A MODIFIED APPROACH TO SMALL AREA ESTIMATION

by

Steven B. Cohen

Department of Biostatistics
University of North Carolina at Chapel Hill
Institute of Statistics Mimeo Series No. 1167

April 1978
A MODIFIED APPROACH TO SMALL AREA ESTIMATION

by

Steven B. Cohen

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

Chapel Hill

1978

Approved by

Co-Advisor

Co-Advisor

Reader
ABSTRACT

STEVEN BRUCE COHEN. A Modified Approach to Small Area Estimation. (Under the direction of Gary G. Koch and William D. Kalsbeek.)

The ever-growing need for good estimates of the health, social political, and economic parameters of local areas has served as the motivating force for new developments in methodology. Due to the constraints of sample size, design, and cost, accessible data from large areas for criterion variables of interest is often used jointly with local data on symptomatic variables. Furthermore, several procedures have derived local area estimators by combining symptomatic information and sample data into a multiple regression format. In those situations where assumptions are too strict or unrealistic, as when a non-linear model is more appropriate, the merits of a more flexible approach are obvious.

Our research focuses upon a further investigation of an alternative strategy for which the most limiting assumption is the availability of good symptomatic information. A more formal representation of the model is developed within the framework of a post-stratification scheme. The methodology involves ratio estimation of the respective stratum means via indicator variables which serve the purpose of classification.

To determine the accuracy of the proposed small area estimator and allow for comparisons of precision with respect to other strategies, we express the relationship between criterion and symptomatic variables by relevant continuous multivariate distributions. Specifically, comparisons are made with the results obtained using a regression estimator
which is applicable to the same general setting. The theoretical framework considers multivariate stratification, where boundary determination is achieved by application of practical methods which use minimum variance stratification as a criteria.

An application of the proposed strategy to real data is also given. Using the Current Population Survey as the source of sample data, we generate state estimates of the 1970 unemployment rate and examine their reliability.
to Paula

and my parents
ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my adviser Dr. Gary Koch for the guidance, support, and direction he has provided during the course of my graduate studies. I am most grateful to him for stimulating my interest in the problem area of the dissertation.

I am deeply indebted to my co-adviser, Dr. William D. Kalsbeek, for his direction, support, and personal involvement in the preparation of the dissertation. It was his previous work in the area that served as the motivation for my research.

Special thanks go to Dr. H.B. Wells, Dr. D.G. Kleinbaum, and Dr. K. Bauman for their participation on my doctoral committee and for their helpful comments, suggestions and encouragement.

I wish to express my heartfelt thanks to Paula, my wife, whose patience, support, encouragement and love made the difficult times endurable.

In addition, I wish to thank the Department of Biostatistics for the opportunity to do graduate work in their demography program. I am particularly grateful for the financial support I have received from the National Institute of Child Health and Human Development (Grant HD-00371) and the U.S. Bureau of the Census (J.S.A.-76-93).

Finally, I wish to thank Jean McKinney for her cheerful assistance and excellent job of typing this manuscript.
TABLE OF CONTENTS

LIST OF TABLES ........................................... Page
LIST OF FIGURES ...........................................

I. INTRODUCTION AND REVIEW OF THE LITERATURE

1.1 Introduction ........................................ 1
1.2 An Additive Model ................................ 3
   1.2.1 An Extension of the Simulmatics Project Technique ................................ 3
   1.2.2 Validation Procedures and Applications ................. 5
   1.2.3 Methodological Problems .......................... 6
1.3 A More General Approach .............................. 8
   1.3.1 The NCHS Model .................................. 8
   1.3.2 Other Methodologies ................................ 10
   1.3.3 Measures of Reliability .......................... 12
   1.3.4 Present Limitations ............................... 16
   1.3.5 Potential Improvements ........................... 17
1.4 Contingency Table Synthetic Estimation .............. 18
1.5 Methods Combining Sample Survey Data and Symptomatic Indicators ............ 21
   1.5.1 A Regression Estimate Approach .................. 21
   1.5.2 A More Complicated Design ........................ 24
   1.5.3 A Multiple Regression Framework .................. 24
   1.5.4 An Alternative Strategy ......................... 28

II. ESTIMATION USING THE KALSBEEK MODEL

2.1 Introduction ........................................ 35
2.2 Notation ............................................. 36
2.3 Representation of the Model ........................... 38
2.4 An Expression for the Mean Squared Error of the Local Area Estimator .......... 42
   2.4.1 Proof for Cov\left( \hat{y}_g, \hat{y}_g \right) \neq 0 ...................... 44
III. A REFORMULATION OF THE KALSBEEK MODEL: SOME ANALYTICAL AND EMPIRICAL INVESTIGATION

3.1 Introduction .............................................. 47
3.2 Determination of Stratum Boundaries ................. 47
3.3 A Reformulation of the Kalsbeek Model ............... 52
3.4 The Theoretical Framework ............................... 55
   3.4.1 Determination of the Bias ....................... 58
3.5 Estimation Using the Ericksen Model .................. 59
3.6 Target Area Estimation for the Trivariate Normal Distribution ............................................. 61
3.7 Target Area Estimation for the Trivariate Logistic Distribution ............................................. 67
3.8 Distribution Specific Results ......................... 75

IV. AN APPLICATION OF THE PROPOSED METHODOLOGY: SMALL AREA ESTIMATION OF UNEMPLOYMENT

4.1 Introduction .............................................. 90
4.2 A Description of the Sampling Framework and the General Methodology ............................................. 92
   4.2.1 The Current Population Survey: A Source for Sample Data ............................................. 92
   4.2.2 Determination of the Stratum Boundaries ........... 93
   4.2.3 The Estimation Procedure .......................... 94
4.3 State Estimates of Unemployment: An Application of the Proposed Strategy and the Ericksen Model ................. 96
4.4 An Interpretation of the Results ......................... 97

V. SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

5.1 Introduction .............................................. 105
5.2 Summary .................................................. 105
5.3 Conclusions ............................................. 107
5.4 Suggestions for Further Research ....................... 109

APPENDIX A. Explicit Expressions for First and Second Order Moments of the Doubly Truncated Bivariate Normal Distribution ............................................. 110

REFERENCES .................................................. 117
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Target Area Estimation for Trivariate Normal Distribution with $R \geq .95$, Range (.20, .80)</td>
<td>81</td>
</tr>
<tr>
<td>3.2</td>
<td>Target Area Estimation for Trivariate Normal Distribution with $R \geq .58$, Range (.20, .80)</td>
<td>82</td>
</tr>
<tr>
<td>3.3</td>
<td>Target Area Estimation for Trivariate Logistic Distribution (corresponding to $R \approx .58$), Range (.20, .80)</td>
<td>83</td>
</tr>
<tr>
<td>3.4</td>
<td>Target Area Estimation for Trivariate Normal Distribution with $R \geq .95$, Range (.05, .95)</td>
<td>84</td>
</tr>
<tr>
<td>3.5</td>
<td>Target Area Estimation for Trivariate Normal Distribution with $R \geq .58$, Range (.05, .95)</td>
<td>85</td>
</tr>
<tr>
<td>3.6</td>
<td>Target Area Estimation for Trivariate Logistic Distribution (corresponding to $R \approx .58$), Range (.05, .95)</td>
<td>86</td>
</tr>
<tr>
<td>3.7</td>
<td>Target Area Estimation for Trivariate Normal Distribution with $R \geq .95$, Range (.35, .95)</td>
<td>87</td>
</tr>
<tr>
<td>3.8</td>
<td>Target Area Estimation for Trivariate Normal Distribution with $R \geq .58$, Range (.35, .95)</td>
<td>88</td>
</tr>
<tr>
<td>3.9</td>
<td>Target Area Estimation for Trivariate Logistic Distribution (corresponding to $R \approx .58$), Range (.35, .95)</td>
<td>89</td>
</tr>
<tr>
<td>4.1</td>
<td>Relative Absolute Deviation of the 1970 Unemployment Rate Estimates for States</td>
<td>101</td>
</tr>
<tr>
<td>4.2</td>
<td>Relative Absolute Deviation of the 1970 Unemployment Rate Estimates for States</td>
<td>103</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.1</td>
<td>Classification of Base Units by Strata and Local Areas</td>
<td>38</td>
</tr>
<tr>
<td>3.8.1</td>
<td>Stratum Boundaries Using Hierarchical Scheme</td>
<td>78</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Determination of Stratum Boundaries Using a Hierarchical Scheme with the CUM/T Rule</td>
<td>98</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION AND REVIEW OF THE LITERATURE

1.1. Introduction

The ever-growing need for good estimates of the health, social, political, economic and demographic parameters of local areas has been rapidly gaining recognition. The allocation of federal aid to both states and municipalities is often dependent upon information pertaining to population, unemployment, median income and residential segregation. Candidates vying for political office are particularly concerned with obtaining reliable estimates of voter preference and participation at the subnational level. Similarly, rather precise small area estimates of retail trade are essential indicators for the commercial sector.

Some useful information has been obtained from sources which include the decennial census and vital registration systems. Generally, federal agencies have relied upon sample surveys to provide estimates of the data they require, though such estimates usually pertain to the United States as a whole or to each of its four broad geographical regions. Estimates of data for small areas are unavailable primarily due to sample size requirements, which are prohibitive with respect to cost and strata designs which often cross state and county limits. Still, the demands at the state, county and municipal level for reliable estimates persists.

