((x_{1k}, x_{1k}^2, x_{2k}, x_{2k}^2, x_{1k} x_{2k})\), then \\
x_{1k} = x_{1k} + c_1, \text{ and } c_1 = 0 \text{ is obtained from} \\
\[
\sum_{k=1}^{n} x_{1k} = \sum_{k=1}^{n} (x_{1k} + c_1) = 0.
\]
Further, $x_{2k}^* = x_{1k}^2 + c_2 x_{1k}^* + c_3$, where $c_2$ and $c_3$ are obtained from

$$\sum_{k=1}^{n} x_{2k}^* = \sum_{k=1}^{n} x_{2k}^* x_{1k}^* = 0;$$

$$x_{3k}^* = x_{2k}^2 + c_4 x_{2k}^* + c_5 x_{1k}^* + c_6,$$

where $c_4$, $c_5$, and $c_6$ are obtained from

$$\sum_{k=1}^{n} x_{3k}^* = \sum_{k=1}^{n} x_{3k}^* x_{2k}^* = \sum_{k=1}^{n} x_{3k}^* x_{1k}^* = 0;$$

$$x_{4k}^* = x_{2k}^2 + c_7 x_{3k}^* + c_8 x_{2k}^* + c_9 x_{1k}^* + c_{10},$$

where $c_7$ through $c_{10}$ are obtained from

$$\sum_{k=1}^{n} x_{4k}^* = \sum_{k=1}^{n} x_{4k}^* x_{3k}^* = \sum_{k=1}^{n} x_{4k}^* x_{2k}^* = \sum_{k=1}^{n} x_{4k}^* x_{1k}^* = 0;$$

and $x_{5k}^* = x_{1k} x_{2k} + c_{11} x_{4k}^* + c_{12} x_{3k}^* + c_{13} x_{2k}^* + c_{14} x_{1k}^* + c_{15}$, where $c_{11}$ through $c_{15}$ are obtained from

$$\sum_{k=1}^{n} x_{5k}^* = \sum_{k=1}^{n} x_{5k}^* x_{4k}^* = \ldots = \sum_{k=1}^{n} x_{5k}^* x_{1k}^* = 0.$$

Now, from (6.20),

$$\hat{b}^* = \begin{bmatrix} x_1^* x_{w11}, x_1^* x_{w12}, \ldots, x_1^* x_{w1q} \\ x_2^* x_{w12}, x_2^* x_{w22}, \ldots, x_2^* x_{w2q} \\ \vdots \\ x_q^* x_{w1q}, x_q^* x_{w2q}, \ldots, x_q^* x_{wqq} \end{bmatrix}^{-1} \begin{bmatrix} x_1^* (\sum_{j=1}^{q} w_{jy_j}) \\ \vdots \\ x_q^* (\sum_{j=1}^{q} w_{jy_j}) \end{bmatrix}$$

(6.25)
We will refer to the inverse matrix on the right-hand side of (6.25) as having \( q \) pseudo rows and \( q \) pseudo columns. But under (6.23), \( x^*_j \) may be written as the partitioned matrix,

\[
x^*_j = (x^*_1 | x^*_2 | x^*_3 | \cdots | x^*_j)
\]

for example, \( x_3 = (x^*_1 | x^*_2 | x^*_3) \) where \( x^*_2 \) is that part of \( x^*_2 \) not contained in \( x^*_1 \), and \( x^*_3 \) is that part of \( x^*_3 \) not contained in \( x^*_1 \) and \( x^*_2 \). Correspondingly,

\[
\hat{b}^{*'}_j = \begin{bmatrix} \hat{b}^{*'}_j(1), \hat{b}^{*'}_j(2), \cdots, \hat{b}^{*'}_j(j) \end{bmatrix}
\]

Hence (6.25) becomes

\[
\begin{bmatrix}
\hat{b}^*_1 \\
\hat{b}^*_2(1) \\
\hat{b}^*_2(2) \\
\vdots \\
\hat{b}^*_q(1) \\
\hat{b}^*_q(2) \\
\vdots \\
\hat{b}^*_q(q)
\end{bmatrix} = \begin{bmatrix}
x^{*'}_1 (\sum_j w^{1j} y_j) \\
x^{*'}_1 (\sum_j w^{2j} y_j) \\
x^{*'}_2 (\sum_j w^{2j} y_j) \\
\vdots \\
x^{*'}_q (\sum_j w^{qj} y_j)
\end{bmatrix} \times \begin{bmatrix}
x^{*'}_1 \\
x^{*'}_1 \\
x^{*'}_2 \\
\vdots \\
x^{*'}_q
\end{bmatrix}
\]

(6.26)
where the vector, \( \hat{b}^* \), which denotes the left-hand side of (6.26) has \((q/2)(q+1)\) pseudo elements and the inverse matrix on the right-hand side of (6.26) has the same number of pseudo rows and columns.

Now, from (6.24), \( \hat{x}_j^* \hat{x}_{j'} = 0 \) for \( j \neq j' \). Further, it is easily seen that if the elements of \( \hat{b}^* \) are rearranged in the following order,

\[
\hat{b}_1^* = \begin{bmatrix}
\hat{b}^*_{2}, \hat{b}^*_{3}, \ldots, \hat{b}^*_{q(1)}, \hat{b}^*_{2(1)}, \ldots, \hat{b}^*_{2(q)}
\end{bmatrix},
\]

then (6.26) becomes

\[
\begin{bmatrix}
\hat{x}^*_1, \hat{x}^*_2, \ldots, \hat{x}^*_q
\end{bmatrix} \cdot \begin{bmatrix}
W_{11} & \ldots & W_{1q}
\end{bmatrix}^{-1} = \begin{bmatrix}
y_1
\end{bmatrix}
\]

times

\[
\begin{bmatrix}
\hat{x}^*_1, \hat{x}^*_2, \ldots, \hat{x}^*_q
\end{bmatrix} \cdot \begin{bmatrix}
W_{21} & \ldots & W_{2q}
\end{bmatrix}^{-1} = \begin{bmatrix}
y_j
\end{bmatrix}
\]

where \( W_{jj} \) of order \((q-j)^2\) is \( W^{-1} \) with the first \( j \) rows and columns deleted, and \( W_{j} = (W_{j} | W_{j+1}) \) of order \((q-j)xq\) is \( W^{-1} \) with the first \( j \) rows deleted, \( j = 1, 2, \ldots, q-1 \). For example, \( W_{q-1, q-1} = W_{qq} \), the lower right-hand element of the matrix \( W^{-1} \).
Clearly, \((x_1^*, x_1^*)^{-1} x_1^* y_1 = \hat{b}_1\), which implies that \((x_1^*, x_1^*)^{-1} x_1^* y_1 = \hat{b}_1\). Further, a typical pseudo element of \(\hat{b}_1\) is

\[
\begin{bmatrix}
\hat{b}_j(j) \\
\hat{b}_{j+1}(j) \\
\vdots \\
\hat{b}_{q}(j)
\end{bmatrix}
\]

\[
= \left[ (x_j^*, x_j^*)^{-1} x_j^* \right]^{-1} \left[ (x_j^*, x_j^*)^{-1} (w_{j-1}, j-1) \right] Y
\]

\[
= \left[ (x_j^*, x_j^*)^{-1} x_j^* \right] \left[ (w_{j-1}, j-1)^{-1} (w_{j-1}, j-1) \right] Y
\]

\[
= \left( \Psi_j \right) \Psi_{j-1} Y
\]

\[
\begin{bmatrix}
\tilde{w}_{j1} \\
\tilde{w}_{j2} \\
\vdots \\
\tilde{w}_{j+\ell, 1} \\
\tilde{w}_{j+\ell, 2} \\
\vdots \\
\tilde{w}_{q1} \\
\tilde{w}_{q2} \\
\vdots \\
\tilde{w}_{q, j-1} \\
\vdots \\
\tilde{w}_{q, j+\ell}
\end{bmatrix}
\begin{bmatrix}
y_1 \\
y_{j-1} \\
y_{j+\ell} \\
y_q
\end{bmatrix}
\]

so that

\[
\hat{b}_{j+\ell}(j) = \tilde{w}_{j+\ell, 1} \Psi_j \Psi_{j-1} + \cdots + \tilde{w}_{j+\ell, j-1} \Psi_j \Psi_{j-1} + \Psi_j \Psi_{j+\ell}.
\]
Hence, for the \(j + l\)th regression line, the vector estimate, \(\hat{b}_{j+l}(j)'\), entails some or all of the \(y\)'s up to and including \(y_{j+l}\), but not \(y_{j+l+1}\). Since the same is necessarily true for \(\hat{b}_{j+l}(j)\), the theorem is proved.

**Corollary:** Under the assumptions of the previous theorem, the test estimate of \(b_1\) is \(\hat{b}_1 = (X_1'X_1)^{-1}X_1'y_1\).

**Proof:** Given during the course of the proof of the previous theorem.

Finally, from (6.20) and (6.16), it follows that

\[
\text{var}(\hat{b}) = \left[X'(I \ast W)^{-1}X\right]^{-1} \sigma^2. \tag{6.27}
\]

Also, when \(X_j = X\) for all \(j\), then (6.27) becomes

\[
\text{var}(\hat{b}) = (X'X)^{-1} \ast \Sigma. \tag{6.28}
\]

The reader is referred to Anderson \(\text{[27]}\) for an alternate derivation of (6.28) and to Williams \(\text{[41]}\) for a specific application thereof.

### 6.3 Estimation of \(\sigma^2\) and \(\Sigma\)

Consider now finding the best unbiased estimate of \(\sigma^2\) in (6.18). Let \(c_j' = (b_{j0}'b_{j1}'); c' = (c_1', c_2', \ldots, c_j', \ldots, c_q'); \hat{c}_j' = (\hat{b}_{j0}', \hat{b}_{j1}'); \hat{c}' = (\hat{c}_1', \hat{c}_2', \ldots, \hat{c}_j', \ldots, \hat{c}_q'); T_j = (1 | X_j); \) and

\[
T = \begin{bmatrix} T_1 & 0 \\ \vdots & \vdots \\ 0 & T_q \end{bmatrix}.
\]
By equating (6.14) and (6.15) for all \(j\) and solving for \(e' = (e'_1, \ldots, e'_j, \ldots, e'_q)\), we have

\[
e = T(c - \hat{c}) + \xi. \tag{6.29}
\]

Letting \(A = I * W\), then from (6.14) and (6.20) it follows that

\[
\hat{c} = c + (T' A^{-1} T)^{-1} T' A^{-1} \xi, \tag{6.30}
\]

and thence,

\[
e = [I - T(T' A^{-1} T)^{-1} T' A^{-1}] \xi. \tag{6.31}
\]

Now

\[
E[e'^t (I * W)^{-1} e] = E(e'^t A^{-1} e)
\]

\[
= E\left[e'^t [A^{-1} - A^{-1} T(T' A^{-1} T)^{-1} T' A^{-1}] \xi\right]
\]

\[
= E\left[e'^t [I - A^{-1/2} T(T' A^{-1} T)^{-1} T' A^{-1/2}] \xi\right],
\]

where, from (6.10), \(\tilde{\xi} = A^{-1/2} \xi : \mathcal{N}(0, \, I \, \sigma^2)\). Then with \(n \times n \times q \times q\) denoting trace,

\[
E(e'^t A^{-1} e) = E(\tilde{\xi}' \tilde{\xi}) - E\left(\tilde{\xi}' [A^{-1/2} T(T' A^{-1} T)^{-1} T' A^{-1/2}] \tilde{\xi}\right)
\]

\[
= nq \sigma^2 - \sigma^2 \text{tr} \left[A^{-1/2} T(T' A^{-1} T)^{-1} T' A^{-1/2}\right]
\]

\[
= nq \sigma^2 - \sigma^2 \text{tr} \frac{q}{\sum_{j=1}^{q} p_j + q} \times (\sum_{j=1}^{q} p_j + q)
\]

\[
= [nq - (\sum_{j=1}^{q} p_j + q)] \sigma^2.
\]
Hence, the best unbiased estimate of $\sigma^2$ is

$$
\sigma^2 = \frac{[e(I*W)^{-1}e]}{[nq - (\sum_{j=1}^{q} p_j + q)]}
= \frac{(y'\Omega^{-1}y - \hat{c}'T'\Omega^{-1}y)/(nq - (\sum_{j=1}^{q} p_j + q))}{(6.33)}
$$

Next we will find the best unbiased estimate of $\Sigma$ when $X_j = X$ for all $j$. By equating (6.14) and (6.15), using 
\[ \hat{c}_j = (T_j' T_j)^{-1} T_j' y_j \] and (6.14), and solving for $e_j$, the solution is

\[ e_j = T_j (c_j - \hat{c}_j) + e_j = [I - T_j (T_j' T_j)^{-1} T_j'] e_j. \quad (6.34) \]

Then

\[ E(e_j' e_j) = E[\varepsilon_j' I - T_j (T_j' T_j)^{-1} T_j' \varepsilon_j] = [n - (p+1)] \sigma_{jj}. \]

Hence the best unbiased estimate of the element $\sigma_{jj}$ of $\Sigma$ is

\[ \hat{\sigma}_{jj} = \frac{e_j' e_j}{n - (p+1)} = \frac{(y_j - T_j \hat{c}_j)'(y_j - T_j \hat{c}_j)}{n - (p+1)}, \quad (6.35) \]

where $T_j = T_j'$.  

6.4 Estimation of $\beta$ When $W$ Is Unknown

Prior to (6.18), $\Sigma$ was rewritten as $W \sigma^2$. In doing so, it was assumed that one had information leading to a known $W$.

Such would be the case if, say, the simple correlations, $\rho_{jj}$, and the ratios of the diagonal elements of $\Sigma$ were known.

However, such knowledge is rarely at hand in practice.
In this connection, with $X_j \neq X$, $\rho_{jj} \neq 0$, and no prior knowledge of $W$, it is the intent to estimate $b$ through an iterative scheme where $W$ is estimated from the data on hand.

If we estimate (by least squares) each $c_j = (b_{j0}, b_j')$ separately by, say,

$$\hat{c}_j(0) = \begin{bmatrix} \hat{b}_{j0} \\ \hat{b}_j \end{bmatrix}$$

(note that the $\hat{b}_{j0}$ are not influenced by the $\rho_{jj}$) when $X_j \neq X$ and $\rho_{jj} \neq 0$, then one finds analogously to (6.31) that

$$e_j = [I - T_j(T_j'T_j)^{-1}T_j'] \xi_j = (I - U_j)\xi_j \text{ (say), (6.36)}$$

where $U_j(n \times n)$ has rank $p_j + 1$. For the expectation of $e_j'e_j'$, we have

$$E(e_j'e_j') = E[\xi_j'(I - U_j)'(I - U_j')\xi_j']$$

$$= E[\xi_j'I\xi_j, - \xi_j'U_j'\xi_j, - \xi_j'U_j'U_j, \xi_j, + \xi_j'U_j'U_j'U_j, \xi_j'] \quad (6.37)$$

$$= \sigma_{jj}, n - \sigma_{jj}, \text{ tr } U_j' - \sigma_{jj}, \text{ tr } U_j, + \sigma_{jj}, \text{ tr } (U_j'U_j').$$

But

$$\text{tr } U_j = \text{tr } [T_j(T_j'T_j)^{-1}T_j'] = \text{tr } [T_j'T_j(T_j'T_j)^{-1}] = p_j + 1,$$

and

$$\text{tr } U_j' = p_j + 1.$$
Also, if \( T_j \subset T_j' \), then \( p_j < p_j' \),

\[
T_j' = \begin{pmatrix} T_j & \tilde{T}_j' \end{pmatrix},
\]

\[
nx(p_j, l) \quad nx(p_j, l + 1) \quad nx(p_j, -p_j)
\]

and

\[
\text{tr}(U_j'U_j) = \text{tr} \left[ T_j (T_j'T_j)^{-1} T_j'/T_j', (T_j', T_j')^{-1} T_j' \right]
\]

\[
= \text{tr} \left[ \begin{bmatrix} T_j (T_j'T_j)^{-1} T_j' & T_j'T_j' \end{bmatrix} \begin{bmatrix} T_j'T_j' & T_j' \end{bmatrix}^{-1} T_j' \right]
\]

\[
= \text{tr} \left[ \begin{bmatrix} T_j'T_j' & T_j' \end{bmatrix} \begin{bmatrix} T_j'T_j' & T_j' \end{bmatrix}^{-1} \right]
\]

\[
= \begin{bmatrix} 1 & (p_j + l)^2 \end{bmatrix}
\]

\[
(p_j + l) \cdot (p_j, -p_j) \quad (p_j, -p_j)^2
\]

\[
\text{tr} \left[ \frac{J_{jj}'}{J_{jj}} \right] = p_j + 1,
\]

where

\[ J_{jj}' = \begin{bmatrix} \tilde{T}_j' & T_j' - \tilde{T}_j' & T_j (T_j'T_j)^{-1} T_j' \tilde{T}_j', (\tilde{T}_j', \tilde{T}_j')^{-1} T_j' \tilde{T}_j' T_j \end{bmatrix} \]

\[ \text{times} \begin{bmatrix} T_j'T_j' & T_j' \tilde{T}_j', (T_j, \tilde{T}_j')^{-1} \tilde{T}_j', T_j \end{bmatrix}.
\]

Hence, if \( T_j \subset T_j' \), then

\[
E(e_j'\tilde{e}_j') = (n - p_j, -1)\sigma_{jj}, \quad (6.38)
\]

so that an unbiased estimate of \( \sigma_{jj} \), is
\[ \hat{\sigma}_{jj'} = \frac{e_j'e_j'}{n - \text{p}_m - 1} \text{ where } \text{p}_m = \max(\text{p}_j, \text{p}_{j'}) \]

\[ = \frac{e_j'e_j'}{n - \text{p}_j - 1} \text{ if } X_j = X_{j'}, \text{ or } j = j'. \]

That is, to obtain an unbiased estimate of \( \sigma_{jj'} \), when \( X_j \neq X_{j'} \), \( \rho_{jj'} \neq 0 \), and \( X_j \subset X_{j'} \), divide the sum of the residual cross products by \( n \) minus the larger number of coefficients occurring in the \( j^{th} \) and \( j'^{th} \) regression lines.

In general, we have

\[ \hat{\sigma}_{jj'} = \frac{e_j'e_j'}{n - \text{p}_j - \text{p}_{j'} - 2 + \text{tr}(U_jU_{j'})}. \]

The weight matrix may then be estimated by

\[ c(\hat{\sigma}_{jj'}) = \hat{W} \text{ (say)} \] (6.39)

for \( j, j' = 1, 2, \ldots, q \) where \( c \) is an arbitrary constant.

Now, for the initial estimate of \( W \), say \( \hat{W}(0) \), we may use \( \hat{c}(0) \) to obtain the residual sum of squares and cross products; i.e.,

\[ e_j'e_{j'} = (y_j - T_j\hat{c}(0))'(y_{j'} - T_{j'}\hat{c}(0)). \]

By substituting \( \hat{W}(0) \) for \( W \) in

\[ \hat{c} = [T'(I*W)^{-1}T]^{-1}T'(I*W)^{-1}y, \] (6.40)

one may proceed to estimate \( c \) by, say, \( \hat{c}(1) \), which is an improved estimate over \( \hat{c}(0) \); improved in the sense that the correlations are now being accounted for. Now use the new
estimate, \( \hat{\sigma}^{(1)} \), to obtain new residual sums of squares and cross products and hence a new weight matrix, say, \( \hat{W}^{(1)} \). Substituting \( \hat{W}^{(1)} \) into (6.40), we obtain another estimate of \( \hat{\sigma} \), say, \( \hat{\sigma}^{(2)} \). This process is continued until \( \hat{W}^{(N)} \) and \( \hat{W}^{(N+1)} \) (and consequently \( \hat{\sigma}^{(N+1)} \) and \( \hat{\sigma}^{(N+2)} \)) are identical up to a desired decimal place, which assumes that convergence takes place at some stage of the iteration cycle.

To illustrate another iterative scheme, suppose we write the normal equations leading to (6.40) as

\[
T_j', \sum_{j=1}^{q} w_j'^{j-1} y_j = \sum_{j=1}^{q} w_j'^{j-1} T_j', T_j \hat{\sigma}^2, \quad T_j = (1 | x_j), \tag{6.41}
\]

for \( j' = 1, 2, \ldots, q \). Further, let \( W_{jj} \) be that matrix including the first \( j \) rows and columns of \( W \) and \( W_{jj}^{-1} = (w_{jj}^{<j}) \).

Now assume that the \( T_j \) of equations (6.41) obey the hypothesis of the theorem in Section 6.2. Then we know that \( y_j \) enters in the least squares estimation of some of the elements of \( \hat{c}_j, \hat{c}_{j+1}, \ldots, \hat{c}_q \), but does not enter in the estimation of \( \hat{c}_{j-1}, \ldots, \hat{c}_1 \). But this implies that \( c_1 \) can be estimated independently of \( y_2, \ldots, y_q, c_2 \) independently of \( y_3, \ldots, y_q, c_3 \) independently of \( y_4, \ldots, y_q \), and so forth. Consequently, the \( c_j \) of (6.41) are attainable in a stepwise fashion as follows.
\[ T'_1 T'_2 \hat{c}_2 = T'_1 y_2 \]  
(6.43)

\[ \tilde{T}'_2 T'_2 \hat{c}_2 = \tilde{T}'_2 [(w^{21}/w^{22})e_1 + y_2] \]  
(6.44)

\[ T'_1 T'_3 \hat{c}_3 = T'_1 y_3 \]  
(6.45)

\[ \tilde{T}'_3 T'_3 \hat{c}_3 = \tilde{T}'_3 [(w^{31}/w^{33})e_1 + (w^{32}/w^{33})e_2 + y_3] \]  
(6.46)

\[ \tilde{T}'_3 T'_3 \hat{c}_3 = \tilde{T}'_3 [(w^{31}/w^{33})e_1 + (w^{32}/w^{33})e_2 + y_3] \]  
(6.47)

\[ T'_1 T'_q \hat{c}_q = T'_1 y_q \]

\[ \tilde{T}'_2 T'_q \hat{c}_q = \tilde{T}'_2 \left[ \sum_{j=1}^{q-1} (w^{2j}/w^{2q})e_j + y_q \right] \]

\[ \tilde{T}'_q T'_q \hat{c}_q = \tilde{T}'_q \left[ \sum_{j=1}^{q} (w^{qj}/w^{qq})e_j + y_q \right], \]

where (6.42) is obtained by setting \( q = 1 \) in (6.41); (6.43) and (6.44) are obtained by setting \( q = 2 \) in (6.41); (6.45), (6.46), and (6.47) are obtained by setting \( q = 3 \) in (6.41), and so on.

Note that with \( W \) unknown, the best estimate \( \hat{c}_1 \) of \( c_1 \) is still available. Then, fixing \( \hat{c}_1 \), we may converge (assuming convergence at the \( N_2^{th} \) iterate) to an estimate \( \hat{c}_2^{(N_2)} \) of \( c_2 \). Next, fixing \( \hat{c}_1 \) and \( \hat{c}_2^{(N_2)} \), we converge to an estimate \( \hat{c}_3^{(N_3)} \) of \( c_3 \), and so forth.
Though this stepwise convergence to estimates of \( c_j \) entails considerably less work than converging to estimates of the \( c_j \) jointly, it should be noted that in the stepwise convergence the variability associated with each progressive \( \hat{c}_j^{(N_j)} \) necessarily increases since it includes the variability of the preceding \( \hat{c}_1^{(N_2)}, \ldots, \hat{c}_{j-1}^{(N_j)} \) which are fixed in determining \( \hat{c}_j^{(N_j)} \). However, if \( q \) is small, the errors of the stepwise estimates will, most likely, be of approximately the same magnitude as the joint estimates.

6.5 A Sufficient Condition for Convergence and an Approximate Variance of the Iterated Estimate

Consider now the question of convergence when using the first iterative scheme of Section 6.4 to obtain an estimate of \( \hat{c} \).

Let \( \hat{c} \) be the best estimate of \( c; \) i.e., if \( \hat{c}^* \) is any other estimate of \( c_\alpha \), then \( \text{var} \hat{c}^* \geq \text{var} \hat{c}_\alpha \). Clearly all the elements of \( \hat{c} \) are attainable if \( W \) is known; i.e.,

\[
\hat{c} = (T'(I*W)^{-1}T)^{-1}T'(I*W)^{-1}y
\]

\[
= \hat{f}(T,y;W) \quad \text{(say)}
\]

where

\[
\hat{c}' = \hat{f}'(T,y;W) = \left[ \hat{f}_1(T,y;W), \ldots, \hat{f}_\alpha(T,y;W), \ldots, \hat{f}_q \sum_{j=1}^{p_j+q} (T,y;W) \right].
\]

Further, since \( W \) does not enter in the joint least squares estimation of the \( b_{j0} \), disregard the latter so that
\[
\hat{b} = \begin{bmatrix}
\hat{b}_1 \\
\vdots \\
\hat{b}_j \\
\vdots \\
\hat{b}_q
\end{bmatrix} = \frac{f}{\sum_{j=1}^{q} p_j x_1} (\hat{x}, \hat{y}; \hat{w}). \tag{6.48}
\]

It was seen that

\[
\hat{b}^{(1)} = f(\hat{w}^{(0)}; \hat{b}_0, \hat{x}, \hat{y}) = h(\hat{b}^{(0)}; \hat{b}_0, \hat{x}, \hat{y}), \tag{6.49}
\]

where one method of obtaining \( \hat{b}^{(0)} \) is through the separate least squares estimation of each \( b_j \). Then (6.49) defines a general recurrence relationship,

\[
\hat{b}^{(N+1)} = h(\hat{b}^{(N)}) \tag{6.50}
\]

or

\[
\hat{b}^{(N+1)} = h(\hat{b}^{(N)}),
\]

which is used in solving the \( \sum_{j=1}^{q} p_j \) nonlinear equations

\[
\hat{b} = h(\hat{b}) \tag{6.51}
\]

in the same number of unknowns. The superscript, \( (N) \), in (6.50) denotes the \( N+1 \)st iterate and for convenience, \( \hat{x}, \hat{b}_0, \hat{y} \) are not written in the functional forms, since they are held constant throughout the iteration process.

Let

\[
\hat{b}^{(N)} = \hat{b}^* + \hat{\theta}^{(N)}, \tag{6.52}
\]
where \( \hat{b}^* \) is a solution of (6.51) and \( \hat{\delta}^{(N)} \) is the error of the \( N+1 \)st iterate. Substituting (6.52) into (6.50) yields

\[
\hat{b}^{(N+1)} = h(\hat{b}^* + \hat{\delta}^{(N)})
\]

or

\[
\hat{b}^{(N+1)} = h_\alpha(\hat{b}^* + \hat{\delta}^{(N)}).
\] (6.53)

Expanding the right-hand side of (6.53) in a Taylor's series about \( \hat{b}^* \), we have

\[
\hat{b}^{(N+1)} = h_\alpha(\hat{b}^*) + \left[ \left( \frac{\partial h_\alpha}{\partial \hat{b}} \right) \bigg|_{\hat{b} = \hat{b}^*} \right] \hat{\delta}^{(N)} + O(\hat{\delta}^2)
\] (6.54)

or generally,

\[
\hat{b}^{(N+1)} = h(\hat{b}^*) + H\hat{\delta}^{(N)} + O(\hat{\delta}^2),
\] (6.55)

where

\[
H = \left[ \left( \frac{\partial h_\alpha}{\partial \hat{b}} \right) \bigg|_{\hat{b} = \hat{b}^*} \right], \quad \alpha = 1, 2, \ldots, \sum_{j=1}^{q} p_j.
\] (6.56)

Using (6.52) along with \( \hat{b}^* = h(\hat{b}^*) \), i.e., substituting the solution \( \hat{b}^* \) into (6.51), then (6.55) takes the form,

\[
\hat{\delta}^{(N+1)} = H\hat{\delta}^{(N)} + O(\hat{\delta}^2),
\]

or, neglecting \( O(\hat{\delta}^2) \), \( \hat{\delta}^{(N)} = H\hat{\delta}^{(N)} \), where the \( \hat{\theta} \)'s are the characteristic roots of the matrix \( H \), and \( H \) is the spectral decomposition of \( H \) (and is determined from \( \hat{b}^{(0)} \)).

Clearly, if the \( \hat{\theta} \) largest in modulus is less than unity, then \( \hat{\delta}^{(N)} \to 0 \) as \( N \to \infty \). This, then, is a sufficient condition for the iteration process to converge to a solution \( \hat{b}^* \), assuming that \( O(\hat{\delta}^2) \) may be neglected. (The preceding argument is taken from reference [37].)

Now consider the approximate variance of such an estimate. Let (6.52) be rewritten in the form,
\[ \hat{b}(N) = \hat{b} + \hat{\delta}(N) = (b + \hat{d}) + \bar{\delta}(N), \] (6.57)

where \( \bar{\delta}(N) \) is the error distance of the \( N+1 \)st iterate from the best estimate, \( \hat{b} \), and \( \bar{d} \) is analogous to \( \bar{c} - \hat{c} \) in (6.30). Substituting the middle term of (6.57) into (6.50) results in

\[ \hat{b}(N+1) = \bar{h}(\hat{b} + \bar{\delta}(N)) \] (6.58)

or

\[ \hat{b}(N+1) = h_{\alpha}(\hat{b} + \bar{\delta}(N)). \] (6.59)

Expanding the right-hand side of (6.59) in a Taylor's series about \( \hat{b} \) yields

\[ \hat{b}(N+1) = h_{\alpha}(\hat{b}) + \left[ (\partial h_{\alpha}/\partial \hat{b})_{\hat{b}=\hat{b}} \right] \bar{\delta}(N) + O(\bar{\delta}^2) \] (6.60)

Further, since \( \hat{b} = b + \bar{d} \),

\[ h_{\alpha}(\hat{b}) = h_{\alpha}(b + \bar{d}) = h_{\alpha}(b) + \left[ (\partial h_{\alpha}/\partial \hat{b})_{\hat{b}=\hat{b}} \right] \bar{d} + O(\bar{d}^2), \] (6.61)

where the latter follows by expanding the middle term of (6.61) in a Taylor's series about \( b \). Substituting the expansion of \( h_{\alpha}(\hat{b}) \) in (6.61) into (6.60) and using (6.57), we have

\[ \hat{\delta}(N+1) = h_{\alpha}(b) - \hat{b} + \left[ (\partial h_{\alpha}/\partial \hat{b})_{\hat{b}=\hat{b}} \right] \bar{\delta}(N) \]

\[ + \left[ (\partial h_{\alpha}/\partial \hat{d})_{\hat{b}=\hat{b}} \right] \bar{d} + O(\bar{\delta}^2) + O(\bar{d}^2). \] (6.62)

Neglecting \( O(\bar{\delta}^2) \) and \( O(\bar{d}^2) \), the result in general is

\[ \hat{\delta}(N+1) = h(b) - \bar{b} + G_1 \bar{\delta}(N) + G_2 \bar{d}, \] (6.63)
where $G_1 = \left[ (\partial h / \partial b) \big| b = \hat{b} \right]$ and $G_2 = \left[ (\partial h / \partial \hat{b}) \big| b = b \right]$ for $\alpha = 1, 2, \ldots, \sum_{j=1}^{q_i} p_j$.