Several procedures have been developed which utilize available
data from large areas, local data on population, and accessible local data on ancillary (symptomatic) variables, in order to produce the desired small area estimates. Synthetic estimation is perhaps the most well-known, defined by the United States Bureau of the Census as "the method of reference to a standard national distribution." Gonzalez (1974) has offered a more comprehensive explanation -- "An unbiased estimate is obtained from a sample survey for a large area; when the estimate is used to derive estimates for subareas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates." In spite of the strict limitations usually placed upon their application to insure validity, the temptation to accept these "synthetic estimates" inevitably prevails. As is often the case, the limitations of use are exceeded and the estimates inappropriately considered.

A substantial amount of the developmental work in this area has been forthcoming from group efforts by the Simulmatics Project under the direction of Pool et al. (1965), the National Center for Health Statistics, and the Bureau of the Census. Notable individual contributions include those made by Woodruff (1966), Levy (1971), Weber (1972), Gonzalez (1973, 1974), Ericksen (1973), Kalsbeek (1973), Koch (1975), and Seidman (1975). There is reason to suspect, however, that the level of common knowledge among these researchers can be improved. It is the intent of this chapter to survey the state of the field, review the statistical properties of the respective small area estimators, and consider their proper applications and limitations.
1.2. An Additive Model

1.2.1. An Extension of the Simulmatics Project Technique

Under the auspices of the Simulmatics Project, Pool et al. (1965) have produced a popular technique to generate synthetic estimates for local areas. Their major concern was to simulate public opinion for individual states by a transformation of national survey data to subnational units. Although its application to public opinion yields estimates of only those characteristics expressed as percentages, the technique's framework lends itself to more general usage.

Procedurally, a number of demographic variables are selected (i.e., region, race, income, sex, age), and when possible, national sample surveys are used to determine estimates of a characteristic of interest for each of the G mutually exclusive and exhaustive domains defined by the respective variable cross-classifications. In practice, this method assumes that estimates for a particular domain can be calculated by simply summing the separate effects of the particular variable classifications that define the domain. As a consequence, the estimate of the characteristic of interest for the $j^{th}$ domain is defined by Weber et al. (1972) as:

$$R_j = A + \sum_{z=1}^{m} (S_z - A)_j$$

(1.2.1)

where $A$ is the national estimate of the characteristic derived from a national survey, $S_z$ is the mean value of the characteristic defined by a single variable classification and $m$ is the number of variables defining a domain.

This model employed to produce synthetic estimates of a characteristic $X^*$ (expressed as a percentage) for subarea $i$ takes the form of
a weighted average:

\[ X^*_i = \sum_{j=1}^{G} P_{ij} R_j \]  \hspace{1cm} (1.2.2)

where \( G \) is the number of mutually exclusive and exhaustive domains defined by all possible variable combinations and \( P_{ij} \) is the proportion of subarea \( i \)'s population represented by domain \( j \) so that \( \sum_{j=1}^{G} P_{ij} = 1 \).

Seidman (1975) notes that \( R_j \) may be rewritten as:

\[ R_j = \left( \sum_{z=1}^{m} S_z \right)_j + (1-m)A \]  \hspace{1cm} (1.2.3)

where \( \left( \sum_{z=1}^{m} S_z \right)_j \) indicates summation over the \( m \) defining variable classifications for domain \( j \). Therefore, \( X^*_i \) may be redefined as:

\[ X^*_i = (1-m)A + \sum_{j=1}^{G} P_{ij} \left( \sum_{z=1}^{m} S_z \right)_j \]  \hspace{1cm} (1.2.4)

Expanding the \( \left( \sum_{z=1}^{m} S_z \right)_j \), collecting terms and setting the constant \( (1-m)A = C \) yields:

\[ X^*_i = C + (P_{ib} + \ldots + P_{iz})S_1 + (P_{ib} + \ldots + P_{iy})S_2 \]

\[ + \ldots + (\ldots)S_t \]  \hspace{1cm} (1.2.5)

where \( t \) is the total number of distinct variable classifications. Within the parentheses are those population proportions of each of the domains which show the variable classification defining the \( S_z \)'s and sum to the proportion of the population having characteristic \( S_z \). By defining that proportion as \( q_{ia} \) for subarea \( i \) and variable classification \( a \), (1.2.5) reduces to

\[ X^*_i = C + q_{i1}S_1 + q_{i2}S_2 + \ldots + q_{it}S_t \]  \hspace{1cm} (1.2.6)
In essence, the assumption of no interaction between the defining variable classification inherent in the formulation of the $R_j$'s allows the synthetic estimate to reduce to an additive linear function of the $S_z$'s in (1.2.6).

1.2.2. Validation Procedures and Applications

In practice it is not unusual to note an inordinate number of domains defined by the respective variable cross-classifications. When national survey data is used directly, several domains will inevitably depend on information derived from a limited number of respondents, and thus, the estimates of $R_j$ may be relatively unstable. An additive model which produces reliable estimates and avoids the pitfalls of inadequate sample size is indeed a tempting alternative. Yet, before one can use the model to derive synthetic estimates of unknown parameter values with any degree of confidence, a number of rigid validation procedures must be satisfied. First and foremost, the assumption of additivity must be realistic. In addition, the model must be uniformly realistic for application to a variety of parameters. In other words, small area estimates derived from the model should consistently provide the best possible fit with actual data.

To their credit, Pool et al. raise the question of whether or not predispositional factors summate. To determine if any significant interactions exist, an analysis of variance framework was considered. Here, attitude for respective voter-type was defined as the dependent variable. It was concluded that the assumption of additivity firmly held for non-party dominated issues, though party identification interacted systematically with the other background factors for the remaining issues. Fur-
thermore, no evidence was present for two factors other than party to interact over all issues. This gave a degree of substance to the generalization that social and demographic factors are additive by nature for large portions of public opinion data.

Rather than test the assumption of additivity more rigorously by themselves, Weber et al. were willing to accept the results advanced by Pool et al. To add substance to this position, the findings of Burnham and Sprague (1970) regarding voting behavior of Pennsylvanian counties were cited. In essence, this was a confirmation on the aggregate level that electoral alignments under stable phenomenon would best be described by an additive model.

The same method has been used by Schneider (1973) to estimate aggregate opinion in small political units. Kim et al. (1975), using a variation of the technique, have derived state estimates of voter turnout; the only predispositional factors they considered were race, age, income, and education. Hinckley (1970) has used the synthetic estimates of statewide party identification derived by Weber to determine the effect of incumbency and presidential vote in Senate elections. Similarly, Weber's estimates of several policy questions has been accepted by Sutton (1973) to evaluate the responsiveness of state government.

1.2.3. Methodological Problems

One is naturally excited by the attractions of the technique. Too often there is a tendency to avoid raising additional questions regarding validity and applicability. Hence, a more critical assessment is warranted.
A major source of concern is the choice of a database. Pool et al. have collected information on "issue clusters" from national cross-section and probability samples. Rather than use only the most recent data in their analysis, they created a five period cumulative database by combining information from the 1952, 1954, 1956, 1958 and 1960 election surveys. Their rationale was that "the benefits of large numbers and broadly based coverage proved to be greater than the benefits of timeliness." Their approach was further confounded by combining the results of several surveys addressing related topics. Both considerations are problematic, suggesting a re-evaluation of the true benefits acquired from a cumulative database.

The impact of inaccurate reporting and other sample survey errors must be isolated, quantified, and treated accordingly. Weber points to the 1960 Survey Research Center's survey, which under-reported national abstention by 15.3 percentage points, as being by no means an exceptional case. Available census data is occasionally used to adjust results. In general, the survey findings hold, though survey error is viewed as a source of the unexplained variation in the simulations.

The greatest source of unexplained variation in the simulations is method error. Perhaps the underlying assumption of additivity upon which this method stands may be a contributing factor. Pool et al. rejected their findings that for "party dominated issues, there is a strong tendency for party identification to interact systematically with other background factors." Soares and Hamblin (1952) noted that "the variation in voting for candidates or parties (i.e., Allende in Chile) should be predicted more completely by multiplicative (interactive) rather than linearly additive models involving a set of socio-economic predictor
variables." Burnham and Sprague (1970) have stated that significant
evidence of interaction among predisposing variables will be found during
periods of ideological polarization and re-aligning elections. Hence,
third party movements outside the leadership organizational structure
are more appropriately analyzed by multiplicative models.

As previously mentioned, simulations using A IPO (American Institu-
tute of Public Opinion) or SRC (Survey Research Center) surveys are
often hard-pressed to produce estimates of parameter values by relying
solely on the few observations falling within each of the respective
domains defined by variable cross-classifications. A reliable assump-
tion of additivity is a welcome alternative. But when its legitimacy
remains in doubt, the results obtained must be dealt with accordingly.

1.3. A More General Approach

1.3.1. The NCHS Model

The technique previously described was quite specific as to the
form synthetic estimates must assume. The merits of a less restrictive
procedure are obvious. A method that produces small area estimates of
desired health parameters has already been developed at the National
Center for Health Statistics (1968). It was initially used to provide
synthetic state estimates of disability from the results of the National
Health Interview Survey (HIS).

The model accepted at NCHS strikingly resembles that defined in
Section 1.2.1. by Weber et al. for estimates of characteristic X for
subarea i. It takes the form of the weighted average described by
Gonzalez (1974) as:
where $G$ and $P_{ij}$ are defined as in (1.2.2) and $X_{.j}$ is a probability estimate of characteristic $X$ for domain $j$ obtained from a national sample.