Note that if $W$ is known, it should follow that $\hat{\phi}^{(N)} = 0$ for all $N$. This is seen by noting that a known $W$ is not a function of $\hat{b}$ or $\hat{\phi}^{(N)}$ so that $G_1 = \left[ (\partial h / \partial W) (\partial \hat{b} / \partial b) \big| b = \hat{b} \right] = 0$, and $G_2 = \left[ (\partial h / \partial \hat{b}) (\partial W / \partial b) \big| b = b \right] = 0$, since $(\partial W / \partial \hat{b}) = (\partial W / \partial \hat{b}) = 0$. Further, with $W$ known, $h(b)$ becomes $f(X,Y;W)$ in (6.48), all of which implies that $\hat{\phi}^{(N)} = 0$. The same can be said for particular $b_\alpha$ for which best estimates are attainable independently of $W$.

Now, (6.63) can be written in the form

$$\hat{\phi}^{(N)} = h(b) - \hat{b} + \hat{\phi}^{(N)} G_1 + G_2 \xi,$$  \hfill (6.64)

where the $\hat{\theta}$'s are the characteristic roots of $G_1$, and $G_1$ is the spectral decomposition of $G_1$. If the largest $\hat{\theta}$ in modulus is less than unity, then $\hat{\phi}^{(N)} \to h(b) - \hat{b} + G_2 \xi$ for $N$ sufficiently large. Further, as $n$ increases, $h(b)$ rapidly approaches $\hat{b}$. (The latter statement is based on the following intuitive argument. If $W = [\text{SSE}_{jj},/(df)_{jj}']$, where df implies degrees of freedom, then when $b$ is known, $(df)_{jj} = n - 1$ rather than $n - p_m - 1$, $p_m = \max(p_j, p_j')$. Clearly, as $n$ increases (say $n \geq 20$), then $W$ should be very close to $W$ so that the difference between $h(b)$ and $\hat{b}$ is negligible.) Then assuming convergence, (which presupposes $0(\hat{\phi}^2), 0(\xi^2)$ negligible) we have

$$\hat{\phi}^{(N)} \sim G_2 \xi,$$  \hfill (6.65)

and by substituting (6.65) into (6.57),
\[
\hat{b}^* = \hat{b} + \delta(N) \approx \hat{b} + G_2X = b + (I+G_2)X.
\]

Since \(E(\delta) = 0, E(\hat{b}^*) = b\) and

\[
\text{var} \hat{b}^* \approx (I+G_2)[X'(I*W)^{-1}X]^{-1}(I+G_2)\sigma^2.
\]  \hspace{1cm} (6.66)

Assuming convergence, it is proposed that

\[
\hat{b}^* \sim N(b, \text{var} \hat{b}^*),
\]  \hspace{1cm} (6.67)

where \(\text{var} \hat{b}^*\) is estimated by substituting \(\hat{b}^*\) for \(b\) and \(\hat{W}(N-1) = \hat{W}^*\) for \(W\) (with convergence occurring at the \(N^{th}\) iterate) in (6.66).

It should be noted that we have neglected in \(\text{var} (\hat{b}^*)\) the variability of \(\hat{b}_0\) which arises in repeated sampling. As such, \(\text{var} (\hat{b}^*)\) in (6.66) will tend to underestimate the true small sample variance of the iterated estimate \(\hat{b}^*\).

6.6 General Comments on Estimation in the Predictive Regression System

This discussion pertains to cases where the \(y_j\) \((j=1,2,\ldots,q)\) depend on different sets of overlapping independent variables with no \textit{a priori} information regarding \(\Sigma\).

If the sole intent is the individual prediction of each \(y_j\) for a given \(x_j\), then separate least squares may be applied in estimating the parametric coefficients and diagonal elements of \(\Sigma\). But for analyses requiring the use of the off-diagonal elements of \(\Sigma\), separate least squares neglecting the correlations may yield misleading results. For example, the \(q\) dimensional confidence ellipse for the average value of \(y\),
\[ A_0 = \begin{bmatrix} x'_1 & 0 \\ x'_2 & \vdots \\ 0 & x'_q \end{bmatrix} \]

is defined by

\[ \left[ \hat{\mathbf{y}} - \mathbf{E}(\hat{\mathbf{y}}) \right]' \left( \frac{1}{n} \hat{W}^* + A_0 V A_0' \right)^{-1} \left[ \hat{\mathbf{y}} - \mathbf{E}(\hat{\mathbf{y}}) \right] / \sigma^2 \]

\[ \approx F_{\alpha} \left[ \sum_{j=1}^{q} p_j + q, nq - \left( \sum_{j=1}^{q} p_j + q \right) \right], \]  

which stems from

\[ \hat{\mathbf{y}} = \hat{b}_0 + A_0 \hat{b} \sim N(b_0 + A_0 b, (C^2/n) \hat{W}^* + A_0 V A_0' \sigma^2), \]

where \( V \) is the coefficient matrix at \( \sigma^2 \) in (6.66). With \( W \) known, \( \hat{W}^* \) is replaced by \( W \) and \( V \) by \( \left[ \mathbf{x}' (I \times W)^{-1} \mathbf{x} \right]^{-1} \). If the correlations in \( \Sigma \) are neglected in (6.68), then the ellipsoidal region defined by the latter will not be "tilted" so that values of \( \mathbf{y} \) given \( A_0 \) may, in repeated sampling, fall outside the confidence region with great regularity. The same can be said for the \( q \) dimensional tolerance ellipse defined by

\[ \left[ \hat{\mathbf{y}} - \mathbf{E}(\hat{\mathbf{y}}) \right]' \left( 1 + \frac{1}{n} \right) \hat{W}^* + A_0 V A_0' \left[ \hat{\mathbf{y}} - \mathbf{E}(\hat{\mathbf{y}}) \right] / \sigma^2 \]

\[ = F_{\alpha} \left[ q, p_j + q, nq - \left( \sum_{j=1}^{q} p_j + q \right) \right], \]

The tolerance hyperellipsoids being such that \( 1 - \alpha \) percent of the vectors \( \bar{\mathbf{y}} \) given \( A_0 \) will fall within it.
Should the simple correlations of Σ be "near zero," then the separate least squares estimation of each $b_j$ may be used. If, however, "near zero" correlations cannot be assumed, then one should go from the restricted regression system (the system in which the models include different sets of overlapping independent variables) to the unrestricted regression system; i.e., $X_j = X$ for all $j$. After regressing each response type on all the independent variables (in which case best estimates are available independently at $\Sigma = \Sigma_0^2$), the result should be that the coefficients not appearing in the restricted regression system should have near zero estimates in the unrestricted system as is apparent from examples 6.1 through 6.3. A test of significance can be made to determine whether the matrix,

$$
\begin{bmatrix}
\hat{b}_1 \\
\vdots \\
\hat{b}_j \\
\vdots \\
\hat{b}_q
\end{bmatrix}
$$

(the estimated coefficient matrix of the unrestricted regression system), differs significantly from $\hat{B}_R$, which is the matrix $\hat{B}_U$ with zeros replacing those $\hat{b}_U$'s not appearing in the restricted system. However, methods such as the likelihood ratio test used for testing $H_0: \hat{B}_U = \hat{B}_R$ (see Anderson (27)) are not sensitive to small departures from the null hypothesis whereas we are particularly interested in detecting small departures. Further, due to the impossibility of
proving $H_0$, a test of significance, particularly in this problem, leaves much to be desired with regard to deciding whether $\hat{b}_R$ is "sufficiently close" to $\hat{b}_U$. On the other hand, one may have no other choice other than to resort to a test of significance.

If $\hat{b}_R$ is "sufficiently close" to $\hat{b}_U$, then the coefficients in the models of the restricted regression system may be estimated by separate least squares and from these estimates, say, $\hat{b}^{**}_j$, the matrix $(SSE_{jj})$ may be determined. The $\hat{b}^{**}_j$ will be unbiased, consistent, and possibly most efficient under the given circumstances.

If those coefficients not appearing in the restrictive regression system have "decidedly" nonzero estimates in the unrestricted system, then one should either

(i) question the models of the restricted regression system, or

(ii) make adjustments by using some iterative scheme to estimate $b$.

With the models having strong theoretical justification, there may be no other choice than to make adjustments through iteration. Then assuming convergence, there is the formidable task of estimated (6.66), the estimated variance of an iterated vector estimate. The likely resolution will be the use of the estimated asymptotic variance of $\hat{b}^*$; namely

$$\left[\hat{X}'(I*\hat{W}*)^{-1}\hat{X}\right]^{-1}\hat{\sigma}^2,$$  \hspace{1cm} (6.69)
even though the variances of the $\hat{b}^*_i$'s may be considerably underestimated, particularly if there are many parameters to estimate through iteration (or rather, by the first iterative scheme of Section 6.4). For more simple systems, such as in example 6.1, it is seen in the forthcoming example (example 6.5) that the estimate of (6.66) is very close to (6.69). This is due possibly to the fact that iteration is required in the estimation of only one parameter, $b_1$, and that

$$\hat{a}_1 = \frac{\sum y_1 x}{\sum x^2}$$

(the least squares estimate of the coefficient not appearing in the restricted system) is not significantly different from zero. Further, it is again stressed that the variance of the iterated solution does not account for the variability contributed by the $\hat{b}_{0,1}$ in $\hat{W}^*$.

In general, if there are very few parameters which must be estimated through iteration, the estimate of (6.66) and the estimated asymptotic covariance matrix (6.69) may be close, particularly if those coefficients in the restricted regression system have nonsignificant (from zero) estimates in the unrestricted regression system. If there exist many parameters and if adjustments need be made, then one should be well aware that the variances of such estimates increase, possibly by very large amounts.

**Example 6.5.** From a bivariate normal population,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} : N \begin{bmatrix} a_0 \\ b_0 + b_1 x \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \sigma^2$$
the data were as follows:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y_1$</th>
<th>$y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>8.8</td>
<td>3.7</td>
</tr>
<tr>
<td>-3</td>
<td>9.6</td>
<td>12.5</td>
</tr>
<tr>
<td>-1</td>
<td>10.6</td>
<td>19.8</td>
</tr>
<tr>
<td>1</td>
<td>8.3</td>
<td>24.4</td>
</tr>
<tr>
<td>3</td>
<td>10.4</td>
<td>31.6</td>
</tr>
<tr>
<td>5</td>
<td>10.1</td>
<td>40.1</td>
</tr>
</tbody>
</table>

Best estimates of $a_0$ and $b_0$ were $\hat{a}_0 = \bar{y}_1 = 9.6333$ and $\hat{b}_0 = \bar{y}_2 = 22.0167$. Even though $\hat{\rho}_1 = (\sum xy_1 / \sum x^2) = (6.6/70)$ was not significantly greater than zero, an adjustment was made by means of

$$\hat{b}_1(N+1) = (\sum xy_2 / \sum x^2) - \hat{\rho}(N) (\sum xy_1 / \sum x^2),$$

where

$$\hat{\rho}(N) = \frac{\text{MSE}_{12}^{(N-1)}}{\text{MSE}_p^{(N-1)}} = \frac{\text{SSE}_{12}^{(N-1)}}{4} / (\text{SSE}_p^{(N-1)}) / 9$$

$$= \frac{9}{4} \frac{\sum (y_1 - \hat{a}_0)(y_2 - \hat{b} - \hat{b}(N-1)x)}{\sum (y_1 - \hat{a}_0)^2 + \sum (y_2 - \hat{b} - \hat{b}(N-1)x)^2},$$

since $\sigma_{11} = \sigma_{12}; \ i.e., \ \sigma_{y_1}^2 = \sigma_{y_2}^2$. Three iterations were required to converge to the third decimal place of $\hat{b}^*$ and the second decimal place of $\hat{\rho}^*$:

$$\hat{b}^{(0)} = 3.4843 \quad \hat{\rho}^{(0)} = 0.6930$$
$$\hat{b}^{(1)} = 3.4190 \quad \hat{\rho}^{(1)} = 0.7677$$
$$\hat{b}^{(2)} = 3.4119 \quad \hat{\rho}^{(2)} = 0.7729$$
$$\hat{b}^{(3)} = 3.4114 \quad \hat{\rho}^{(3)} = 0.7703.$$
Then
\[ \hat{\sigma}^2 \left\langle (N+1) \sigma_{b_1} \right\rangle = \left[ -\left( \sum xy_1 / \sum x^2 \right) \left( \sigma_{\hat{\rho}^2} (N) / \sigma_{b_1} \right) \right] \left\langle b_1 = \hat{b}_1 \right\rangle = 0.0639, \]
so that the variance of \( \hat{b}^* \) is estimated to be \( (1.0639)^2 = 1.1319 \) times the estimated asymptotic variance of \( \hat{b}^* \). The asymptotic covariance matrix of the estimates is
\[
\text{var} \begin{bmatrix} \hat{a}_0 \\ \hat{b}_0 \\ \hat{b}^* \\ \hat{\rho}^2 \\ \hat{\theta}^2 \end{bmatrix} = \begin{bmatrix} 1/n & \rho/n & 0 \\ \rho/n & 1/n & 0 \\ 0 & 0 & (1-\rho^2)/\sum x^2 \end{bmatrix} \sigma^2,
\]
which is estimated by
\[
\begin{bmatrix}
1/6 & 0.77/6 & 0 \\
0.77/6 & 1/6 & 0 \\
0 & 0 & 1-(0.77)^2/70
\end{bmatrix}
\hat{\sigma}^2 = \begin{bmatrix}
0.167 & 0.128 & 0 \\
0.128 & 0.167 & 0 \\
0 & 0 & 0.006
\end{bmatrix}
\]
where \( \hat{\sigma}^2 = 1.329 \). The estimate of the covariance matrix given by (6.66) is
\[
\begin{bmatrix}
0.167 & 0.128 & 0 \\
0.128 & 0.167 & 0 \\
0 & 0 & 0.007
\end{bmatrix}
\hat{\sigma}^2.
\]
Note that the zero's in the above matrix are fictitious, since \( \hat{b}^*_1 \) is dependent on \( \hat{a}_0 \) and \( \hat{b}_0 \); i.e., \( \hat{b}^*_1 \) depends on \( \hat{\rho}^* \) which in turn depends on \( \hat{a}_0 \) and \( \hat{b}_0 \). Also, the variance of \( \hat{b}^*_1 \) would be somewhat larger, since the variation of \( \hat{a}_0 \) and \( \hat{b}_0 \) is neglected in \( \hat{\rho}^* \).
6.7 A Note on the Variance of a Vector Coefficient Estimate Obtained Through Iteration in a Single Rational Regression Model

The initial iteration scheme of Section 6.4 is well known. See, for example, applications by Williams [41] and Turner [37]. Then assuming convergence as discussed in the previous section (though convergence is rarely discussed in statistical applications using iteration processes), the usual recourse is to use the asymptotic variance to describe the variability of the resulting estimates. In doing so, the variability of the estimated weight matrix, \( \hat{W}^* \), is neglected. Consequently, probability statements, such as those associated with confidence intervals for the parametric coefficients, may be grossly in error.