Two pronounced differences exist, however, between the respective methods. In (1.2.2) $R_j$ is always determined as an additive function of the separate effects of each distinct defining variable classification. Contrarily, $X_{.j}$, which corresponds to $R_j$, must always be directly estimated from a national survey. When the $G$ cross-classification domains are so numerous as to require sample sizes prohibitive with respect to cost in order to determine $X_{.j}$, the domains are collapsed into $G'$ subgroups, for which reliable estimates can be made. In addition, the method is not limited to the estimation of characteristics expressed as percentages, including those in terms of rates, ratios, and averages.

The more detailed estimating equation (referred to as NCHS Method II), which includes a regional adjustment, takes the form:

$$X_{Ri}^{**} = \left[ \frac{\sum_{j=1}^{G} Y_{ji} \frac{X_{.j}}{Y_{.i}}}{Z_{.i}} \right] \left[ \frac{X_{R}'}{\sum_{i=1}^{R} X_{i}' P_{i}} \right]$$

(1.3.2)

where $X_{Ri}^{**}$ = estimate of $X$-characteristic for the $i^{th}$ subarea (here, state);

$Y_{ji}$ = the most recent Census population in the $i^{th}$ subarea in the $j^{th}$ domain;

$Y_{.i}$ = $\sum_{j=1}^{G} Y_{ji}$ = most recent Census population of the $i^{th}$ subarea;

$Z_{.i}$ = current independent estimate of civilian non-institutional population in the $i^{th}$ subarea;
\[ X_{ij} = \frac{\sum_{n,m} W_{mn} X_{mnj}}{\sum_{n,m} W_{mn}} \quad \text{usual HIS final estimate for the } j^{th} \text{ domain over all the United States;} \]

\[ W_{mnj} = \text{the HIS final weight for the } mn^{th} \text{ person;} \]

\[ Z_{mnj} = 1 \text{ if the } mn^{th} \text{ person is in the } j^{th} \text{ domain, 0 otherwise;} \]

\[ \frac{X_{ji}}{\sum_{i} X_{ji} p_i} = \text{ratio adjustment factor at the regional level,} \]

\[ x_{R} = \text{HIS published regional estimate;} \]

\[ X_{i} = \sum_{j=1}^{G} \frac{Z_{ij}}{Y_{ij}} Y_{i} \quad X_{i} = \text{subarea estimate before ratio adjustment;} \]

\[ P_{i} = \text{proportion of region } R^{'}s \text{ current estimated population belonging to the } i^{th} \text{ subarea.} \]

1.3.2. **Other Methodologies**

Before deciding to accept the estimator of (1.3.2), NCHS considered three other potential methods. Method I is a nearly unbiased procedure which allocates the HIS stratum estimates in the following manner:

\[ x_{i}^{(I)} = \sum_{h=1}^{357} b_{ih} x_{ih} \quad (1.3.3) \]

where \( b_{ih} \) is the proportion of the \( h^{th} \) stratum population coming from the \( i^{th} \) state and \( x_{ih} \) is the HIS final estimate for the \( h^{th} \) stratum.

Method III is an adaptation of Woodruff's procedure (1966) whereby estimates are made for every PSU in the sample or not. The model is presented as:
\[ X_{i(III)}^{(III)} = \sum_{h=1}^{357} \sum_{k} \lambda_{ikh} \hat{X}_{kh} \]

where \( X_{kh} \) exists for every primary sampling unit (PSU) in the universe with \( k \) representing the PSU,

\[ \hat{X}_{kh} = \alpha_{kh} X_{kh}^{*} + \frac{X_{kh}^{*}}{Y_{kh}^{*}} (Y_{kh} - Y_{.h}) \]

and \( \lambda_{ikh} \)

= 1 if any part of the \( kh \) is in \( i \), except that non-
self-representing PSU's crossing state lines are
allocated to one state;

= 0 otherwise.

\( \alpha_{kh} \)

= \( b_{ih} \) as in (1.3.3) if \( k \) is a self-representing PSU;

= 0 otherwise.

\( X_{kh}^{*} \)

= \( X_{.h} \) as in (1.3.3) if \( k^{th} \) PSU is in sample;

= 0 otherwise.

\( X^{*} \)

= regular HIS estimate for region.

\( Y_{kh} \)

= most recent Census population in the \( k^{th} \) PSU in the
\( h^{th} \) stratum.

\( Y_{.h}^{*} \)

= \( \frac{P_{.h}}{P_{kh}} \) \( Y_{kh}^{*} \) if \( k^{th} \) PSU in \( h^{th} \) stratum is in sample;

= 0 otherwise.

\( \frac{P_{.h}}{P_{kh}} \)

= reciprocal of first stage probability of selection.

And, \( Y^{*} \)

= \( \sum_{h=1}^{R} Y_{.h}^{*} \) are \( R \) states in \( R^{th} \) region.

Finally, Method IV attempts to combine the advantages of both
(1.3.2) and (1.3.4). It appears to be most suited for the largest sub-
areas (states) which contain the large number of PSU's. The equation
for the composite estimate is

\[ X_{i(IV)}^{(IV)} = w X_{i(III)}^{(III)} + (1-w) X_{Rd}^{**} \]
where \( w \) is an appropriate proportion which is high for estimates of populous subareas (states).

To determine the most promising technique, comparisons were made on the basis of available data on population, employment and disability. The following criteria were used for the purpose of evaluation: consistency both within and between methods, plausibility, and comparisons with data from an external source. Method II produced the most acceptable results.

1.3.3. Measures of Reliability

Due to the nature of their derivation, the synthetic estimates obtained from the NCHS model (1.3.1) are biased. Consequently, a popular measure used to assess their reliability is the mean square error. This can be expressed as the sum of the variance and square of the bias:

\[
\text{MSE}(x_{i}^{**}) = \sum_{j=1}^{G} p_{ij}^{2} \sigma_{x_{i}^{*}}^2 + (x_{i}^{**} - x_{i})^2
\]  

(1.3.6)

where \( \sigma_{x_{i}^{*}}^2 \) is the sampling variance of estimate \( x_{i}^{*} \),

\( x_{i} \) is the "true value" of the statistic for subarea \( i \), and

\( x_{i}^{**} \) is the expected value of the synthetic estimate for subarea \( i \).

The estimate given in (1.3.6) assumes that

the \( p_{ij} \)'s are fixed and measured without error, and

the \( \text{cov}(x_{i}^{*}, x_{j}^{*}) = 0 \), for \( j \neq i \).

Since the values of \( x_{i} \) are rarely known, one is often hard-pressed to obtain the MSE of \( x_{i}^{**} \) from (1.3.6). As an alternative, it is possible to produce an estimate of the average MSE of \( x_{i}^{**} \) from the sample. Consider
\[
E \left[ \frac{1}{M} \sum_{i=1}^{M} (X_{1i}^{**} - X_i)^2 \right] = \alpha, \quad (1.3.7)
\]

with the average calculated over all M subareas defined by the survey population. Gonzalez and Waksberg (1973) have derived the following approximation for \(\alpha\):

\[
\hat{\alpha} = \frac{1}{M} \sum_{i=1}^{M} (X_{1i}^{**} - \frac{G}{\sum_{j=1}^{G} p_{ij} X_{1j}})^2
- \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{G} p_{ij}^2 (1 - 2f_{ij}) \sigma_{X_{1ij}}^2 \quad (1.3.8)
\]

where \(X_{1j}\) is the estimated statistic from the sample for domain \(j\) and subarea \(i\), \(f_{ij}\) is the sample estimate of the proportion of the total for the \(j\)th domain that is in the \(i\)-th subarea, that is,

\[
f_{ij} = \frac{n_{ij}}{\sum_{i=1}^{M} n_{ij}}.
\]

and \(\sigma_{X_{1ij}}^2\) is the sampling variance of estimate \(X_{1j}\).

Similarily, the average MSE of the usual survey estimates can be expressed as:

\[
E \left[ \frac{1}{M} \sum_{i=1}^{M} (X_{1i} - X_i)^2 \right] = \beta \quad (1.3.9)
\]

and approximated as:

\[
\hat{\beta} = \frac{1}{M} \sum_{i=1}^{M} \sigma_{X_{1i}}^2 \quad (1.3.10)
\]

where \(\sigma_{X_{1i}}^2\) is the sampling variance of estimate \(X_{1i}\).

This provides a useful mechanism for comparing the accuracy of synthetic estimates with the usual survey estimate. By examining the results obtained in (1.3.7) and (1.3.9), one determines when the NCHS model is most preferable. Naturally, sample estimates of \(\sigma_{X_{1i}}^2\) and \(\sigma_{X_{1ij}}^2\)
are required.

Confidence bounds can be derived from the square root of the average MSE if the deviations of the estimates from the true values are approximately normally distributed. Unlike the sampling errors of unbiased estimates which satisfy this condition, no parallel theory exists for synthetic estimates. Their distributions are subject only to empirical examination.

Using 1960 data from the Census of Housing (units dilapidated with all plumbing facilities), Gonzalez (1973) compared the empirical distribution of the biases of state synthetic estimates with the normal distribution. No major discrepancies were evident, though more than the expected number of outliers were observed.

A number of other procedures have been used to determine the reliability of the synthetic estimates. In the study just cited, Gonzalez examined the distribution of relative biases for state synthetic estimates of dilapidated housing units with all plumbing facilities. It was observed that the proportion of estimates with large relative biases diminished as the size of the synthetic estimate increased. The highly variable nature of these estimates was partially confirmed by the fact that the average root mean square divided by the mean was about 20 percent. In an earlier study, Gonzalez and Waksberg (1973) compared the reliability of synthetic and unbiased estimates of unemployment for SMSA's. Here, monthly, quarterly, and annual estimates were obtained from Current Population Survey data. Using relative errors and coefficients of variation when applicable to measure reliability, the findings were mixed. The stable synthetic estimates were preferable for monthly data, though unbiased estimates proved superior for annual averages. For quarterly data, no discernable difference in
was evident. Additional information regarding the biases of synthetic estimates of unemployment was forthcoming from Gonzalez and Hoza (1975). Due to the nature of their derivation, the estimates were observed to cluster near the mean for a specific geographical region rather than the true expected values for the county. Therefore, the biases increased for those counties whose unemployment experience diverged most notably from the norm for the region.

The availability of mortality data from the United States Vital Statistics Annual Volumes by cause of death for all states enabled Levy (1971) to determine the accuracy of related synthetic estimates at the state level. Agreement between these estimates and the true number of deaths was measured by the percentage absolute difference (PAD) defined as:

\[
\left| \frac{X_i - X_{ri}^{**}}{X_{ri}^{**}} \right| \times 100
\]

(1.3.11)

where \(X_i\) is the actual number of deaths (by specific cause) for the \(i^{th}\) subarea (i.e., state), and \(X_{ri}^{**}\) is a regionally adjusted synthetic estimate of deaths. The accuracy of the synthetic estimates varied considerably for the four causes of death examined. While agreement was good for deaths due to major cardiovascular diseases and suicide, as measured by respective median PAD's of 6.4% and 10.2%, it was generally poor for motor vehicle accidents (14.6%) and tuberculosis (32.0%). When comparing the synthetic estimator's effectiveness to a crude regionally adjusted estimator \(X_{ri}^{-}\), obtained for each state by multiplying the state's population by the crude death rate for the respective geographical region, no striking differences in performance were evident.

A similar investigation involving state estimates of work disability
was undertaken in a later study by Namekata, Levy and O'Rourke (1975).

1.3.4. Present Limitations

Although the NCHS model is a most welcome alternative to the one developed by Weber et al., it is assumed information pertaining to the demographic composition of the respective local areas is readily available (The $P_{ij}$'s are fixed and measured without error.) When the variable cross-classifications depend on a limited number of predispositional variables (i.e., age, sex race), one is more confident that the condition is satisfied. Contrarily, when as many as seven defining variables are considered, the results are questionable. At NCHS, the estimates were derived from available data only after the G-domains were collapsed. Schneider (1973) assumed that each of the selected indicators were independent of the others. In this manner the proportion of the population belonging to a particular socio-economic type could reasonably be estimated "as the product of the proportion in each of the indicators which make up the S.E.T." Neither treatment is completely satisfactory to the issue at hand.

Predispositional variables which are absent from the model are treated as having insignificant effects on the characteristic to be estimated. In addition, the $X_i^{**}$'s for two distinct subareas differ only with respect to differences in demographic composition - the $P_{ij}$'s. The impact of such internal forces as politics, policy, social structure and environment are not directly considered at the local level. Consequently, dramatic differences between parameter estimates for particular subareas will occasionally be masked. To be more explicit, consider the forces operating in a political campaign. If a candidate spent
more time in a number of select communities than an opponent with similar views, only the regional component of the model would reflect this. Otherwise, the time factor is virtually ignored when deriving estimates of opinion for the respective subareas.

1.3.5. Potential Improvements

Levy (1971) proposed a method of using local variables related to the parameter being estimated in conjunction with the NCHS estimator to yield an improved estimator. The following model was considered:

\[ Y_i = a + BZ_i + \epsilon_i \]  \hspace{1cm} (1.3.12)

where \( Z_i \) is the value of an ancillary variable \( Z \) for the \( i^{th} \) subarea;

\[ Y_i = \frac{X_1 - X_{Ri}^{**}}{X_{Ri}^{**}} \times 100 \]

\( X_{Ri}^{**} \) = the synthetic estimate of the \( X \) characteristic for the \( i^{th} \) subarea;

\( X_i \) = the true value of the \( X \) characteristic for the \( i^{th} \) subarea;

\( \epsilon_i \) = a term representing random error;

and \( a \) and \( B \) are regression parameters to be estimated.

In (1.3.10) the percentage difference between the synthetic estimator and the true value is treated as a linear function of some related variable, \( Z_i \). Were the estimates \( \hat{a} \) and \( \hat{B} \) available and \( \epsilon_i \) omitted, an estimator \( \hat{X}_i \) of \( X_i \) could be derived from (1.3.10), taking the form:

\[ \hat{X}_i = X_{Ri}^{**} \left( \frac{\hat{a} + \hat{B}Z_i}{100} + 1 \right) \]  \hspace{1cm} (1.3.13)
It is assumed that \( Z_i \) is available for every subarea, but since \( Y_i \) depends on the true value \( \chi_i \), which is unknown, the coefficients cannot be estimated. Nevertheless, one can obtain unbiased estimates \( X_c^* \) of \( \chi_c \) where \( \chi_c \) is the value of \( X \) for the \( c \)-th stratum combination, where strata are generally counties or groups of counties. As before, the ancillary variable \( Z_i \) and the synthetic estimate \( X_{Ri}^{**} \) can easily be determined for each strata combination. By dividing the population into \( C \) strata combinations, \( a \) and \( B \) can be estimated by least squares from the data pairs \((Z_c^*, Y_c^*)\) where

\[
Y_c^* = \frac{X_c^* - X_{c}^{**}}{X_{c}^{**}} \times 100
\]  

(1.3.14)

c = 1, ..., C

and substituted back into (1.3.11) to yield estimates \( \hat{\chi}_i \) for each subarea. The method can be extended to consider \( Z_i \) as a vector of ancillary data, whereby \( \hat{\chi}_i \) is treated as a multiple regression estimator.

1.4. Contingency Table Synthetic Estimation

Contingency table synthetic estimation, also known as "raking," is another potentially useful procedure, frequently used by the Bureau of the Census. One extension of the technique, suggested by Koch and Freeman (1975), allows for the adjustment of the observed table corresponding to a sample from a specific population to produce estimators for other target populations. These include various local (county or state) sub-divisions of a nationally sampled population in addition to other local, national or international populations which occa-
sionally overlap a sampled population. To insure the methodology's validity, the following assumptions are required:

1. Certain marginal distribution referred to as the "allocation structure" is known on the basis of census or other sample survey data for the target population.

2. The higher order interactions across the subsets in (1), referred to as the "association structure," are the same for both sampled and target populations.

Specifically, consider a set of \( d \) attributes where \( j \in \{1, 2, \ldots, L\} \) indexes the response categories for the \( g \)-th attribute where \( g = 1, 2, \ldots, d \), and \( j = (j_1 \ldots j_d) \) denotes the vector response profile. Let

\[
\pi = \begin{bmatrix}
\pi_{11} & \ldots & 1 \\
. & & . \\
. & & . \\
\pi_{L_1} & \ldots & L_d
\end{bmatrix} \tag{1.4.1}
\]

represent the parameter vector characterizing the response profile distribution for the sample population, \( \hat{\pi} \) the corresponding estimator, \( \pi_T \) the parameter vector for the target population, \( A_T \) a matrix of coefficients whose columns characterize the "allocation structure," and \( \xi_T \) their known values. Consequently, assumption (1) requires \( \pi_T \) to satisfy

\[
A_T^\top \pi_T = \xi_T \tag{1.4.2}
\]

where \( A_T^\top \) is of full rank and

\[
1^\top \pi_T = 1 \tag{1.4.3}
\]

Similarly, assumption (2) requires \( \pi_T \) to satisfy
\[ \kappa^\top \log_e(\pi_T) = \kappa^\top \log_e(\pi) \] (1.4.4)

where \( \kappa \) is an ortho-complement matrix to \( \pi_T \) and the "association structure" is represented in terms of log-linear contrast functions.

These assumptions imply the marginal adjustment (raking) estimator \( \pi_T \) is characterized by the equations:

\[ \pi_T \hat{\pi}_T = \xi_T \]
\[ \kappa^\top \log_e(\hat{\pi}_T) = \kappa^\top \log_e(p) \] (1.4.5)

Here, the estimator \( \hat{\pi}_T \) and hence the estimated table for the target population can be determined by applying the Deming–Stephan Iterative Proportional Fitting (IPF) algorithm which preserves the "association structure" while adjusting the initial estimator to the respective marginal configurations represented by the "allocation structure."

To obtain the asymptotic covariance matrix, one considers an extension of the method based on the first order Taylor series (\( \delta \)-method). Consequently, the asymptotic covariance matrix is represented by:

\[ \nabla_\pi \hat{\pi}(\pi) = \kappa^\top \left[ D_{\pi} \kappa^\top D_{\pi}^{-1} \right] \kappa^\top D_{\pi}^{-1}[\nabla(\pi)] D_{\pi}^{-1} \kappa^\top \left[ D_{\pi} \kappa^\top D_{\pi}^{-1} \right] \kappa^\top \] (1.4.6)

where \( \nabla(\pi) \) denotes the covariance matrix of the estimator \( \pi \) and \( D_{\pi}^{-1} \) denotes the inverse of a diagonal matrix with the elements of \( \pi \) on the diagonal.