Turner [37] considered the rational regression model,

\[
y = \left[ A_p(x)/B_p(x) \right] + \epsilon, \quad (6.70)
\]

where \( A_p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_p x^p \), and \( B_q(x) = 1 + b_1 x + b_2 x^2 + \ldots + b_q x^q \). The problem is one of estimating the \( a \)'s and \( b \)'s under the assumption that the \( \epsilon \)'s are NID(0,\( \sigma^2 \)). It is readily seen that even in the simple case, \( p=0 \) and \( q=1 \), the maximum likelihood equations are intractable due to there being no systematic method of iteration. To counter this difficulty, (6.70) is put in another form; namely,

\[
y = A_p(x) - b_1 xy - \ldots - b_q x^q y + B_q(x) \epsilon. \quad (6.71)
\]
If one were to neglect the obvious defect that the "independent" variables are correlated with the error term, \( B_q(x) \hat{\epsilon} \), then estimates of the \( a \)'s and \( b \)'s are attainable (assuming convergence) through weighted least squares by using the first iterative scheme of Section 6.4; i.e.,

\[
\begin{bmatrix}
\hat{a}(N+1) \\
\hat{b}(N+1)
\end{bmatrix}
= \left[ T_0(\hat{W}_0(N))^{-1}T_0 \right]^{-1}T_0(\hat{W}_0(N))^{-1}y \\
= \hat{f}(\hat{W}_0(N); T_0, y) = \hat{h}(\hat{b}(N); T_0, y),
\]

where

\[
T_0 = \begin{bmatrix}
1, x_1, x_1^2, \ldots, x_1^p, -x_1y_1, \ldots, -x_1^py_1 \\
\vdots & \vdots & \ddots & \vdots \\
1, x_n, x_n^2, \ldots, x_n^p, -x_ny_n, \ldots, -x_n^py_n
\end{bmatrix},
\]

\[
\hat{a}(N)^\prime = (a_0(N), \ldots, a_p(N)),
\]

\[
\hat{b}(N)^\prime = (b_1(N), \ldots, b_q(N)),
\]

and \( \hat{W}_0(N)^{-1} \) is a diagonal matrix with typical element, \( (1 + \hat{b}_1(N)x_k + \ldots + \hat{b}_q(N)x_k)^{-2} \), \( k = 1, 2, \ldots, n \). The estimated asymptotic covariance matrix of the coefficient estimates, \( \hat{a}^*, \hat{b}^* \), is

\[
\text{var} \begin{bmatrix}
\hat{a}^* \\
\hat{b}^*
\end{bmatrix} = (T_0^\prime \hat{W}_0^{-1}T_0)^{-1} \sigma^2,
\]

where \( \sigma^2 = \hat{\epsilon}^\prime \hat{W}_0^{-1} \hat{\epsilon} / (n-p-q-1) \) and \( \hat{\epsilon} = y - T_0 \hat{b}^* \).
If we proceed as in (6.51), the recurrence relationship (6.72) is used in obtaining solutions to

\[
\begin{bmatrix}
\hat{\mathbf{a}} \\
\hat{\mathbf{b}}
\end{bmatrix} = h(\hat{\mathbf{b}}).
\]  

(6.73)

But (6.73) is a set of \(p + q + 1\) equations in \(q\) unknowns so that attempts to find the approximate variance of \(\hat{\mathbf{a}}^*\) and \(\hat{\mathbf{b}}^*\) by methods of Section 6.5 are futile.

Suppose, however, that the model (6.70) is altered so as to take the form

\[
y = \frac{A_p(x)}{B_q(x) + A_p(x)} + \epsilon
\]

under the same assumptions following (6.70). Multiplying through (6.74) by \(B_q(x) + A_p(x)\) and rearranging terms, we have

\[
y = -y \sum_{i=1}^{q} b_i x^i + (1-y) \sum_{j=0}^{p} a_j x^j + \left[ B_q(x) + A_p(x) \right] \epsilon
\]

(6.75)

so that, again, one may proceed with the weighted least squares iteration scheme in estimating the \(a's\) and \(b's\).

Instead of (6.72), we have the new vector estimate (assuming convergence)

\[
\begin{bmatrix}
\hat{\mathbf{a}}^{**} \\
\hat{\mathbf{b}}^{**}
\end{bmatrix} = \left( 0_{T_0} \hat{\mathbf{W}}^{**-1} 0_{T_0} \right)^{-1} 0_{T_0} \hat{\mathbf{W}}^{**-1} \epsilon,
\]

(6.76)
where,

$$0^T 0 = \begin{bmatrix}
1-y_1, x_1(1-y_1), \ldots, x_1^p(1-y_1), -x_1 y_1, \ldots, -x_1^q y_1 \\
\vdots \quad \vdots \quad \vdots \\
1-y_n, x_n(1-y_n), \ldots, x_n^p(1-y_n), -x_n y_n, \ldots, -x_n^q y_n
\end{bmatrix}$$

and $\hat{W}^{**}-1$ is a diagonal matrix with typical element,

$$(1 + \sum_{i=1}^{q} \hat{b}_k x_k^i + \sum_{j=0}^{p} \hat{a}_i x_k^j)^{-2}.$$ 

Now, as in (6.51), the iteration scheme is used to obtain the solutions to

$$\begin{bmatrix}
\hat{a} \\
\hat{b}
\end{bmatrix} = \hat{h}(\hat{a}, \hat{b})$$

under the model (6.74) and in precisely the same manner as in Section 6.5, it is seen that

$$\text{var} \begin{bmatrix}
\hat{a}^{**} \\
\hat{b}^{**}
\end{bmatrix} \overset{\Delta}{=} \begin{bmatrix}
I + \left( \frac{\partial \hat{h}^*}{\partial (\hat{a}, \hat{b})} \right) \begin{bmatrix}
\hat{a} \\
\hat{b}
\end{bmatrix} = \begin{bmatrix}
\hat{a} \\
\hat{b}
\end{bmatrix}
\end{bmatrix} \left[ 0^T 0 \hat{W}^{**}-1 0^T 0 \right]^{-1} \left[ I + \left( \frac{\partial \hat{h}^*}{\partial (\hat{a}, \hat{b})} \right) \begin{bmatrix}
\hat{a} \\
\hat{b}
\end{bmatrix} = \begin{bmatrix}
\hat{a} \\
\hat{b}
\end{bmatrix}
\right]^{0^2},$$

$\Delta=1,2,\ldots,p+q+1$. Then (6.77) is estimated by setting $\sigma^2 = \hat{\sigma}^2$ and evaluating $(\hat{a}, \hat{b})$ at $(\hat{a}^{**}, \hat{b}^{**})$.

By setting $p=0$ and $q=1$ in (6.68), there results the rectangular hyperbola, $y = \frac{a_0}{1+b_1 x} + \xi$, with asymptotes at $x = -(1/b_1)$ and $y = 0$. But under the model (6.68), it is difficult to measure the worth of those estimates obtained through iteration, other than by using the asymptotic covariance matrix.
If, instead, we refer to model (6.74) and set p=0 and q=1, then \[ y = \left[ \frac{a_0}{(b_0+b_1 x)} \right] + \epsilon, \] where \( b_0 = a_0 + 1 \) and the asymptotes existing at \( x = -\frac{b_0}{b_1} \) and \( y = 0 \). In this case, different estimates will, most likely, be obtained, but one will at least be able to gain a better measure of their variation in small samples.

Note that the disadvantage of using the model (6.74) instead of (6.70) is that additional parameters are introduced, thus possibly increasing the variability of the iterated estimate. Hence, while the gain in using (6.74) is possibly a more realistic covariance matrix for the estimates, the loss may be a larger variability associated with the estimates due to the additional parameters. It should also be noted that another alternative is to generate small sample estimates of the parameters by omitting single observations, pairs of observations, and so forth. In this way, estimates of the small sample dispersion for the parameter estimates are available. We have chosen another course in attempting to find an approximate expression for small sample dispersion.
7.0 SOME UNCONSTRAINED EXTREME VALUE PROBLEMS

7.1 The Estimation of Stationary Points of Linear Functions of Linear Regressions

Referring to the p columns of X in (6.13), let \( x_i = \bar{x}_i = \frac{1}{n} \sum_{k=1}^{n} o_{ik} \), where \( \bar{x}_i \) is a potentially observable value in the range \( \min_k o_{ik} \leq x_i \leq \max_k o_{ik} \) as opposed to \( o_{ik} \) which has already been observed. Let \( \mathbf{Z}'=\begin{bmatrix} Z_1, \ldots, Z_t, \ldots, Z_h \end{bmatrix} \) be the h dimensional vector \( (t=1,2,\ldots,h) \) including those \( x \)'s in \( x'=(x_1,\ldots,x_i,\ldots,x_p) \) which are functionally independent; e.g., if \( x'=(x_1,x_2,x_1^2,x_2^2) \), then \( \mathbf{Z}'=(Z_1,Z_2)=(x_1,x_2) \).

If an estimated linear function of interest, say,

\[
\hat{\phi}(\hat{y}) = \mathbf{l}' \hat{y} = \mathbf{l}' \hat{B} \mathbf{x},
\]

is differentiable where \( \mathbf{l}'=(l_1,l_2,\ldots,l_q) \) is a non-null vector of preselected constants and the \( \hat{y}_j \) taken in linear combination are in the same unit of measure, then by taking partial derivatives of \( \hat{\phi} \) with respect to the \( Z \)'s and solving the h equations,

\[
\frac{\partial \phi(\hat{y})}{\partial \mathbf{Z}} = 0
\]

simultaneously, we obtain estimates, say, \( \hat{Z} \) of those \( Z \) values for which \( \phi[E(\hat{y})] \) attains stationary values, assuming the existence of at least one stationary point.
7.2 A Confidence Region for $E(\hat{\gamma})$ When $\emptyset(\hat{\gamma})$ is a
Second-order Linear Function of the Z's

At this point and for the remainder of Chapter 7, we shall restrict $\emptyset(\hat{\gamma})$ to the role of a second-order linear function of the Z's.

With $\hat{Z}$ denoting the Z value for which $\emptyset(\hat{\gamma})$ attains its stationary value, let $E(\hat{Z})$ be the Z value for which $\emptyset[E(\hat{\gamma})]$ attains its stationary point. But $E(\hat{\gamma})$ is unknown, so that a natural consequence is to determine an h dimensional confidence region for $E(\hat{Z})$. Such a region is useful in situations where, say, the achievement of $\hat{Z}$ is difficult, making a range or region of Z values desirable.

One approach to its attainment is to define the confidence region as that region in the h dimensional Z space such that, with probability $1-\alpha$, the null hypothesis,

$$H_0 : \frac{\partial \emptyset[E(\hat{\gamma})]}{\partial Z} = 0$$

(7.3)

is not rejected. Consequently, the problem is one of finding the region of Z values which is in concordance with the null hypothesis.

If $X_j \neq X$ for some $j$, then $\frac{\partial \emptyset(\hat{\gamma})}{\partial Z}$ is of the form
\[
\frac{\partial \hat{\boldsymbol{y}}}{\partial Z} = \begin{bmatrix}
\frac{\partial \hat{y}}{\partial Z_1} \\
\vdots \\
\frac{\partial \hat{y}}{\partial Z_t} \\
\vdots \\
\frac{\partial \hat{y}}{\partial Z_h}
\end{bmatrix} = \begin{bmatrix}
\begin{bmatrix}
\tau_1^{t_1}, \tau_2^{t_2}, \tau_2^{t_2}, \ldots, \tau_q^{t_p}
\end{bmatrix} \\
\begin{bmatrix}
lp_1, lq_1, lq_2, \ldots, lq_p
\end{bmatrix}
\end{bmatrix}
\]

or

\[
\frac{\partial \hat{\boldsymbol{y}}}{\partial Z} = C \left[ \begin{array}{c}
\hat{b}_1 \\
\vdots \\
\hat{b}_q \\
\end{array} \right], \quad (7.4)
\]

where, for example, if \( q = 2 \); \( \hat{x'} = (x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2) \); \( y_1 \) is regressed on \( x' \) and \( y_2 \) on \( (x_1, x_2, x_1x_2, x_1^2, x_2^2) \); then \( Z' = (x_1, x_2, x_3); t = 1, 2, 3 = h \); \( t_{1-t_1} = t_1 (1, 0, 0, x_2, x_3, 0, 2x_1, 0, 0) \); \( t_{2-t_2} = t_2 (1, 0, x_2, 2x_1, 0, 0); \) and so forth.

Under (7.3),

\[
\frac{\partial \hat{\boldsymbol{y}}}{\partial Z} = \hat{C}_b : N(0, CV'C \sigma^2), \quad (7.5)
\]

where from (6.27), \( V = [X'(1*W)^{-1}X]^{-1} \) if \( W \) is known and for \( W \) unknown, \( V \) is the coefficient matrix of \( \sigma^2 \) given in (6.66).

If \( \sigma^2 \) is known, then

\[
\frac{\partial \hat{y}}{\partial Z} \left( CV'C \right)^{-1} \frac{\partial \hat{y}}{\partial Z} = \chi^2 \quad (7.6)
\]

is distributed as \( \chi^2 \) with \( h \) degrees of freedom under (7.3).
Since a quadratic form
\[ r'Ax^{-1}r = \begin{vmatrix} 0 & r' \\ r & A \\ A \end{vmatrix}, \quad (7.7) \]

then we may write (7.6) as
\[ -\begin{vmatrix} 0 & (\partial\phi(\hat{\theta})/\partial z)' \\ \partial\phi(\hat{\theta})/\partial z & \text{CVC'} \end{vmatrix} \chi^2_h(\alpha)\sigma^2 = \chi^2_2(\alpha)\sigma^2 \text{CVC'} \]
or
\[ \begin{vmatrix} \sigma^2\chi^2_h(\alpha) & (\partial\phi(\hat{\theta})/\partial z)' \\ \partial\phi(\hat{\theta})/\partial z & \text{CVC'} \end{vmatrix} = 0, \quad (7.8) \]

where \( \chi^2_h(\alpha) \) is the upper \( \alpha \) percentage point of the \( \chi^2 \) distribution with \( h \) degrees of freedom.

The \((h+1)\times(h+1)\) determinant in (7.8) is a function of the \( Z \)'s which defines an exact confidence region for \( E(\hat{Z}) \). This, then, is the region of all \( Z \) values such that, with probability \( 1-\alpha \), the null hypothesis in (7.3) is not rejected.

If \( \sigma^2 \) is unknown, then,
\[ \frac{(\partial\phi(\hat{\theta})/\partial z)'(\text{CVC'}^{-1})(\partial\phi(\hat{\theta})/\partial z)}{h\hat{\sigma}^2} = F, \quad (7.9) \]
where \( F \) follows the \( F \) distribution with \( h \) and \( nq-(\sum_{j=1}^{q} p_j+q) \) degrees of freedom. In this case, the exact confidence region for \( E(\hat{Z}) \) is determined from
\[
\begin{bmatrix}
\hat{h}^2_{F_{\lambda}}[h, nq - (\sum_{j=1}^{q} p_j + q)] & \partial \hat{\theta}(\hat{y}) / \partial Z \\
\vdots & \vdots \\
\partial \hat{\theta}(\hat{y}) / \partial Z & CVC'
\end{bmatrix} = 0. \quad (7.10)
\]

When \(X_j = X\) for all \(j\); i.e., all the \(\hat{y}_j\) are second-order regression in the same \(Z\)'s,

\[
\partial \hat{\theta}(\hat{y}) / \partial Z = \begin{bmatrix}
t_1^1 \quad t_1^1 \quad t_2 \quad t_1^1 \quad \cdots \quad t_q \quad t_1^1 \\
pxl \quad pxl \quad pxl \quad pxl \quad \cdots \\
t_1 \quad t_1^h \quad t_2 \quad t_1^h \quad \cdots \quad t_q \quad t_1^h \\
pxl \quad pxl \quad pxl \quad pxl \\
\hat{b}_1 \\
\hat{b}_q \\
\end{bmatrix}
\]

or

\[
\partial \hat{\theta}(\hat{y}) / \partial Z = \hat{C}_B,’
\]

and \(\tilde{CVC}'\) becomes

\[
\tilde{CVC}' = \begin{bmatrix}
\tilde{t}_1^1 (X'X)^{-1} \tilde{t}_1^1, \cdots, \tilde{t}_1^1 (X'X)^{-1} \tilde{t}_1^h \\
\vdots \\
\tilde{t}_1^h (X'X)^{-1} \tilde{t}_1^1, \cdots, \tilde{t}_1^h (X'X)^{-1} \tilde{t}_1^h
\end{bmatrix} \tilde{t}' \sum \tilde{t} = \Omega \tilde{t}' \sum \tilde{t}, \quad (7.11)
\]

where from (6.28) \(\hat{b} = (X'X)^{-1} \Sigma\). From (6.34) and (6.10) we have \(\sum_{j=1}^{q} t_j E_j : N(Q, (I-U) t' \sum \tilde{t}), \) where

\[
U = \begin{bmatrix} 1 & X \end{bmatrix} \begin{bmatrix} 1' \\ X' \end{bmatrix} \begin{bmatrix} 1' \\ X \end{bmatrix}^{-1} \begin{bmatrix} 1' \\ X' \end{bmatrix}.
\]

And since \(I-U\) is idempotent, it necessarily follows that
\[
\left( \sum_{j=1}^{q} t \ e_j \right) \left( \sum_{j=1}^{q} t \ e_j \right) = \frac{t'(e_i e_j)}{\sum t} = (n-p-1) \frac{\sum t}{\sum t}
\]

is distributed as \( \chi^2 \) with \( n-p-1 \) degrees of freedom. Hence,

\[
\left[ \frac{\partial \hat{\phi}(\hat{y})/\partial \alpha}{\partial \hat{y}/\partial \beta} \right]^{-1} \left[ \frac{\partial \hat{\phi}(\hat{y})/\partial \beta}{\partial \hat{y}/\partial \alpha} \right] = F
\]

(7.12)

is distributed as \( F \) with \( h \) and \( n-p-1 \) degrees of freedom, and the exact confidence region for \( E(\hat{z}) \) is formed from

\[
\left| \begin{array}{c}
\frac{\partial \hat{\phi}(\hat{y})/\partial \alpha}{\partial \hat{y}/\partial \alpha} \\
\vdots \\
\frac{\partial \hat{\phi}(\hat{y})/\partial \beta}{\partial \hat{y}/\partial \beta}
\end{array} \right| = 0.
\]

(7.13)

The determinant in (7.13) is precisely the confidence region derived by Box and Hunter \( \chi^2 \), if \( t'y \) is regressed on \( x \). This is so because the equation resulting from the latter and \( t'\hat{y} \) are identical as was noted earlier. However, by dealing with linear functions of regressions rather than one regression of a linear combination of the observations on \( x \), it may be possible to explain and/or remedy irregularities in the confidence region by omitting certain \( y_j \). Irregularities in the confidence region will be discussed in the next section.