Reasonable estimators of \( \nabla(\pi) \) may be determined by substituting a consistent estimator \( \nabla(p) \) for \( \nabla(\pi) \) and replacing \( \pi \) and \( \pi_T \) by \( p \) and \( \pi_T \), respectively. Additional information regarding the statistical proper-
ties of the technique is noted in Causey (1972), Bishop, Feinberg and Holland (1975), Scheuren (1975), and Koch and Freeman (1976).

1.5. Methods Combining Sample Survey Data and Symptomatic Indicators

1.5.1. A Regression Estimate Approach

The demand for producing small area estimates of a criterion variable \( Y \), expressible in terms of a population total, often surfaces at the subnational level. A relatively uncomplicated method used to derive these small area estimates has been suggested by Woodruff (1966). It is essentially an extension of the regression estimate approach considered by Hansen et al. (1953) and later, by Cochran (1963).

Initially, one of the four censal regions is selected and a simple random sample of \( m_{(R)} \) counties is drawn without replacement from the \( M_{(R)} \) counties in the \( R \)th region (\( R = 1, ..., 4 \)). A within-county sample is then drawn to produce an unbiased estimate of the criterion variable for each sample county - \( Y_{(R)i}^- \). The final model contains an estimate of \( Y \) in addition to a correlated auxiliary variate \( X \), whose value for each county, \( X_{(R)i} \), must be accessible. Here, the best (lowest variance) linear unbiased estimate \( Y_{(R)}^- \) for region \( R \) is:

\[
Y_{(R)}^- = Y_{(R)}^- + B(X_{(R)}^- - X_{(R)})
\]

where

\[
Y_{(R)}^- = \frac{m_{(R)}}{m_{(R)}} \sum_{i=1}^{m_{(R)}} Y_{(R)i}^-
\]
\[ X^*(R) = \frac{m(R)}{M} \sum_{i=1}^{m(R)} X(R)_i \]
\[ X(R) = \sum_{i=1}^{M} X(R)_i \]
\[ B = \frac{\sigma^2_{X(R)} Y(R)}{\sigma^2_{X(R)}} \], which is computed from full-county totals and assumed to be known at this point.

Equation (1.4.1) may be restated as:

\[ Y^*(R) = \sum_{i=1}^{M} Y(R)_i + B(X(R)_i - X^*(R)_i) \quad (1.5.2) \]

where \( Y(R)_i \):
- if county \( i \) in region \( R \) falls in sample,
- \( = 0 \) otherwise;

\( X(R)_i \) = census total for county \( i \) in region \( R \),

\( X^*(R)_i \):
- if county \( i \) in region \( R \) falls in sample,
- \( = 0 \) otherwise.

The variance of \( Y^*(R) \) takes the form:

\[ \sigma^2_{Y^*(R)} = M(R)^2 \left( \frac{(M(R) - m(R))}{(M(R) - 1)m(R)} \right) \sigma^2_{Y(R)} \left( 1 - \rho_{X(R)Y(R)}^2 \right) + \frac{M(R)}{m(R)} \sum_{i=1}^{M} \omega^2_{(R)_i} \]

\( (1.5.3) \)

where \( \sigma^2_{Y(R)} \) is the between-county component of variance for the full county totals;

\( \rho_{X(R)Y(R)} \) is the correlation between \( X(R)_i \) and \( Y(R)_i \).
and \( W_{(R)i}^2 \) is the within county component of variance for the \( i^{th} \) county in variate \( Y\)\(_{(R)i}\).

The principle of making estimates of variance of complex estimates with a single set of squared deviations is discussed by Keyfitz (1975). Consequently, one may estimate \( \sigma^2 \)\(_{(Y)}\) from the sample as:

\[
\hat{\sigma}^2 = \frac{m_{(R)}}{m_{(R)} - 1} \sum_{i=1}^{m_{(R)}} (Z_i - \bar{Z})^2
\]  
(1.5.4)

where

\[
Z_i = \frac{M_{(R)}}{m_{(R)}} Y\)\(_{(R)i}\) - B \frac{M_{(R)}}{m_{(R)}} X\)\(_{(R)i}\)
\]

and

\[
\bar{Z} = \frac{\sum_{i=1}^{m_{(R)}} Z_i}{m_{(R)}}
\]

For those counties in the sample, the value of expression (1.5.2), which serves as a county total estimate, is

\[
\frac{M_{(R)}}{m_{(R)}} Y\)\(_{(R)i}\) + B(1 - \frac{M_{(R)}}{m_{(R)}}) X\)\(_{(R)i}\)
\]  
(1.5.5)

Similarly, (1.5.2) reduces to \( BX\)\(_{(R)i}\) for those counties not represented. One must only sum the relevant estimated county totals in (1.5.5) or \( BX\)\(_{(R)i}\) to generate an estimate for any geographic area within the region. These estimates will be consistent though biased when \( B \) is estimated from the available data. In practice, Woodruff advocates using the ratio of \( Y \) to \( X \) as calculated from the sample at the regional level to estimate \( B \). The variance of these area estimates may be determined from (1.5.4) by setting \( Z_i \) to zero for each of the counties not included.
When only a small number of counties are considered, the estimated variance will have a tendency to be high.

1.5.2. A More Complicated Design

While the original model assumes no stratification within region, one is never so limited in practice. Another often considered design used by the Bureau of the Census divides all the counties in the U.S. into L primary sampling units (PSU's). Counties within a PSU must be contiguous, but can be found in more than one state. Homogeneous PSU's within a region are then combined into S strata and a single PSU is drawn from each with probability proportionate to size. In this situation it is more difficult to use available county data and estimated stratum totals in determining a criterion variable's estimated total for a specific geographic area. Although the method previously discussed may be used, $M(R)/m(R)$ must be replaced by $1/P_i$, where $P_i$ is the probability of drawing the PSU. Since the PSU is usually composed of a combination of counties, the largest must be selected as the "principal" county. For non-principal counties, $Y_{(R)i}^{(R)}$ and $X_{(R)i}^{(R)}$ in (1.5.2) are always set to zero. The same is done for principal counties within PSU's not represented in the sample. Otherwise, $Y_{(R)i}^{(R)}$ and $X_{(R)i}^{(R)}$ in (1.5.2) are determined by the product of the complete PSU totals and $1/P_i$ when the PSU is in the sample. If the PSU's are not drawn with certainty, variance estimates for geographical areas should be determined by grouping strata.

1.5.3. A Multiple Regression Framework

Ericksen (1974) developed a new technique for computing local
estimates which combined symptomatic information and sample data into a multiple regression format. Referred to as the regression-sample data method of local estimation, the three basic steps of the procedure are:

1. Compute sample estimates of the criterion variable for the respective primary sampling units in the sample. Although data is usually collected by a two-stage sampling frame, probability sampling is sufficient.

2. Collect symptomatic information for both sample and non-sample PSU's. Typical predictor variables are the number of births, deaths, and school enrollment.

3. Compute the linear least squares regression estimate using data for the sample PSU's only. Estimates for all subareas are then determined by substituting values of the symptomatic indicators, whether included in the respective sample or not.

Although the method is applicable for estimating any parameter for which the sample and symptomatic data is available, attention has been directed to post-censal estimates of population growth. To reduce the variability and skewness of the distribution, it is suggested that variables be written in ratio form. The procedure resembles the ratio-correlation technique first introduced by Snow (1911) and developed by Schmitt and Crosetti (1954), which estimates the multivariate relationship among population growth and predictor variables. Post-censal estimates derived using the ratio-correlation method requires the fitting of a linear model to selected variables represented in terms of a ratio of measurements taken at the endpoints of the immediately preceding intercensal period. The availability and inclusion of information pertaining to each subarea of the total population is essential. In addition, satisfactory results can only be expected
when the functional form of the actual and predicted models vary only slightly. Assuming the stability of relationships between the inter-
censal and post-censal periods, desired small area estimates are ob-
tained by entering the respective post-censal changes in the values of
the symptomatic variables into the resulting equation. Ericksen's
procedure uses data which is exclusively post-censal and obtained from
sample surveys. Consequently, fewer restrictions are specified for
the method to yield reliable results.

The model assumes the availability of criterion variable esti-
mates for each of n sample PSU's and the values of p symptomatic indi-
cators for the universe of N local areas. It takes the matrix repre-
sentation:

\[ \mathbf{Y} = \mathbf{XB} + \mathbf{u} \]  \hspace{1cm} (1.5.6)

where \( \mathbf{Y} \), an \( n \times 1 \) vector, is the criterion variable consisting
of a set of actual unobserved values \( \mathbf{X} \), an \( n \times (p+1) \)
matrix denoting the set of predictor variables;

\( \mathbf{B} \), the \( (p+1) \times 1 \) vector of regression coefficients;

and \( \mathbf{u} \), an \( n \times 1 \) vector, a stochastic error term.

Under the assumption of linearity, \( \mathbf{B} \) could be estimated by
ordinary least squares regression were the \( \mathbf{Y} \) values observed. Because
the individual observations of \( \mathbf{Y} \) are affected by sampling variability,
the model may be revised to explain the within-PSU sampling error in
the following manner:

\[ \mathbf{\tilde{Y}}_0 = \mathbf{XB} + \mathbf{\tilde{u}} + \mathbf{\tilde{y}} \]  \hspace{1cm} (1.5.7)

where \( \mathbf{\tilde{y}} \) is an \( n \times 1 \) vector of sampling error deviations,
and \( \mathbf{\tilde{Y}}_0 \), the observed values.
The regression equation is then computed, substituting the observed values of $Y_0$ for $Y$. Hence, the regression coefficients are unbiased in the absence of correlations between $y$ and $\hat{y}$. The mean square error of the regression estimates is expressed as:

$$E(\hat{Y} - \bar{Y})(\hat{Y} - \hat{Y})/n = [(n - p - 1) \sigma_u^2/n] + [(p + 1) \sigma_v^2/n]$$

(1.5.8)

where $\sigma_u^2$ is the between-PSU variance unexplained by the predictor variables,

and $\sigma_v^2$ is the within PSU variance.

When $n/p$ is large, the within PSU component of error is reduced for increases in $n$. Increases in $p$ and large reductions in $n$ have the opposite effect. The mean square error could be reduced, were the reductions of the first term (obtained by adding symptomatic indicators or stratifying the sample into smaller, more homogeneous groups) great enough to counterpoise the resulting increases in the second term.

Because the true values of the criterion variable are unknown, the M.S.E. cannot be estimated directly. However, by considering the mean square difference between regression estimates,

$$E(\hat{Y}_0 - \hat{Y})(\hat{Y}_0 - \hat{Y})/n = [(n - p - 1)/n] (\sigma_u^2 + \sigma_v^2)$$

(1.5.9)

the mean square error (1.5.8) can be obtained by subtracting an allowance for the within-PSU component of error $[(n-2p-2)/n] \sigma_v^2$. Though it primarily serves as a measure of the reliability for sample PSU's, it provides an indication of the errors for counties. To determine the best model, one would compare the estimated MSE for all possible regression equations defined by the potential symptomatic indicators.
Inclusion of an alternate estimator (i.e., ratio correlation) is advised.

1.5.4. **An Alternative Strategy**

The method advanced by Eriksen is most feasible when the linearity assumption is satisfied and the observed multiple correlation is high. But what decision is reached when the multiple correlation level is moderate (.5-.8) and a non-linear model is more suitable? The inclusion of all possible symptomatic variables into the regression would increase the $R^2$ but most probably at the expense of an "over-fit" model which increases the mean square error of the final estimate. More generally, in those situations where assumptions are too strict or unrealistic, the need for a more flexible approach is most obvious. Kalsbeek (1973) has developed one such procedure in which the most limiting assumption is the availability of good symptomatic information.

It has usually been common practice to treat the local area units as the smallest level for which the estimates are made. Contrarily, Kalsbeek suggests breaking up the local unit into constituent geographical sectors called "base units," such as townships, enumeration districts, or other geographical subunits of a county. The local area for which a variable of interest is to be estimated is referred to as the "target area" and further subdivided into "target area base units." Unlike other methods which use symptomatic information directly for the purposes of estimation, Kalsbeek's procedure uses the information to group base units (sample base units) from the total population. The symptomatic information is also used to classify "target area base units" into the appropriate group.
Initially, a random sample of \( n \) base units is selected from the total population of \( N \) base units. The sample base units (possibly including some "target area base units") are required to possess both symptomatic and criterion information. These units are divided into \( K \) groups (strata) using either or both types of the information available. The object is to form groups which are most homogeneous within while dissimilar between themselves. It is noteworthy that the respective groups may be defined by either rectilinear or non-rectilinear boundaries. To be more explicit, consider the \((n \times p)\) matrix \( \bar{X} \) of symptomatic information where

\[
\bar{X} = \begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_n
\end{bmatrix}_{(n \times p)}
\]

and \( \bar{X}_i \) is a \((1 \times p)\) row vector of available symptomatic information for the \( i \)th sample base unit. We want to partition the \( \bar{X}_i \) into \( K \) groups, where \( K \) is specified, which are non-empty, non-overlapping and exhaustive:

\[
C_1, C_2, \ldots, C_K \quad \text{where} \quad C_g \neq \emptyset,
\]

\[
C_g \cap C_{g'} = \emptyset
\]

\[
\bigcup_{g=1}^{K} C_g = \{ \bar{X} \} \quad (p \times n)
\]

If \( n_g \) is the number of sample base units in the \( g \)th group, \( \bar{X}_g \) is defined as the sample mean of the \( g \)th group so that

\[
\bar{X}_g = \frac{1}{n_g} \sum_{\bar{X}_i \in C_g} \bar{X}_i
\]

\((p \times 1)\)
and $\overline{\tilde{x}^*}$ is defined as the overall sample mean, where

$\overline{\tilde{x}^*} = \frac{1}{K} \sum_{g=1}^{K} n_g \overline{\tilde{x}^*_g} / n$

(p×1)

Here, the total sum of squares is expressed as

$$(SS)_T = \sum_{i=1}^{n} (x_i - \overline{x}) (x_i - \overline{x})$$

Similarly, the sum of squares for the $g^{th}$ group is

$$S_g = \sum_{x_i \in C_g} (x_i - \overline{\tilde{x}_g}) (x_i - \overline{\tilde{x}_g})$$

so that the pooled sum of squares within groups takes the form

$$(SS)_W = \sum_{g=1}^{K} S_g = \sum_{g=1}^{K} \sum_{x_i \in C_g} (x_i - \overline{\tilde{x}_g}) (x_i - \overline{\tilde{x}_g})$$

The between sum of squares for groups takes the form

$$(SS)_B = \sum_{g=1}^{K} n_g (\overline{\tilde{x}_g} - \overline{x}) (\overline{\tilde{x}_g} - \overline{x})$$

and

$$(SS)_T = (SS)_W + (SS)_B$$

The groups that are formed take the shape of clouds in $p$ dimensional space. To form groups which are optimally homogeneous within while heterogeneous between, one must decide upon criteria which minimize $(SS)_W$ and maximize $(SS)_B$. Two useful criteria are:

1) the trace of $(SS)_W^{-1} (SS)_B$

2) the trace of $(SS)_T^{-1} (SS)_B$

where (1) is to be maximized and (2) is to be minimized. It is evi-
dent that the evaluation of all possible partitions will be quite expensive when \( n \) is large. Consequently, a number of useful iterative procedures in cluster analysis (i.e., Automatic Interaction Detection (A.I.D.), Multivariate Iterative K-Means Cluster Analysis (MIKCA)) serve as reasonable alternatives which consider the same criteria.

All "target area base units" belonging to the local area in question are then assigned (classified) to one of the \( K \) groups with respect to symptomatic information. Consequently, each "target area base unit" is associated with a group of base units both similar to itself and internally homogeneous. An estimate for each of the "target area base units" with respect to the criterion variable is obtained from the sample base units in the group to which it has been assigned. These estimates are then pooled to arrive at a final estimate for the respective target area.

Familiarity with the following notation is essential to a more explicit presentation of the estimation procedure:

- \( Y \) - the criterion variable to be estimated.
- \( \bar{Y} \) - an estimate of the true mean of \( Y \).
- \( n \) - number of units.
- \( w \) - weighting factor assigned to unit.
- \( P \) - probability that the unit is in the sample survey.
- \( M \) - measure of size.
- \( * \) - association with the target area data, whereas its absence indicates association with the sample data.
- \( s \) - number of symptomatic variables.
- \( k \) - number of groups formed.
- \( g \) - the \( g \)th group.
- \( gh \) - the \( h \)th base unit in the \( g \)th group.
- \( ghc \) - the \( c \)th item in the \( gh \)th base unit.
The sampling framework considered strongly resembles that of the multi-stage variety and is treated accordingly. Hence the estimates for each base unit are:

$$\bar{Y}_{gh} = \frac{\sum_{c=1}^{n} w_{ghc} Y_{ghc}}{L}$$  \hspace{1cm} (1.5.10)

where

$$w_{ghc} = \left(P_{ghc} \sum_{c=1}^{n} 1/P_{ghc}\right)^{-1}$$  \hspace{1cm} (1.5.11)

Were the $P_{ghc}$ in each group equal, then

$$w_{ghc} = \frac{1}{n_{gh}}$$  \hspace{1cm} (1.5.12)

The estimate for the $g^{th}$ group takes the form

$$\bar{Y}_g = \frac{\sum_{h=1}^{n} w_{gh} \bar{Y}_{gh}}{L}$$  \hspace{1cm} (1.5.13)

where, as one possibility,

$$w_{gh} = \left(P_{gh} \sum_{h=1}^{n} 1/P_{gh}\right)^{-1}$$  \hspace{1cm} (1.5.14)

The final estimate is

$$\bar{Y}^* = \frac{k}{\sum_{g=1}^{k} w^* Y_g}$$  \hspace{1cm} (1.5.15)

where, as one possibility,

$$w^*_g = \frac{M^*_g}{\sum_{g=1}^{k} M^*_g}$$  \hspace{1cm} (1.5.16)

The availability of Census data on population and per capita income for 1970 allowed for an examination of the method's accuracy. State estimates of population growth (from 1960 to 1970) and per capita income were generated by Kalsbeek, using the Current Population Survey
(a national multi-stage probability sample of the U.S. conducted monthly) as the source of sample information. Here, the sample base units corresponded to the first stage sampling units (PSU's) in the C.P.S., which are counties or groups of contiguous counties. The symptomatic variables considered when estimating population growth included total school enrollment, live births, and deaths, all expressed in ratio form (1970 total/1960 total). Those considered in the per capita income example included the percent natural increase in population between 1960 and 1965, the 1960 per capita aggregate income, and the 1964 percent of the population on public assistance (all obtained from the 1967 County-City Data Book).