7.3 Remarks Concerning the Confidence Region

Following the reasoning and terminology of Box and Hunter \( \chi^2 \), the region allowed by (7.8), (7.10), or (7.13) will now be discussed.
The size and shape of the confidence region depends not only on the adequacy of the model, \( \phi[E(\hat{y})] = \hat{\alpha}^T \hat{B} x \), but also on the nature of the response surface itself.

Roughly speaking, the experimental error influences the size of the region, while the nature of the response surface is ascendant with regard to shape. It is evident that a large error mean square stemming from either an inadequate model or insufficient error degrees of freedom will give rise to a large confidence region. If, on the other hand, the experimental error is small and the response surface portrays, say, a stationary ridge, then clearly, the confidence region for \( \hat{E}(\hat{z}) \) must certainly be attenuated in the direction of the ridge. Further, consider the following situation where the nature of the surface influences the size of the region. If the surface is near symmetrical concave paraboloid with the curvature in the vicinity of \( \phi(\hat{y}|\hat{z}) \) being sufficiently small, then there would exist many values close to the true maximum such that in repeated sampling diverse \( \hat{z} \) values may occur even with a small experimental error. In such a situation, a large confidence region may result.

At this point we will consider the detailed aspects of the shape of the confidence region as determined by the nature of the surface. An aspect greatly influencing the shape of the region is whether the equations

\[
\frac{\partial \phi(\hat{y})}{\partial \hat{z}} = \Delta, \tag{7.14}
\]
where
\[ \Delta = \frac{\partial \hat{E}(\hat{y})}{\partial \hat{Z}} - \frac{\partial \Phi(\hat{y})}{\partial \hat{Z}}, \]
are well or poorly conditioned. In explanation of this term, we first rewrite \( \Phi(\hat{y}) \) in the well-known form,
\[ \Phi(\hat{y}) = r'Z + Z'RZ, \quad (7.15) \]
assuming, of course, that \( \Phi(\hat{y}) \) is a second-degree function of all the \( Z \)'s. Then \( \partial \Phi(\hat{y})/\partial Z = r + RZ \).

The quadratic function, \( Q \), defined by
\[ Q = \left[ \frac{\partial \Phi(\hat{y})}{\partial Z} \right] \left[ \frac{\partial \Phi(\hat{y})}{\partial Z} \right]' \quad (7.16) \]
is the surface which Box and Hunter call the conditioning surface in the \( h+1 \) dimensional space of \( Q,Z_1,Z_2, \ldots, Z_h \).

For \( |R| \neq 0 \) and for fixed values of \( Q \), (7.16) defines hyper-ellipsoids in the \( Z \) space. Neglected in (7.16) is the covariance matrix of \( \left[ \frac{\partial \Phi(\hat{y})}{\partial Z} \right] \) which is usually placed "between" the product vectors so that well-known distributions are applicable. Clearly, any extreme elongations in any of the \( h \) dimensions of the hyperellipsoid defined by fixing \( Q \) must induce the same effect when the covariance matrix is accounted for. Hence, an examination of (7.16) allows a critical study of the confidence region apart from any lack of fit of one or more of the regressions taken in linear combination.

Suppose now we refer the hyperellipsoids to their principal axes, say, \( \hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_h \). First translate the center of the \( Z \) axes to \( \hat{Z} \), the center of the hyperellipsoids, so that
\[ Q = (\hat{Z}^T R' R \hat{Z}) \text{, where } \hat{Z} = -R^{-1}r \text{ by setting } \partial \Theta(\hat{Z})/\partial Z = 0. \] 

Next apply the orthogonal transformation which transforms the above quadratic into

\[
Q = \hat{Z}^T \Theta \hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_h) \begin{bmatrix}
\theta_1 & 0 & \cdots & 0 \\
0 & \theta_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \theta_h
\end{bmatrix} \begin{bmatrix}
\hat{Z}_1 \\
\hat{Z}_2 \\
\vdots \\
\hat{Z}_h
\end{bmatrix} = \sum_{\ell=1}^{h} \theta_\ell \hat{z}_\ell^2,
\]

where the \(\theta_\ell\) are the \(h\) characteristic roots of \(R' R\), \(Z_\ell = u_\ell^T (Z - \hat{Z})\), and \(u_\ell\) is the characteristic vector corresponding to the root \(\theta_\ell\). Notice now that if one or more of the characteristic roots, say, \(\theta_1, \ldots, \theta_s\), \(s < h\), is small compared with the remainder of the roots, then the conditioning surfaces are attenuated in the direction of the axes corresponding to the smaller roots.

As such, it is possible for values of \(Z\), which differ greatly from the correct solution, to correspond to points near a hyperplane passing through the axes of \(\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_s\) such that there result small values of \(Q\) and hence of \(\Delta = \partial \Theta(\hat{Z})/\partial Z\). From this it is seen that there will exist many "nearly correct" solutions to the equations (7.2). In this case, Box and Hunter call the equations (7.14) ill or poorly conditioned as opposed to well conditioned equations where all the roots of \(R' R\) are equal or nearly so.

Turing proposed \(g(R' R) = (1/h)(\sum_{\ell=1}^{h} \theta_\ell \sum_{\ell=1}^{h} \theta_\ell^{-1})^{1/2}\) as a criterion for conditioning of equations. Clearly, if \(g(R' R) = 1\), then \(\theta_\ell = \theta\) for all \(\ell\) (well conditioned equations).
On the other hand, if some of the $\theta_k$ are small compared with
the remainder of the roots, then $g(R'R)$ is small, implying
poorly conditioned equations.

If $\hat{\phi}(\hat{\mathbf{y}})$ is a linear combination of second-order regres-
sion functions in the same $Z$'s, then $\hat{y}_j = a_j'Z + Z'A_jZ$, and

$$\hat{\phi}(\hat{\mathbf{y}}) = \mathbf{t}'\hat{\mathbf{y}} = \sum_{j=1}^{q} \mathbf{t}_j a_j'Z + \sum_{j=1}^{q} \mathbf{t}_j Z'A_jZ,$$

(7.17)

so that $\partial \hat{\phi}(\hat{\mathbf{y}})/\partial Z = \sum_{j=1}^{q} \mathbf{t}_j a_j + \sum_{j=1}^{q} \mathbf{t}_j A_j Z_j$. Consequently, $r$ and $R$ in
(7.15) are $r = \sum_{j=1}^{q} \mathbf{t}_j a_j$ and

$$R = \sum_{j=1}^{q} \mathbf{t}_j A_j.$$

(7.18)

Noting that the roots of $R'R$ are the squares of the roots
of $R$ with the characteristic vectors being the same, then it
is easily seen that the nature of the response surface as
written in (7.15) or (7.17) reflects the conditioning of the
equations in (7.14).

Under (7.18), if $g(R)$ and consequently $g(R'R)$ is small,
then the ill conditioning is due either

(i) to the inherent nature of the surface, $\hat{\phi}(\hat{\mathbf{y}})$, even
though each set of equations $\partial \hat{y}_j/\partial Z = \Delta_j$, $j=1,2,\ldots,q$,
is well conditioned; i.e., a linear combination of
well conditioned equations may yield poorly con-
ditioned equations;

(ii) or to one or more of the sets of equations

$\partial \hat{y}_j/\partial Z = \Delta_j$ being ill conditioned such that when taken
in linear combination, the resultant equations in (7.14) are ill conditioned; i.e., a linear combination of well and poorly conditioned equations may result in poorly conditioned equations.

On the other hand, a linear combination of poorly conditioned equations may produce well conditioned equations. In case (ii), $g(R)$ would be small and the diversity of the roots of $R = \sum \lambda_j A_j$ is due to one or more of the $\lambda_j A_j$. A recent theorem by Ostrowski [24] allows a study of linear functions of poorly and well conditioned equations (resulting in poorly conditioned equations due to the poorly conditioned equations taken in linear combination) and may possibly direct one to that/those $A_j$ with roots distinctly different from $R$ without laborious computations.

**Theorem.** Let $H = (h_{ij})$, $\hat{H} = (\hat{h}_{ij})$ be two nxn matrices with $\bar{\nu}_H(\Theta) = |H - \Theta I| = 0$, $\bar{\nu}_H^\circ(\Theta) = |\hat{H} - \Theta I| = 0$ being the corresponding characteristic equations of $H$ and $\hat{H}$, respectively. Denote the roots of $\bar{\nu}_H(\Theta)$ and $\bar{\nu}_H^\circ(\Theta)$ by $\theta_i$ and $\hat{\theta}_i$, respectively. Set

$$m = \max(|h_{ij}|, |\hat{h}_{ij}|), i, j = 1, 2, \ldots, n, \quad (1/nm) \sum_i \sum_j |h_{ij} - \hat{h}_{ij}| = d.$$ 

Then for at least one pair, $\theta_i, \hat{\theta}_i$, we have

$$|\theta_i - \hat{\theta}_i| \leq (n+2)md^{1/n} \quad (7.19)$$

and for all $\theta_i, \hat{\theta}_i$,

$$|\theta_i - \hat{\theta}_i| \leq 2(n+1)^2md^{1/n}. \quad (7.20)$$

Now, let the inclusion of any subscript on $R$, say $R_j$, imply that the corresponding $A_j$ is not taken in linear combination.
For example, \( R_{1,2,3} = \sum_{j=4}^{q} \tau_j A_j = R - (\tau_1 A_1 + \tau_2 A_2 + \tau_3 A_3) \). Denote by \( R_{(-)} \) the exclusion of at least one \( A_j \).

Suppose we find upper bounds on the difference between at least one pair and all pairs of roots of \( R \) and \( R_{(-)} \) by applying (7.19) and (7.20), respectively. Suppose further that (7.20) yields a small value (though in examples taken, (7.19) and (7.20) tend to give conservative upper bounds; i.e., they are too large). Then if the roots of \( R \) are close, so must be the roots of \( R_{(-)} \). Hence, \( R \) and \( R_{(-)} \) both correspond to sets of well conditioned equations. If, on the other hand, the upper bound on the difference between at least one pair of roots of \( R \) and \( R_{(-)} \) is large when \( g(R) \) is small, then there is the indication that \( R_{(-)} \) corresponds to a set of poorly conditioned equations (though this is not necessarily true).

In another example, suppose the \( A_j \) are to be taken in linear combination one at a time, or the inclusion of each \( A_j \) is through a stepwise procedure. Starting with \( A_1 \) and supposing its roots are close, we then form \( A_1 + \tau_2 A_2 \). If (7.19) is small and (7.20) is relatively small for the matrices, \( A_1 \) and \( A_1 + \tau_2 A_2 \), then \( A_2 \) corresponds to a set of well conditioned equations. Assuming this to be true until the \( i \)th step where (7.19) is relatively small and (7.20) is excessively large for the matrices, \( \sum_{j=1}^{i=1} \tau_j A_j \), \( \sum_{j=1}^{i} \tau_j A_j \), then it necessarily follows that \( \tau_i A_i \) in

\( \sum_{j=1}^{i} \tau_j A_j \) corresponds to a set of poorly conditioned equations,
since those associated with \[ \sum_{j=1}^{i} l_j A_j \] are well conditioned. In this situation one may possibly omit \( A_i \) from consideration after comparing desired analyses including and excluding \( A_i \).

Finally, mention is given to open and closed confidence regions. The \( \hat{Z}_i \) determine the type of conic the second-degree function, \( \phi(y) \), describes. The reader is referred to reference \( \text{[citation]} \) for illustrations of the basic types of contours (for \( h=2 \) and \( h=3 \)) arising from fixing \( \phi(\hat{y}) \). If the contours are closed with the \( \hat{Z}_i \) being of the same sign, and if the plausible variation associated with each coefficient is such that none of the \( \hat{Z}_i \) will change sign, then the confidence region will be closed and will describe a region for \( Z \) values giving maximum (all the \( \hat{Z}'s \) are positive) attainment for \( \phi(\hat{y}) \). If, on the other hand, the plausible variation associated with the coefficients is such that the \( \hat{Z}_i \) may change sign (so that some are positive and some negative) then open contours result, and an open confidence region may result. The latter may be true for, say, a saddle point in a saddle type surface.

7.4 Intersection Region Confidence Procedures in
Finding an Approximate Confidence Region for \( E(\hat{Z}) \)

Consider the null hypothesis given in (7.3). For any non-null, preselected hxl vector \( k \), \( H_0 \) is rejected if

\[
H_{0k} : k' \left( \frac{\partial \phi[E(\hat{y})]}{\partial Z} \right) = 0
\]

(7.21)

is rejected. With the customary normality assumption, no
mathematical difficulties are encountered in defining confidence regions or intervals. However, with $k$ dependent on the data, only approximate methods were available prior to Scheffe's work in multiple comparisons \cite{29}, and Roy's extensive developments in simultaneous confidence bounds \cite{26}. Wallace \cite{40} draws on this theory in simplifying the confidence region defined by (7.10) and (7.13).

The algebra necessary for obtaining the confidence regions in Section 7.2 is, even with two $Z$ variables, cumbersome to the point of possibly curtailing its use. Fully aware of this, Wallace states "what is important is that the region be represented geometrically or analytically so that the user can comprehend its size, shape and location. Approximations to the region which simplify this representation will be valuable so long as they do not greatly change the confidence level" \cite{40}, p. 455. Motivated by Scheffe's method of making all contrasts in the ellipsoidal type confidence region, Wallace approximates nonellipsoidal regions as (7.10) or (7.13) by paralleloipipids or convex polyhedra.