The grouping of the sample base units (PSU's) was done using the Automatic Interaction Detector, version II, which is essentially a clustering algorithm that uses both symptomatic and criterion value information. Since the respective groups (strata) formed have rectilinear boundaries, the "target base units" (here counties) are assigned to the group whose boundaries include the observation's symptomatic values. Hence, each target base unit takes on the group estimate of the criterion variable to which it is assigned. For the \( i \)th state, one considers the respective target base unit alignment, and weights the group (strata) estimators by the proportion of the state's population in the target base units so classified. Here, the 1960 county populations were used.

The method was compared with Ericksen's procedure since both are applicable under essentially the same circumstances. The criterion for measuring the accuracy of the estimates was the relative absolute
deviation from the true value: $\frac{\text{Estimated} - \text{True}}{\text{True}}$. Ericksen's procedure would be expected to give better results for the population growth example due to the inclusion of three symptomatic variables with a high level of multiple correlation and an underlying linear relationship. Still, the method of Kalsbeek yielded more accurate estimates for more than 25 percent of the states considered (11 out of 42). It was observed that the results tended to improve with increases in population size for both methods. Kalsbeek's method did much better in generating state estimates of per capita income, yielding more accurate results in 29 of the 47 states considered.
CHAPTER II
ESTIMATION USING THE KALSBEEK MODEL

2.1. Introduction

Several of the methods proposed for estimating the parameters of local areas have been observed to be rather limited in scope. This is primarily due to their restrictive underlying assumptions. The need for a less restrictive though equally precise approach has already been raised. The method considered at NCHS improves upon the Simulmatics Project technique, though not without its share of constraints. By assuming that small areas share the same characteristics as a standard national distribution, they can only be distinguished by their respective demographic configurations. Consequently, the method is not particularly sensitive to many of the internal forces operating at the local level. It has been suggested that the problem might be resolved by adjusting the NCHS estimator with available data on local area auxiliary (symptomatic) variables, though only at the risk of confounding the final estimates. The method advanced by Ericksen makes direct use of symptomatic information and is most reasonable when the assumption of linearity holds. Often, however, the relationship between criterion and symptomatic variables is more appropriately characterized by a non-linear model. The method introduced by Kalsbeek is particularly attractive in that no functional model between criterion and symptomatic variables must be specified. Here, the most limiting
assumption is the availability of good symptomatic information. In
text, it most closely resembles the NCHS model, though like Erickson's
approach, direct use is made of symptomatic information. Estimates
for the base units of the "target area" (respective local area) are
available as a by-product of the technique. Finally, the method per-
forms reasonably well even for linear setting, though here it would
be better to choose Erickson's approach.

Both the need for good small area estimates and the merits of
Kalsbeek's approach not only justify but motivate further research in
this area. In this chapter, a more formal representation of the model
is developed for a single-stage cluster sample design and a two-stage
sampling design with equal probabilities of selection. In addition,
a general analytic expression for the mean square error of the resul-
tant estimators is derived.

2.2. Notation

Consider a population consisting of \( L \) local areas, indexed by
\( \ell = 1, 2, \ldots, L \), which have further been subdivided into constituent
geographical sectors called "base units." There are \( N_{\ell} \) base units in
the \( \ell \)th local area, and

\[
\sum_{\ell=1}^{L} N_{\ell} = N
\]  

(2.2.1)
in the population, individually indexed by \( i = 1, 2, \ldots, N_{\ell} \), to
denote the \( i \)th base unit from the \( \ell \)th local area. When the local area
reference is dropped, each base unit is indexed by \( i = 1, 2, \ldots, N \).
Furthermore, each base unit \( i \) consists of a cluster of \( M_{i} \) smaller units
referred to as elements. Hence, there are $M_L = \sum_{i=1}^{N_L} M_i$ elements in the $l^{th}$ local area and $M_i = \sum_{l=1}^{L} M_L = \sum_{i=1}^{N} M_i$ elements in the population. Let $y_{ij}$ represent the observed value of the criterion variable for the $j^{th}$ element within the $i^{th}$ base unit, where

$$y_i = \sum_{j=1}^{M_i} y_{ij} \quad (2.2.2)$$

is the $i^{th}$ unit total.

Here, we assume a two-stage sampling design, whereby a simple random sample of $n$ base units (first stage units) is initially drawn from the $N$ base units in the population. A subsample of $m$ out of the $M_i$ elements is then selected with equal probabilities of selection from each of the chosen sample base units. The subunits are chosen independently in different base units. The units are then divided into $K$ groups (strata), indexed by $g = 1, 2, \ldots, K$, by one of the aforementioned procedures (Section 1.4.5). Consequently, estimates of the group means are obtained by a method which closely resembles post-stratification. To determine the criterion variable estimator for the $l^{th}$ local area, each "target base unit" is assigned to the group most similar with respect to symptomatic information. Thus, we have a two-way classification of all base units in the population by respective strata and local areas. This can be observed in the following LxK figure.
### Classification of Base Units by Strata and Local Areas

<table>
<thead>
<tr>
<th>Strata (g)</th>
<th>Total number of base units in the ( k )th local area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( N_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( N_2 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( g )</td>
<td>( N_{kg} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( K )</td>
<td>( N_K )</td>
</tr>
</tbody>
</table>

Total number of base units assigned to the \( g \)th strata

where \( N_{kg} \) is the total number of base units in the \( g \)th strata from the \( k \)th local area.

### 2.3. Representation of the Model

The local area estimator of the criterion variable may be expressed in terms of an average, a proportion, or a total. Initially, we direct attention to the mean per element representation.

Assuming a two-stage sampling design with sub-units of unequal sizes, we define

\[
y_1 = \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}
\]  

(2.3.1)
as the sample mean per element in the \(i^{th}\) base unit and
\[
\bar{y}_i = \frac{\sum_{j=1}^{M_i} y_{ij}}{M_i}
\]  
(2.3.2)
as the overall mean per element in the \(i^{th}\) base unit. To obtain an estimate of the \(g^{th}\) stratum mean per element, we also define the indicator variables \(I_{gi}\) (once more dropping the local area reference), such that
\[
I_{gi} = 1 \text{ if the (first stage) base unit falls in the } g^{th} \text{ stratum;}
\]
\[
= 0 \text{ otherwise}
\]
for \(g = 1, 2, \ldots, K\) and \(i = 1, 2, \ldots, N_i\). Here, \(\sum_{i=1}^{n} I_{gi} = n_g\), the number of sample base units belonging to the \(g^{th}\) stratum, and
\[
\sum_{i=1}^{N} I_{gi} = N_g
\]
Consequently, let
\[
\hat{y}_g = \frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} I_{gi} M_i} = \frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}
\]  
(2.3.3)
(summed only over the \(n_g\) sample base units from the \(g^{th}\) stratum) be our (post-stratified) estimator of the \(g^{th}\) stratum mean per element.

Since \(\hat{y}_g\) is a ratio estimator of
\[
\hat{y}_g = \frac{\sum_{i=1}^{N} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{N} I_{gi} M_i} = \frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}
\]  
(2.3.4)
(where the sum is over the \( N_i \) base units assigned to the \( g^{th} \) stratum), it is biased to the order of \( 1/n \). Yet, when \( n \) is large (i.e., \( n > 100 \)), the bias is negligible and the expectation of \( \hat{y}_g \) is approximately equivalent to \( \bar{y}_g \),

\[
E(\hat{y}_g) = \bar{y}_g, \quad g = 1, 2, \ldots, K.
\]  

(2.3.5)

Returning to the \( l^{th} \) local area, we focus attention on the "target base unit" alignment in order to weight appropriately the stratum estimators \( \hat{y}_g \) by the proportion of elements in the base units so classified. Therefore, the estimator of the criterion variable for the \( l^{th} \) local area takes the following form:

\[
\hat{y}_l = \sum_{g=1}^{K} \frac{M_g}{M_l} \hat{y}_g
\]

(2.3.6)

such that

\[
E(\hat{y}_l) = \sum_{g=1}^{K} \frac{M_g}{M_l} E(\hat{y}_g) = \sum_{g=1}^{K} \frac{M_g}{M_l} \bar{y}_g
\]

(2.3.7)

when \( n \) is large. Often the sizes of \( M_{lg} \) and \( M_l \) are only known approximately. When this occurs, the respective estimators of the stratum means are weighted by the ratio of available estimates \( M_{lg} \) and \( M_l \), or by the cruder ratio \( N_{lg}/N_l \).

Due to the nature of its derivation, the local area estimator \( \hat{y}_l \) of \( \bar{y}_l \) is biased. The observed value of the criterion variable mean per element is

\[
\bar{y}_l = \frac{\sum_{i=1}^{N_l} \bar{y}_i}{\sum_{i=1}^{N_l} M_i} = \frac{\sum_{i=1}^{N_l} M_i \bar{y}_i}{\sum_{i=1}^{N_l} M_i}
\]

(2.3.8)
summed across only those base units in the $i^{th}$ local area. The bias,

$$B = [E(y_{i,x}) - \bar{y}_{i,x}]$$

(2.3.9)

can be approximated by

$$B' = \left[ \sum_{g=1}^{K} \frac{M_{i,x}}{M_{i'}} \bar{y}_{i,g} - \frac{\sum_{i=1}^{Ng} M_{i} \bar{y}_{i}}{M_{i'}} \right]$$

(2.3.10)

Similarly, to express the local area estimator in terms of a proportion, $y_{ij}$ is redefined, so that

$$y_{ij} = 1 \text{ when the } j^{th} \text{ element in the } i^{th} \text{ base unit}
\text{ has the characteristic of interest};
$$

$$= 0 \text{ otherwise},$$

so that

$$\frac{M_{i}}{\sum_{j=1}^{N_{i}} y_{ij}} = \bar{y}_{i}$$

(2.3.11)

is the total number of elements in the $i^{th}$ base unit with the characteristic of interest.