From (7.5), it follows that for preselected $k \neq 0$ under $H_{0k}$, $k'\left[ \frac{\partial \Theta(\hat{\chi})}{\partial Z} \right] : N(0, k'A_k)$, where $A=CVC'\hat{\sigma}^2$ under (7.5) and $A=\bigodot l'\sum_i l$ under (7.11). The set or region of all $Z$ such that $H_{0k}$ is not rejected is found from

$$p \left\{ \frac{k'\left[ \frac{\partial \Theta(\hat{\chi})}{\partial Z} \right]}{\left( k'\hat{A}_k \right)^{1/2}} \leq t_\alpha \right\} = 1 - \alpha, \quad (7.22)$$

where $\hat{A}=CVC'\hat{\sigma}^2$ or $\bigodot l'\sum_i l$ and $t_\alpha$ is the upper $\alpha$ percentage
point of Student's distribution with degrees of freedom,

\[
f = nq - \left( \sum_{j=1}^{p} p_j q \right) \quad \text{if } \hat{A} = \text{CVC}' \hat{\sigma}^2
\]

\[= n - p - 1 \quad \text{if } \hat{A} = \Omega \sum \hat{\ell} \]

Or (7.22) implies that

\[
p \left[ \frac{\left( \frac{\partial \hat{\phi}(\hat{y})}{\partial \hat{Z}} \right)^2 \hat{\sigma}^2}{\hat{\phi}^2 \hat{\sigma}^2} \right] \leq \frac{1}{F_\alpha(1,f)} = 1 - \alpha. \quad (7.23)
\]

Consider the preselected non-null vectors, \(k_1, k_2, \ldots, k_L, \ldots, k_h\). With each \(k_L\) there is associated a confidence region or interval via

\[
\frac{\left[ k_L \left( \frac{\partial \hat{\phi}(\hat{y})}{\partial \hat{Z}} \right)^2 \hat{\phi}^2 \hat{\sigma}^2 \right]}{(k_L \hat{\phi}^2 \hat{\sigma}^2)} \leq \frac{1}{F_\alpha(1,f)}.
\]

Suppose that the \(k_L\) were selected after an examination of the data. Clearly, then, the probability in (7.23) is not equal to \(1 - \alpha\). In this case, one may attempt to find the exact distribution of the above statistic. However, when \(k_L\) is functionally dependent on the data, attempts in this direction may be futile. Due to a theorem by Scheffe \(29\), another recourse is available. Here the distribution of

\[
\sup_k \left[ \frac{k_L \left( \frac{\partial \hat{\phi}(\hat{y})}{\partial \hat{Z}} \right)^2 \hat{\phi}^2 \hat{\sigma}^2 \right] / k_L \hat{\phi}^2 \hat{\sigma}^2 \leq \frac{1}{F_\alpha(1,f)} \quad (7.24)
\]

is found; i.e., the distribution of the least upper bound
of the statistic,

\[
\left( \frac{k_t^\prime \frac{\partial \Phi (\hat{y})}{\partial Z}}{k_t} \hat{A} k_t, \right)^2
\]

is obtained over all non-null \( k_t \). Clearly, the main disadvantage of this approach is over-protection or the possibility of drastic conservatism.

Proceeding in detail, \( F^* \) is to be found such that

\[
P \left( \sup_{k_t} \frac{(7.24)}{F^*} \right) = 1 - \alpha. \quad (7.25)
\]

It is easily seen that (Roy /267, p. 95) (7.24) is equal to \( \left[ \frac{\partial \Phi (\hat{y})}{\partial Z} \right] ^\prime \hat{A}^{-1} \left[ \frac{\partial \Phi (\hat{y})}{\partial Z} \right] \). But if \( \hat{A} = C V C^\prime \hat{\sigma}^2 \), then from (7.9) it follows that (7.24) is distributed as hF where \( F \) is an \( F \) variable with \( h \) and \( n q - \sum_{j=1}^q p_j + q \) degrees of freedom; further, if \( \hat{A} = \bigcap L \bigcup L \), then from (7.12) it is seen that (7.24) is distributed as hF where \( F \) is an \( F \) variable with \( h \) and \( n-p-1 \) degrees of freedom.

Hence (7.25) \( \iff \) (7.26) where

\[
P \left( \sup_{k_t} \left[ \frac{\partial \Phi (\hat{y})}{\partial Z} \right]^2 \right) \leq hF_{\alpha} (f_1, f_2) \quad (7.26)
\]

for all non-null \( k; \ i.e., \) (7.25) implies (7.26) and (7.26) implies (7.25) so that (7.25) \( \iff \) (7.26).

The significance of (7.26) is that with each \( k_t \neq 0 \) defining a confidence region or interval, the probability of the infinity of such intervals or regions being true
simultaneously is \(1 - \alpha\). Note that \(k\) is no longer required to be independent of the data. However, if only \(h\) non-null \(k\) vectors are chosen (resulting in \(h\) intervals or regions), the probability of the \(h\) statements holding simultaneously is necessarily \(\geq 1 - \alpha\). (See Roy \(267\), p. 105, or Wallace \(407\), p. 461.)

Wallace writes

\[
h_t(Z) = \left( k_t^T \left[ \frac{\partial \theta(\hat{y})}{\partial Z} \right] \right)^2 - k_t^T \hat{A} k_t h \Phi_\alpha
\]

and observes that the latter is a quadratic equation in the \(Z\)'s which defines a region \(R^{(t)}\) "between the sheets of a two-sheeted hyperboloid, the exterior of an ellipse, or limiting and transitional forms of these. The boundary, \(D_t\), can never be one of the one-sheeted hyperboloids." The region formed by the intersection of the \(h_t(Z), t=1,2,\ldots,h\), is defined as the intersection confidence region and is denoted by \(R^{(I)}\). Observe that the confidence region is approximate in the sense that \(P \geq 1 - \alpha\) with finite \(h\) where \(P \rightarrow 1 - \alpha\) as \(h\) increases indefinitely.

If, say, \(h=2\), then from \(k_1\) and \(k_2\) result \(h_1(Z)\) and \(h_2(Z)\). Letting both quadratic equations define two-sheeted hyperbolas which intersect, then an example of the intersection confidence region, \(R^{(I)}\), is depicted in Figure 7.1 as transformed axes, \(v_1\) and \(v_2\).
Figure 7.1 The intersection confidence region, \( R^{(I)} \), for \( E[(z_1, z_2)] \) for specific \( k_1, k_2 \) vectors.

Wallace gives three types of approximations to the region, \( R^{(I)} \). The first approximation region, \( R^{(1)} \), is obtained by connecting the 2\(^{h} \) intersection points of the \( h_{k}(Z)=0 \), assuming there exist 2\(^{h} \) intersection points. Figure 7.2 illustrates \( R^{(1)} \) as the area enclosed by lines joining the four intersection points.

Figure 7.2 The \( R^{(1)} \) approximation to \( R^{(I)} \).
The second approximation region, $R^{(2)}$, for hyperboloids (two-sheeted), and probably the most useful in practical situations, is formed by the intersection of hyperplanes tangent to each of the pairs of surfaces (hyperboloids), $D_4$, the tangencies occurring at points of intersection of the hyperboloids and the transformed axes. Figure 7.3 depicts $R^{(2)}$ as the closed area formed by the intersection of tangent lines to the pairs of curves, $D_1$ and $D_2$.

![Diagram](image)

**Figure 7.3** The $R^{(2)}$ approximation to $R^{(1)}$

Finally, the third type of approximation, or the parallelepiped approximation region, $R^{(3)}$ (again for hyperboloids), is formed by approximating the pair $D_4$ by hyperplanes parallel to $k_j \left[ \partial \phi \partial (\phi) / \partial z \right] = 0$. Figure 7.4 depicts this region in reference to the other figures. The reader is referred to Wallace \[40\] for the detailed mathematical and geometrical aspects of these regions.
Figure 7.4 The $R^{(3)}$ approximation to $R^{(1)}$

For purposes of simplifying the formulation of $R^{(1)}$, $R^{(2)}$, and $R^{(3)}$, new coordinates, $v_1, v_2, \ldots, v_h$, are defined by $v_i = k_i \frac{\partial \phi(\nu)}{\partial \nu}$ in the Z space for specific choices of $k_1, k_2, \ldots, k_h$. Then,

$$h_i(Z) = H_i(v) = v_i^2 - k_i \hat{\lambda} k_i \mu \phi(x)(f_1, f_2)$$

is a quadratic in the $v_i$. Further, letting $P_i$ be the hyperplane, $v_i = 0$, then "the pair of points lying on $D_i$ and every $P_i'$, $i \neq i'$, have all v coordinates but the $i$th zero, and that one given by the roots $(r_i' < r_i''$) of the quadratic, $H_i(0, \ldots, 0, v_i, \ldots, 0) = 0."$ If $r_i' < 0 < r_i''$, then the equation of the tangent hyperplane approximation, $R^{(2)}$, to the part of $D_i$ with $v_i > 0$ is
\[
\sum_{l \neq \ell} v_l \left[ \frac{\partial H_{\ell}(v)}{\partial v_l} \right] v' = (0, \ldots, r''_l, \ldots, 0) \\
+ (v_\ell, -r''_\ell) \left[ \frac{\partial H_{\ell}(v)}{\partial v_\ell} \right] v' = (0, \ldots, r''_\ell, \ldots, 0)
\]

(see Taylor \cite{327}, p. 151). Thus, with \( H_{\ell}(v) = 0 \) given, \( R^{(2)} \) is easily formed. It follows that \( R^{(3)} \) is given as \( r''_l \leq v_\ell \leq r''_l \), \( l=1, \ldots, h \). From this point, one may transform the inequalities back to the Z's.

With regard to the choice of \( k_1, \ldots, k_h \), if \( k_\ell = U_\ell / (\theta_\ell)^{1/2} \), where \( U_\ell \) and \( \theta_\ell \) were previously defined as the characteristic vectors and roots of \( R' R \) or \( U_\ell \) and \( \theta_\ell^{1/2} \) are the vectors and roots of \( R \), \( R \) being defined in (7.14)\cite{7}, then

\[
v_\ell = \left[ \frac{U_\ell}{(\theta_\ell)^{1/2}} \right] \left[ \frac{\partial \theta(\hat{v})}{\partial Z} \right] = \left[ \frac{U_\ell}{(\theta_\ell)^{1/2}} \right] (r + RZ) \\
= \left[ \frac{U_\ell r}{(\theta_\ell)^{1/2}} \right] + \left[ \frac{U_\ell RZ}{(\theta_\ell)^{1/2}} \right] \\
= \left[ \frac{U_\ell r}{(\theta_\ell)^{1/2}} \right] + U_\ell Z = U_\ell (Z - \hat{Z}),
\]

the last step following by equating (7.15) to zero, which gives \( r = -\hat{Z} \) so that \( U_\ell r / (\theta_\ell)^{1/2} = -U_\ell (Z - \hat{Z}) \).

But from Section 7.3, \( \hat{Z}_\ell = U'_\ell (Z - \hat{Z}) \). Hence the transformation to the new coordinate \( v \) system is precisely the same as referring \( \theta(\hat{v}) \) to canonical form (or is the same as referring the hyperellipses, \( Q = [\partial \theta(\hat{v})/\partial Z]'[\partial \theta(\hat{v})/\partial Z] \), to their principal axes). Then from the canonical form of \( \theta(\hat{v}) \), one can proceed to the formulation of \( H_{\ell}(v) = H_{\ell}(\hat{Z}) \), \( l=1, 2, \ldots, h \), and the various approximate confidence regions.
8.0 SOME PROBLEMS OF CONSTRAINED EXTREMA

8.1 Uses of Lagrangian Multipliers in a System of Linear Regression Equations

When linear functions of the $\hat{y}_j$ were first considered in Section 7.1, it was assumed that all $\hat{y}_j$ taken in linear combination (for purposes, say, of maximization or minimization) were in the same or comparable units of measure. Developments under this assumption are restrictive since, even in the present problem, toxoplasma exudate yields three characteristic response types in two units of measure--volume and number. Maximizing $(\hat{y}_T-\hat{y}_L)+\hat{y}_V$ is inappropo as $\overline{Z}$ is not invariant under transformation of scale unless $\hat{y}_T-\hat{y}_L$ and $\hat{y}_V$ attain maximums for the same $\overline{Z}$ value, or approximately so, in which case no further developments are necessary.

The intent in this chapter is to consider extremal relationships between response types in different units of measure. This concept is best exemplified with the toxoplasma problem.

Suppose that $y_T-y_L$ and $y_V$ can be described adequately by concave parabolic surfaces with $\hat{y}_T-\hat{y}_L$ and $\hat{y}_V$ attaining maximums for distinctly different $\overline{Z}$ values, say, $\overline{Z}_{T-L}$ and $\overline{Z}_V$. If that amount and dilution of inoculum which maximizes $\hat{y}_T-\hat{y}_L$ also produces an unacceptably low volume of exudate, and conversely, if that amount and dilution of inoculum which maximizes $\hat{y}_V$ also produces an unacceptably small difference in opposing organisms, then the bacteriologist is forced to compromise, assuming that the inoculation of a greater number of
mice is unfeasible. Compromise values of $Z_i$, estimated by, say $\hat{Z}_i$, fall in the range

$$\min(\hat{Z}_{T-L}, \hat{Z}_V) \leq \hat{Z}_i \leq \max(\hat{Z}_{T-L}, \hat{Z}_V);$$

i.e., compromise values of $Z_1$ stem from the interval

$$\min(\hat{Z}_{1,T-L}, \hat{Z}_{1,V}) \leq \hat{Z}_1 \leq \max(\hat{Z}_{1,T-L}, \hat{Z}_{1,V}),$$

and those of $Z_2$ from

$$\min(\hat{Z}_{2,T-L}, \hat{Z}_{2,V}) \leq \hat{Z}_2 \leq \max(\hat{Z}_{2,T-L}, \hat{Z}_{2,V}),$$

since $\hat{V}_{T-L}$ and $\hat{V}_V$ describe concave parabolic surfaces. A compromise implies the acceptance of a lower volume of exudate and a smaller difference between opposing organisms. The sacrifice inherent in such a compromise is, in the former case, fewer sera determinations, and, in the latter, a lower precision in the LD 50.

One approach to a proper selection among the $\hat{Z}_i$ is as follows. Between the functions $\hat{V}_V$ and $\hat{V}_{T-L}$, one must be chosen as the constraining function and the other as the function to be maximized. If both functions describe concave paraboloids, the choice is somewhat arbitrary. If not, then the function which lends itself to maximization most conveniently within the range of interest becomes (possibly) the function to be maximized. Let $\hat{V}_{T-L}$ and $\hat{V}_V$ take the roles of maximizing function and constraining function, respectively. Select a lower bound for the acceptable exudate volumes, say,
where $C_M$ is the lowest acceptable volume of exudate. With the surfaces $\hat{y}_T - \hat{y}_L$ and $\hat{y}_V$ being concave paraboloids, then $\hat{z}_\beta$ is that $Z$ value such that $\hat{y}_T - \hat{y}_L$ is maximum for $Z$ on the boundary of $\hat{y}_V = C_\beta$. Consequently, by varying $C_\beta$ under (8.1), the set

$$\left[ \hat{y}_V = C_\beta, \hat{z}_\beta, \text{max}(\hat{y}_T - \hat{y}_L \mid C_\beta) \right]$$

includes all possible compromise solutions. From physical considerations,

$$\left[ \hat{y}_V = C_0, \hat{z}_0, \text{max}(\hat{y}_T - \hat{y}_L \mid C_0) \right]$$

is selected and defined to be the optimal solution with $\hat{y}_V = C_0$ and $\text{max}(\hat{y}_T - \hat{y}_L \mid C_0)$ being the optimal values of $\hat{y}_V$ and $\hat{y}_T - \hat{y}_L$, and $\hat{z}_0$, the optimal $Z$ vector. For example, the selection may possibly involve associating the precision of the LD 50 with values of $\text{max}(\hat{y}_T - \hat{y}_L \mid C_\beta)$. An acceptable precision (specified from the risk of deciding that a serum is negative when in fact it is positive) would then determine the optimal solution.

The process is depicted in Figure 8.1. The closed contours surrounding the point $Q_4$, which corresponds to $\hat{z}_V$, stem from dropping the boundaries formed by the intersection of the paraboloid $\hat{y}_V$ and the planes $C_M \leq \hat{y}_V = C_\beta$ onto the $Z$ axes. The maximum values of $\hat{y}_T - \hat{y}_L$ for $Z$ on the boundaries of $C_M \leq \hat{y}_V = C_\beta$ are on the curve, $Q_1Q_2$, where the $Z$ vectors defined by $Q_1Q_3$ comprise the entire set $(\hat{z}_\beta)$. Max$(\hat{y}_T - \hat{y}_L \mid C_M)$ is denoted by point $Q_1$ which represents the largest difference in opposing
Figure 8.1 A geometric example of compromise solutions and the optimal solution
organisms at the lowest acceptable volume of exudate. After a study of (8.2) and through other considerations, (8.3) is selected as the optimal solution where, say, \( Q' = \max(\hat{\gamma}_T - \hat{\gamma}_L | C_0) \).

From the above discussion, it is clear that an expedient method of determining values of \( Z_\beta \) is by the use of Lagrangian multipliers, since one function is maximized, subject to a varied constraint.

Proceeding with a general formulation of the problem, let each of the \( q \) response types be in one of \( t(t \leq q) \) units of measure. Let \( \mathcal{L}^*(y) \), a linear combination of \( y \)'s in one unit of measure, be the function to be maximized. Let \( \mathcal{L}_\gamma(y) \) be a function of those \( y \)'s in unit of measure \( \gamma \), \( \gamma = 1, 2, \ldots, t-1 \), where the \( \mathcal{L}_\gamma \) will assume the role of constraining functions. If

\[
y_j = f_j(Z), \quad j = 1, 2, \ldots, q,
\]

then \( \mathcal{L}_\gamma(y) = \mathcal{L}_\gamma[f(Z)] = L_\gamma(Z) \) and \( \mathcal{L}^*(y) = L^*(Z) \); e.g.,

\[
\hat{y}_T - \hat{y}_L = L^*(Z) = Z'r_1 + Z'R_1Z \quad \text{and} \quad \hat{y}_V = L_1(Z) = Z'r_2 + Z'R_2Z
\]

where \( L \) is estimated by \( \hat{L} \).

Assume that \( L^* \) and \( L_\gamma(Z) - C_\beta \gamma \geq 0 \) are regular functions for \( Z \in K_1 \), and \( h \) dimensional region in the \( Z \) space, where the \( C \)'s are constants such that

\[
C_{\beta 1} \geq C_{M1} \\
\ldots \\
C_{\beta \gamma} \geq C_{M\gamma} \\
\ldots \\
C_{\beta, t-1} \geq C_{M, t-1}
\]
or for brevity

\[ C_\beta \succ C_M \quad (8.5) \]

Further assume that \( h > t - 1 \); i.e., the number of constraints is less than the dimensionality of \( Z \). Denote by \( \mathcal{K}_\beta \) the region defined by

\[ \bigcap_{\gamma = 1}^{t-1} \left[ L_{\gamma}(Z) - C_{\beta \gamma} = 0 \right] ; \quad (8.6) \]

e.g., \( \mathcal{K}_\beta \) is that region or the ellipse in Figure 8.1 formed by the boundary of \( \hat{y}_\gamma = C_\beta \).

To determine the extremal positions of \( L^* \) subject to the constraints \( L_{\gamma}(Z) - C_{\beta} \succ 0 \), we use the Lagrangian form

\[ \Psi(Z, \lambda_\beta) = L^*(Z) - \sum_{\gamma = 1}^{t-1} \lambda_{\gamma} L_{\gamma}(Z) - C_{\beta \gamma} \quad (8.7) \]

where \( \lambda_\beta = (\lambda_{\beta 1}, \ldots, \lambda_{\beta \gamma}, \ldots, \lambda_{\beta, t-1}) \) is a vector of Lagrangian multipliers, take partial derivatives of \( \Psi \) with respect to \( Z \) and \( \lambda_\beta \), then equate the \( h + (t-1) \) expressions \( \partial \Psi / \partial Z \) and \( \partial \Psi / \partial \lambda_\beta \) to zero; i.e.,

\[ \frac{\partial \Psi}{\partial Z} = \frac{\partial L^*}{\partial Z} - \sum_{\gamma = 1}^{t-1} \lambda_{\gamma} \frac{\partial L_{\gamma}}{\partial Z} = \frac{0}{hx1} \quad (8.8) \]

\[ \frac{\partial \Psi}{\partial \lambda_\beta} = L(Z) - C_\beta = \frac{0}{(t-1)x1} \quad , \quad (8.9) \]

where \( L'(Z) = [L_1(Z), \ldots, L_{\gamma}(Z), \ldots, L_{t-1}(Z)] \).

Solving equations (8.8) and (8.9) simultaneously, we find among the \( Z \) vectors on the boundary of \( \mathcal{K}_\beta \) (or rather, that...
part of the boundary of $\mathcal{K}_\beta$ in $\mathcal{K}_1$) those $s_\beta$ (say) roots (vectors) for which $L^*(\mathcal{Z})$ is extremal. Denote these $s_\beta$ vectors by

$$s_\beta(\mathcal{Z}) = (\mathcal{Z}_1, \mathcal{Z}_2, \ldots, \mathcal{Z}_s).$$  \hfill (8.10)

Ideally, we would like $\mathcal{K}_M$, defined by putting $\beta = M$ in (8.6), to be a convex region such that $\mathcal{K}_\beta \subset \mathcal{K}_M$ for all $\beta \neq M$ under (8.5), for then, the problem of estimating those $\mathcal{Z}$ for which $L^*$ is optimal is narrowed to a certain bounded locale of the $\mathcal{Z}$ space, namely, on and within $\mathcal{K}_M$. For example, in Figure 8.1, $\mathcal{K}_M$ is the region formed by $\mathcal{Z}_v = C_M$ with $\mathcal{K}_\beta \subset \mathcal{K}_M$ under (8.6) for $\beta \neq M$. Certainly one may proceed with solving (8.8) and (8.9) for, say, $\mathcal{K}_M$ as the outside and boundary of a hyperellipse (so that $\mathcal{K}_M$ is closed); however, the choice of which pattern of $\mathcal{C}$'s to choose under (8.5) for substitution into (8.9) such that a representative set of compromise solutions is found may be difficult.

If the $L_V(\mathcal{Z})$ are concave functions, then each forms a convex region in the $\mathcal{Z}$ space. Since the intersection of a finite number of convex regions is convex, $\mathcal{K}_M$ is convex, and it follows that $\mathcal{K}_\beta \subset \mathcal{K}_M$ for all $\beta$ under (8.5). Then external values of $L^*$ for $\mathcal{Z}$ on and within $\mathcal{K}_M$ are estimated by varying $\mathcal{C}_\beta$ and hence $\mathcal{K}_\beta$ within $\mathcal{K}_M$. Clearly, the infinity of regions $\mathcal{K}_\beta$ "fill" the region of $\mathcal{K}_M$.

Then through the variation of $\mathcal{C}_\beta$ under (8.5), the set of solutions,

$$[s_\beta(\mathcal{Z})],$$  \hfill (8.11)
is generated which, upon substituting each $\hat{Z}$ into $L^*$, gives
the extremal values of $L^*$

(i) for $Z \in \mathcal{K}_M$ under the assumptions of the last para-
graph,

(ii) for $Z$ on the boundary of $\mathcal{K}_\beta$, otherwise.

The set,

$$
\left[ L(Z) = C_\beta, s_\beta(\hat{Z}), L^*(\hat{Z}_1), \ldots, L^*(\hat{Z}_{s_\beta}) \right], \quad (8.12)
$$

includes all compromise solutions with only one, say,

$$
\left[ L(Z) = C_0, \hat{Z}_0, L^*(Z_0) \right] \quad (8.13)
$$

being the optimal solution.

The reader is referred to Hancock (397, Chapter 6, par-
ticularly pp. 115-116) for determining whether $L^*$ attains a
maximum or minimum for any particular vector of the set
(8.11).

The general aim of the literature dealing with problems
of constrained extrema is that of finding the $Z$ vector, say
$\hat{Z}_+$, in (8.11) such that $L^*(Z)$ is extremal over all $Z \in \mathcal{K}_M$
(for example, $\hat{Z}_+$ is the vector which gives rise to $Q_1$ in
Figure 8.1), or such that $L^*(Z)$ is extremal on the boundary
of $\mathcal{K}_\beta$. Oftentimes equations (8.8) and (8.9) are considered
in a different light so as to render general theories ap-
licable. For example, the method of steepest descent (417)
is one by which $\hat{Z}_+$ is sought among those $Z$ vectors minimizing

$$
\tilde{\Upsilon} = \left[ (\partial \Upsilon / \partial Z)' , (\partial \Upsilon / \partial \lambda) \right]^T \left[ \begin{array}{c}
\partial \Upsilon / \partial Z \\
\partial \Upsilon / \partial \lambda
\end{array} \right], \quad (8.14)
$$
the solution being found by iterating along the normal to
surfaces of constant $\mathbf{v}$.

In those problems where $L^*$ is a quadratic function of
the Z's, sufficient conditions for the existence of a $\mathbf{\hat{z}}_+$
(for which $L^*$ is supremum) are given in a theorem by Frank
and Wolf [407]: If $L^*(Z)$ is a quadratic function and is
bounded on a polyhedral convex region, $\mathcal{K}_M$, then $L^*$ achieves
its supremum on $\mathcal{K}_M$. If, further, $L^*$ is concave and the con-
ditions of the Kuhn-Tucker theorem [427] are met with regard
to the constraints, then the constrained extrema problem
falls into the realm of concave programming. In this case,
solving for $\mathbf{\hat{z}}_+$ reduces to estimating the saddle point of a
Lagrangian form [427].

In the concave programming situation, $\mathbf{\hat{z}}_+$ is termed the
optimal vector among the set of vectors (8.11). This would
necessarily lead to $L^*(\mathbf{\hat{z}}_+)$ as the optimal value of $L^*$.
Though our definition and the econometric definition of op-
timality may coincide, they are in general different. With
optimality as defined earlier in this section, one is not
led to $L^*(\mathbf{\hat{z}}_0)$ or to $L^*(\mathbf{\hat{z}}_0)$ if $L^*$ is estimated directly,
but rather, a selection process must take place with respect
to (8.12) after its determination. The selection may be in-
fluenced by, say, considerations of economy, expediency, or
some combination thereof.

For instances where $L^*$ or one or more of the $L_\gamma$ are esti-
mated by least squares, the writer is unaware of any publication
using this method of solution except for one application by Umland and Smith [45].

8.2 Errors in the Optimal Solution

Errors in any of the entities of the optimal solution (8.13) will tend to be more serious than those associated with the econometric optimal solution, since the true value of (8.13) or the expected value of (8.13) if $L^*$ and some or all of the $L_\gamma$ are unknown and estimated, say, jointly by weighted least squares may not have been chosen as the optimal solution, whereas the econometric optimal solution entails no selection process.

If $L^*$ and the $L_\gamma$ are known in exact form, then the only error entering in (8.13) is through the estimation of the roots of (8.8) and (8.9) when the latter describe a system of nonlinear equations.

To illustrate this error, suppose we were to estimate the roots of the nonlinear system (8.8) and (8.9) directly, by some iterative process, without reworking them into the framework given by (8.14). Suppose further that equations (8.8) and (8.9) can be rewritten in the form

$$
\frac{p}{h + t - 1} = \begin{bmatrix} s \\ \lambda \end{bmatrix} = F(s, \lambda) = F(p),
$$

(8.15)

where $F'(p) = [F_1(p), \ldots, F_\ell(p), \ldots, F_{h+t-1}(p)]$. Then in iterating for each root of (8.15), if $p(0)$ is the initial estimate, so result $p(1), p(2), \ldots, p(0), \ldots$ by means of the recurrence relation,

$$
p^{(N+1)} = F(p^{(N)}),
$$

(8.16)
Analogous to (6.52), let

$$\bar{p}^{(N)} = \bar{\Pi} + \bar{\delta}^{(N)} ,$$  \hspace{1cm} (8.17)

where \( \bar{\Pi} \) is a root of (8.15) and \( \bar{\delta}^{(N)} \) is the error of the \( N \)th iterate. Substituting (8.17) into (8.16) and expanding the right-hand side in a Taylor's series around \( \bar{\Pi} \) results in

$$\bar{\delta}^{(N+1)}_t = \bar{\delta}^{(N)} \frac{\partial F_t}{\partial \bar{\Pi}} + \frac{1}{2} \left[ \frac{\partial (\bar{\delta}^{(N)})'}{\partial \bar{\Pi}} \right] F_t$$

$$+ \ldots + \frac{1}{r!} \left[ \frac{\partial (\bar{\delta}^{(N)})'}{\partial \bar{\Pi}} \right] F_t + o(\delta^{r+1}),$$

\( t=1,2,\ldots,h+t-1 \), where

$$\left[ \frac{\partial (\bar{\delta}^{(N)})'}{\partial \bar{\Pi}} \right] F_t = \left[ \frac{\partial (\bar{\delta}^{(N)})'}{\partial p} \right] F_t \bigg|_{\bar{p} = \bar{\Pi}}.$$

If the terms, \( \left[ \frac{\partial (\bar{\delta}^{(N)})'}{\partial \bar{\Pi}} \right] F_t \), in (8.18) vanish up to and including \( r-1 \) for all \( t \), then Hartree \( \underline{14} \) defines the iterative process "convergent to order \( r \)." Bodmer \( \underline{34} \) has shown that the generalized Newton-Raphson process is convergent to the second order; i.e., the matrix,

$$\left( \begin{array}{c} \frac{\partial F_t}{\partial p_{t'}} \\ \vdots \\ \frac{\partial F_t}{\partial p_{t'}} \\ \end{array} \right) \bigg|_{p_{t'} = \Pi_{t'}} = 0,$$

\( t,t'=1,2,\ldots,h+t-1 \), when \( F(p) \) is defined to be

$$F(p) = p - \left[ \frac{\partial (\bar{\delta}^{(N)})'}{\partial p_{t'}} \right]^{-1} \left[ \frac{\partial (\bar{\delta}^{(N)})'}{\partial p} \right] \hspace{1cm} (8.19)$$

where \( (\partial^{(N)} / \partial p)' = \left[ (\partial^{(N)} / \partial \bar{Z})' (\partial^{(N)} / \partial \lambda)' \right] \) in (8.8) and (8.9). Note that from (8.19), the equations in (8.15) are equivalent to (8.8) and (8.9).
If in (8.19), $O(\delta^2)$ is negligible, then the recurrence relationship,

$$p^{(N+1)} = p^{(N)} - \left[ \frac{\partial \psi / \partial p}{\partial p^{(N)}} \right]^{-1} \left[ \frac{\partial \psi}{\partial p^{(N)}} \right],$$

(8.20)

is used to estimate the roots of (8.19) or (8.15), and hence of (8.8) and (8.9). Unless $O(\delta^2)$ is sufficiently small, convergence (to a given decimal place) to the roots of (8.15) is only "seemingly." That is, we want to find the roots of one system of equations, (8.8) and (8.9), but find it more convenient to work with a second system, (8.19) or (8.15) (whose roots are the same as the first system), since the latter have associated systematic, iterative scheme, (8.20).

But in the iterative scheme a quantity, $O(\delta^2)$, is assumed negligible, and if the latter is, in fact, substantial, then the roots obtained may differ greatly from those of (8.19), thus resulting in pseudo roots. However, by neglecting terms of, say, $O(\delta^3)$ or $O(\delta^4)$ in (8.18) rather than $O(\delta^2)$ and/or obtaining a closer initial estimate of $\Pi$, the pseudo roots may be corrected.

When $O(\delta^2)$ is sufficiently small, the error may be negligible so long as one iterates a sufficient number of times, especially when the roots of (8.19), and consequently of (8.8) and (8.9), are close. Then it could be concluded that the true value of (8.13) would have been chosen as the optimal solution.

Consider next the case when $L^*$ and some or all of the $L_\gamma$ are unknown in exact form and must be estimated by, say,
\( L^* \) and \( L_\gamma \). Here we consider those cases for which the \( L \) can be estimated by least squares (or by weighted least squares when the errors within the experimental units are correlated), as is described in Chapter 6; i.e., \( y_j = f_j(Z) \) in (8.4) can be adequately approximated by a Taylor's expansion up to a certain order. In this case, the error associated with each member of the optimal solution may be real and therefore cannot be discounted as previously, since besides the error of iteration, we have the errors stemming from the estimation of the \( L \)'s. As such, it is not unlikely that the expected value of the optimal solution,

\[
\left\{ E\left( \hat{L}(Z) - L_0 \right), E\left( \hat{Z}_0 \right), E\left[ \hat{L}^*(\hat{Z}) \right] \right\}, \tag{8.21}
\]

may not have been chosen as the optimal solution, unless (8.13) and (8.21) are sufficiently close. But there exist no means of measuring this closeness aside from the narrowness of associated confidence bounds for each member of (8.21) (assuming the existence of unbiased statistics), which are difficult, if not impossible, to construct from \( E(\hat{Z}_0) \), except in possibly very simple problems.

Regardless of the usual indetermination of closeness just mentioned, there are situations where there is an especially high risk of choosing the optimal solution when, in fact, the expected value of the optimal solution would not have been desired. This occurs when \( \hat{L}^*(\hat{Z})_0 \) is at a point on the surface of \( \hat{L}^* \) where the curvature is large, for then, small changes in \( \hat{Z} \) produce large changes in \( \hat{L}^*(\hat{Z}) \). Further, for cases where \( \hat{L}^* \)
is neither concave nor convex and/or the \( \hat{L}_\gamma \) are not concave, the latter mentioned risk is high since small changes in \( C_\beta \) could produce large changes in \( \hat{Z} \). If in the immediate neighborhood of \( \hat{Z}_0 \) and \( C_0 \) the distance, \( \left[ (\hat{Z}_\beta^0 - \hat{Z}_\beta^1, \hat{Z}_\beta^1 - \hat{Z}_\beta^0) \right]^{1/2} \), is large compared with any of the distances, \( |C_\beta - C_\beta'| \), where \( \beta \neq \beta' \), \( \hat{Z}_\beta^0, < \hat{Z}_0 < \hat{Z}_\beta^1 \), and \( C_\beta, < C_0 < C_\beta \), then one should consider a sufficient number of interpolated values of \( C_\beta \) in the neighborhood of \( C_0 \) before finally selecting the optimal solution.

In passing, it is worth noting that for \( \hat{L}^* \) roughly planar over the constraint region, one might approximate \( \hat{L}^* \) by a plane (hyperplane) and simplify the computations considerable.

8.3 An Approximate Confidence Interval for an Individual Lagrangian Multiplier

Suppose that the q response types are in one of two units of measure; i.e., \( t=2 \). Let \( \hat{L}^*(Z) \) be a linear combination of \( q' \) linear regression functions in one unit of measure and \( \hat{L}(Z) \) a linear function of the \( q-q' \) linear regression functions in the other unit of measure. Then (8.7) becomes

\[
\psi(Z, \lambda) = \hat{L}^*(Z) - \lambda[\hat{L}(Z) - C].
\]

(8.22)

Assuming there exists at least one value of \( Z \) on the boundary defined by \( L(Z) = C \) for which \( L^* \) is extremal, then this point (points) is found by setting

\[
\frac{\partial \psi}{\partial Z} = (\frac{\partial \hat{L}^*}{\partial Z}) - \lambda(\frac{\partial \hat{L}}{\partial Z}) = 0,
\]

(8.23)
and \( L(Z) = C \). The \( t^{th} \) equation of (8.23) is

\[
\hat{\lambda} = \left( \frac{\partial L}{\partial Z} \right) \bigg/ \left( \frac{\partial L}{\partial Z} \right) = \frac{\hat{k}'_{t1} \hat{b}_{t1} / \hat{k}'_{t2} \hat{b}_{t2}}{I},
\]

(8.24)

where \( \hat{b}' = (\hat{b}_{t1}, \hat{b}_{t2}) \), \( \hat{b}_{t1} = (\hat{b}_{1}, \ldots, \hat{b}_{q}) \), \( \hat{b}_{t2} = (\hat{b}_{q+1}, \ldots, \hat{b}_{q}) \), and \( \hat{\lambda} \) is the appropriate estimate of \( \lambda \). From (8.24), \( \hat{\lambda} \) may also be written in the form,

\[
\hat{\lambda} = \left[ \frac{\hat{b}_1}{\hat{b}_2} \right] = \left[ \frac{\hat{k}'_{t1} (\hat{b})}{\hat{k}'_{t2} (\hat{b})} \right] (8.25)
\]

By a proper multiplication and rearrangement in (8.24), where \( \hat{k}'_{t1} (\hat{b}) \) implies that the elements of \( \hat{k}'_{t1} \) and \( \hat{k}'_{t2} \) depend on \( \hat{b} \). Clearly, these dependencies are in nonlinear forms, and except in trivial cases, it is difficult, if not unfeasible, to obtain the \( \hat{k}'_{t} \) in explicit forms. With \( \lambda = \hat{k}'_{t} (\hat{b}) \), it may be evident to the reader at this point that we intend to use the argument of Section 7.4 in finding a confidence interval for \( \lambda \), since in (8.25) \( \hat{k}'_{t} \) is functionally dependent on the data which is implied by \( \hat{k}'_{t} (\hat{b}) \).

Now, if \( X_j \neq X \) for some \( j=k, 2, \ldots, q \), then \( \text{var} (\hat{b}) = V \sigma^2 \), where \( V \) is given in (7.5) and discussed thereafter; i.e.,

\[
V = \left[ \hat{X}' (I \ast W)^{-1} \hat{X} \right]^{-1} (8.26)
\]

\[
\sum_{j=1}^{q} p_j \times \sum_{i=1}^{q} p_i
\]

if \( W \) is known and

\[
V = (I + G_2) \left[ \hat{X}' (I \ast \hat{W})^{-1} \hat{X} \right]^{-1} (I + G_2) \quad \text{if } W \text{ is unknown (assuming convergence).}
\]
If \( k \) is a preselected non-null vector,

\[
k'\hat{b} : N(k'b, k'\nu k\sigma^2).
\]  

(8.28)

Then using the same argument given in Section 6.4, we find, similar to (7.24) and (7.25) that

\[
\sup_k \left[ k'(\hat{b} - b) \right]^2 / \hat{\sigma}^2 k'\nu k = (\hat{b} - b)'V^{-1}(\hat{b} - b) / \hat{\sigma}^2
\]

(8.29)

is distributed as \( \rho_1 F_{\alpha}(f_1, f_2) \) where \( F_{\alpha} \) is the upper \( \alpha \) percent point of the \( F \) distribution with \( f_1 = \sum_{j=1}^{q} p_j, \ f_2 = qn - (\sum_{j=1}^{q} p_j) + q \), \( \rho_1 = f_1 / f_2 \), and \( \hat{\sigma} \) is the proper estimate of \( \sigma \). Then

\[
P \left[ \frac{k'(\hat{b} - b)^2}{k'\nu k \hat{\sigma}^2} \leq \rho_1 F_{\alpha}(f_1, f_2) \right] = 1 - \alpha
\]

(8.30)

for all non-null \( k \). Now, the statement,

\[
\left[ k'(\hat{b} - b)^2 \right] \leq \rho_1 F_{\alpha}(f_1, f_2) k'\nu k \hat{\sigma}^2,
\]

(8.31)

for all non-null \( k \) and with probability \( 1 - \alpha \), implies that

\[
\left[ k'(\hat{b} - b)^2 \right] \leq \rho_1 F_{\alpha}(f_1, f_2) \hat{\sigma}^2 \sup_k (k'\nu k)
\]

(8.32)

subject to \( k'k \) constant and with probability \( \geq 1 - \alpha \); i.e.,

(8.31) \( \Rightarrow \) (8.32), but (8.32) \( \Rightarrow \) (8.31). But subject to \( k'k \) constant, \( \sup (k'\nu k) = \text{ch} \ V \) so that (8.32) implies

\[
\max \left[ k'(\hat{b} - b)^2 \right] \leq \rho_1 F_{\alpha}(f_1, f_2) \hat{\sigma}^2 \text{ch} \ V
\]

(8.33)

subject to \( k'k \) constant and with probability \( \geq 1 - \alpha \); i.e.,

(8.32) \( \Leftarrow \) (8.33). Since (8.32) and (8.33) hold for all non-null \( k \) subject to \( k'k \) constant, we may certainly choose
\[ k = \tilde{K}_\ell(\hat{b}) \] so that \( \tilde{K}_\ell(\hat{b})[\hat{b} - b] = \lambda_0 - \tilde{K}_\ell(\hat{b})b = \lambda - \lambda_0 \). Hence from (8.33),
\[
(\lambda - \lambda_0)^2 \leq \rho_1 F_{\alpha}(f_1, f_2) \sigma^2 \text{ch } V \text{ with probability } 1 - \alpha \text{ so that } \max
\]
\[
P \left[ \hat{\lambda} - \rho_1 F_{\alpha}(f_1, f_2)^{1/2} \sigma \text{ch } V \right] \leq \lambda_0 \leq \hat{\lambda} \leq \max
\]
\[
+ \rho_1 F_{\alpha}(f_1, f_2)^{1/2} \text{ch } V \right] \leq 1 - \alpha
\]
(8.34)
defines an exact confidence region for \( \lambda_0 = \tilde{K}_\ell(\hat{b})b \). If we assume that \( \lambda_0 = \lambda \) and replaced \( \lambda_0 \) by \( \lambda \) in (8.34), then the (8.34) will define an approximate confidence interval for \( \lambda \).

If \( X_j = X \) for all \( j \), consider the regression of \( t'y \) on \( x \), so that the resultant coefficient vector estimate of, say, \( \hat{c} \), is \( \hat{c} = t_1 \hat{b}_1 + t_2 \hat{b}_2 + \ldots + t_q \hat{b}_q \), and, from previously, \( \text{var}(\hat{c}) = (X'X)^{-1}t's't \). Then in (8.25) we will have \( \hat{\lambda} = \tilde{K}_\ell(\hat{c})\hat{c} \), so that following through with the entire argument just given, one arrives at
\[
P \left[ \hat{\lambda}^* - \rho_1 F_{\alpha}(f_1^*, f_2^*)^{1/2} (t^s_\ell s_\ell) \right. \leq \lambda_0 \leq \hat{\lambda}^* \leq \max
\]
\[
+ \rho_1 F_{\alpha}(f_1^*, f_2^*)^{1/2} (t^s_\ell s_\ell) \right] \geq 1 - \alpha,
\]
(8.35)
where \( f_1 = p, f_2 = q, f_1^* = f_2^* = q, \rho_2 = f_1^*/f_2^* \), and \( \lambda_0 = \tilde{K}_\ell(\hat{c})\hat{c} \). As before, if we assume \( \lambda^* = \lambda^* = \tilde{K}_\ell(\hat{c})\hat{c} \) and replace \( \lambda^* \) by \( \lambda^* \) in (8.35), then there results an approximate confidence interval for an individual Lagrangian multiplier, \( \lambda^* \), whose estimate is \( \hat{\lambda}^* \).

Since the determination of \( \text{ch } V, \text{ch } (X'X)^{-1} \) may involve \( \max \) \( \text{ch } V, \text{ch } (X'X)^{-1} \) may involve \( \max \) tedious computations should either of these matrices be of large order, one may use the following approximation. The
The largest characteristic root of a matrix is less than the largest sum of the moduli of the terms in any row or column.

Taking this value as the largest characteristic root, one might at the same time increase the value of $\kappa$ by some amount which reduces $F_\kappa$ to counteract the possible increase in $\text{ch} \ V$ or $\text{max} \ \text{ch} \ (X'X)^{-1}$ due to its approximation.

$\text{max}$
9.0 SUMMARY AND GENERAL REMARKS

Concepts in various parts of this article have long been known and applied. However, to the writer's knowledge, no one has considered various aspects of Chapter 6. Its purpose is to allow the construction of an adequate system of linear regressions in order to cope with problems dealing with certain relationships between dependent response types.

First, a phenomenon of interest (exudate extracted from inoculated mice) is characterized by q response types each of which are, in turn, expressed in terms of other variables. The latter includes certain exogenous variables (which are assumed to be at least controlled) and certain endogenous variables (the response types). The reduced equations or predictive regression system is assumed to include only linear models depending on the same or different sets of overlapping independent variables; e.g., if $x_1$ and $x_2$ denote the exogenous variables, then the response type $y_j$ is regressed on the independent variables, say, $x_1, x_2, x_1^2, x_1x_2,$ and $x_2^2$. While another response type, $y_j$, is regressed on, say, $x_1$ and $x_1^2$. It is stressed that while the hypothesized structural regression system may be incorrect, the predictive regression system may lead to valid inferences.

Should each $y_j$ be regressed on precisely the same set of independent variables, then $\Sigma$, the covariance matrix of the $q$ errors stemming from an experimental unit, does not enter in the joint least squares estimation of the parametric
coefficients of the regression system. If $\Sigma$ is known (or at least the simple correlations and the ratios of the diagonal elements of $\Sigma$ are known), and the $y_j$ depend on different sets of overlapping independent variables, then best estimates of the coefficients are available. Otherwise ($\Sigma$ unknown), an iterative scheme is proposed for finding joint estimates of the coefficients. Further, an approximate covariance matrix is given for the iterated vector estimate. In this connection, a computer study should now be made to determine whether the approximate covariance matrix is a realistic measure of dispersion arising in small samples.

Then linear functions, say $\theta(\hat{y})$, of the predicted values, $\hat{y}_j$ (in the same or comparable units of measure), are taken with the objective being the estimation of the points, say $\hat{Z}$, in the functionally independent factor space such that $\theta[E(\hat{y})]$ attains stationary points, assuming the existence of at least one. Should $\theta(\hat{y})$ be of second order in the $Z$'s, then exact and approximate confidence regions are given for $E(\hat{Z})$.

Finally, if the $\hat{y}_j$ are in different units of measure, functions of certain $\hat{y}_j$ are maximized (minimized) subject to varied constraints, the latter being defined by other $\hat{y}_j$.

It is unfortunate that the iterative methods and approximate covariance matrix given in Chapter 6 are not applicable to all single nonlinear models such as the "revised" rational regression model of Section 6.7. Nonlinear regression models are gaining increasing prominence and rightly so. The real world is not linear, but rather nonlinear and by referring to
realistic nonlinear models, one is, in effect, coming closer to nature. Further work is needed to determine the worth of parameter estimates of nonlinear models since large sample variances tend to underestimate the actual dispersion. On the other hand, the more basic problem lies in estimation. Nonlinear estimation by known methods oftentimes requires some type of iteration while iteration, in turn, may induce a large variability in the estimates. Perhaps at a future date, some method of estimation will be devised voiding the "necessity" of iteration in nonlinear estimation.

The theorem by Scheffe \( \hat{\Sigma} \) was used in finding out only an approximate confidence region for \( E(\hat{\Sigma}) \), but also an approximate confidence interval for an individual Lagrangian multiplier. Though no other recourse is usually available with such a sound mathematical basis, the application of this theorem presupposes nature to react at her worse, while, in fact, she does not always do so or may rarely do so. Consequently, to obtain a relatively simple solution to a very complex problem, one is forced to become ultra-conservative. With regards to needed research, few statistical problems hold such high priority as the one dealing with overprotection. In other words, when applying the aforementioned theorem, to what extent should \( \Sigma \) be increased (see Section 8.3) in order that realistic confidence intervals (regions) and tests of significance are attained.
10.0 LIST OF REFERENCES


292. Bhattacharya, P. K. Some properties of the least square estimator in regression analysis when the 'independent' variables are stochastic. June, 1961.