When a single-stage cluster sampling design is assumed, a simple random sample of $n$ base units (cluster units) is drawn from the $N$ base units in the population. To obtain an estimator of the $g^{th}$ stratum mean per element, let

$$\hat{y}_{g} = \frac{\sum_{i=1}^{n} I_{gi} M_{i} \bar{y}_{i}}{\sum_{i=1}^{n} I_{gi} M_{i}} = \frac{\sum_{i=1}^{n} I_{gi} M_{i} \bar{y}_{i}}{\sum_{i=1}^{n} M_{i}}$$

(2.3.12)
Since $\hat{y}_g$ is also a ratio estimator of $\bar{y}_g$, $E(\hat{y}_g) = \bar{y}_g$ for large $n$, $g = 1, 2, ..., K$. As before, our local area estimator takes the form:

$$\hat{y}_l. = \sum_{g=1}^{K} \frac{M_{lg}}{M_l} \hat{y}_g$$

(2.3.13)

such that its expectation and bias have the same form as the two-stage sampling design.

2.4. An Expression for the Mean Squared Error of the Local Area Estimator

It has already been observed that the local area estimator $\hat{y}_l.$ is biased. Consequently, the mean squared error term takes the form:

$$E[(\hat{y}_l. - \bar{y}_l.)^2] = E(\hat{y}_l. - E(\hat{y}_l.))^2 + (E(\hat{y}_l.) - \bar{y}_l.)^2$$

$$= \text{Variance}(\hat{y}_l.) + \text{(Bias)}^2.$$  

(2.4.1)

Since

$$E(\hat{y}_l.) = \sum_{g=1}^{K} \frac{M_{lg}}{M_l} \bar{y}_g,$$

where $\bar{y}_l.$ is a linear combination of the ratio estimators $\hat{y}_g$, $g = 1, 2, ..., K$ (with negligible bias), the variance of $\hat{y}_l.$ can be approximated by

$$\text{Var}(\hat{y}_l.) = \sum_{g=1}^{K} \left( \frac{M_{lg}}{M_l} \right)^2 \text{Var}(\hat{y}_g)$$

$$+ \sum_{g \neq g'} \sum_{g' \neq g} \left( \frac{M_{lg}}{M_l} \right) \left( \frac{M_{lg'}}{M_l'} \right) \text{Cov}(\hat{y}_g, \hat{y}_g')$$

(2.4.2)
If we also assume

\[
\sum_{i=1}^{n} \frac{I_{g1} M_{1}}{g_{i} M_{1}} \bar{y}_{i} = \frac{\sum_{i=1}^{n} I_{g1} M_{1} (\bar{y}_{i} - \bar{y})}{n} - \frac{\sum_{i=1}^{n} I_{g1} M_{1} (\bar{y}_{i} - \bar{y})}{n \left( \frac{N_{..}}{N} \right)}
\]

then

\[
\text{Var}(\hat{y}_{g}) = \frac{(N_{..} - n)}{n N_{..}} \left( \frac{N_{..}}{M_{g}} \right)^2 \frac{\sum_{i=1}^{N} I_{g1} M_{1} (\bar{y}_{1} - \bar{y}_{g})^2}{(N_{..} - 1)}
\]

\[
+ \frac{N_{..}}{n M_{g}^2 \frac{m_{i}}{M_{1}}} \left( 1 - \frac{m_{i}}{M_{1}} \right) \frac{\sum_{i=1}^{M_{i}} (y_{ij} - \bar{y}_{1})^2}{(m_{i} - 1)}
\]

This is the standard form of the approximate variance of a ratio estimator for a two-stage sampling design where the base units have equal probabilities of selection. Here, the first term represents the between base unit component of the variance, whereas the second denotes the within-base unit contribution. A nearly unbiased sample estimate of \(\text{Var}(\hat{y}_{g})\) takes the form:

\[
\text{Var}(\hat{y}_{g}) = \left( \frac{N_{..} - n}{n N_{..}} \right) \left( \frac{N_{..}}{M_{g}} \right)^2 \frac{\sum_{i=1}^{n} I_{g1} M_{1} (\bar{y}_{1} - \hat{y}_{g})^2}{(n - 1)}
\]

\[
+ \frac{N_{..}}{n M_{g}^2 \frac{m_{i}}{M_{1}}} \left( 1 - \frac{m_{i}}{M_{1}} \right) \frac{\sum_{i=1}^{m_{i}} (y_{ij} - \bar{y}_{1})^2}{(m_{i} - 1)}
\]
Since our two-stage sampling design requires the independent selection of subsamples from different sample base units, and the respective strata estimators are defined in terms of the indicator variables \( I_{g1} \), it can be shown that \( \text{Cov}(\hat{\gamma}_g, \hat{\gamma}_g') = 0 \). Hence, the mean squared error of our small area estimator can be expressed as:

\[
\text{MSE}(\hat{\gamma}_g) = \sum_{g=1}^{K} \left( \frac{M_g}{M_i} \right)^2 \text{Var}(\hat{\gamma}_g) + (\text{Bias})^2
\]

(2.4.6)

When a single-stage cluster sampling design is considered, the variance of \( \hat{\gamma}_g \) can be approximately expressed as

\[
\text{Var}(\hat{\gamma}_g) \approx \sum_{i=1}^{N} \left( \frac{M_i}{M} \right)^2 \left( \bar{y}_{i1} - \bar{y}_{ig} \right)^2 \left( 1 - \frac{n}{N} \right) \frac{M_i}{M} \left( \frac{N - 1}{N} \right)^2 \frac{1}{n}
\]

(2.4.7)

Here, it can also be shown that \( \text{Cov}(\hat{\gamma}_g, \hat{\gamma}_g') = 0 \). Consequently, the mean square error of the small area estimator has the same form as that expressed above.

2.4.1. Proof for \( \text{Cov}(\hat{\gamma}_g, \hat{\gamma}_{g'}) = 0 \)

The proof considers the two-stage sampling design first.

Since

\[
E(\hat{\gamma}_g) = \bar{y}_g \quad \text{for} \quad n \geq 100 \quad (g = 1, 2, \ldots, K),
\]

the covariance of \( \hat{\gamma}_g \) and \( \hat{\gamma}_{g'} \) \((g \neq g')\) can be written as

\[
\text{Cov}(\hat{\gamma}_g, \hat{\gamma}_{g'}) = E[(\hat{\gamma}_g - \bar{y}_g)(\hat{\gamma}_{g'} - \bar{y}_{g'})]
\]

(2.4.8)
If we also assume \( \sum_{i=1}^{n} I_{g_i} M_i = n \left( \frac{\bar{y}_g}{\bar{y}} \right) \) , then

\[
\text{Cov}(y_g', y_{g'}) = E \left[ \sum_{i=1}^{n} I_{g_i} M_i (\bar{y}_i' - \bar{y}_g') \sum_{i=1}^{n} I_{g_i} M_i (\bar{y}_i - \bar{y}_g) \right] \frac{M}{n(\frac{\bar{y}_g}{\bar{y}})} \frac{M}{n(\frac{\bar{y}_g}{\bar{y}})}
\]

\[
= E \left[ \sum_{i=1}^{n} I_{g_i} M_i \frac{2}{M} (\bar{y}_i - \bar{y}_g)(\bar{y}_i' - \bar{y}_g') \right] \frac{M}{n^2(\frac{\bar{y}_g}{\bar{y}})(\frac{\bar{y}_g}{\bar{y}})}
\]

\[
+ E \left[ \sum_{i<j} (I_{g_i} I_{g_j} + I_{g_i} I_{g_j}) M_i M_j (\bar{y}_i - \bar{y}_g)(\bar{y}_j - \bar{y}_g) \right] \frac{M}{n^2(\frac{\bar{y}_g}{\bar{y}})(\frac{\bar{y}_g}{\bar{y}})}
\]

\[(2.4.9)\]

But the \( I_{g_i} \)'s are defined such that \( I_{g_i} \cdot I_{g_i} = 0 \) for all \( i \); thus,

\( E(\text{Term 1}) = 0 \). Also,

\[
E(\text{Term 2}) = E_{\text{all}} E_{\text{base}} E_{\text{base}} E_{\text{units in units}} E_{\text{units in sample}} [(\text{Term 2})]
\]

where \( E_{\text{base}} [(\text{Term 2})] =
\[
\sum_{i<j} (I_{g_i} I_{g_j} + I_{g_i} I_{g_j}) M_i M_j (\bar{y}_i - \bar{y}_g)(\bar{y}_j - \bar{y}_g) \]
\[
\frac{M}{n^2(\frac{\bar{y}_g}{\bar{y}})(\frac{\bar{y}_g}{\bar{y}})}
\]

\[(2.4.10)\]

since the subunits are selected independently in different base units.

To find \( E_{\text{all}} E_{\text{base}} [(\text{Term 2})] \), we average over all possible selections
of the \( \sum_{i<j}^{n} \) \( = n(n-1)/2 \) base units, assuming equal probabilities of selection. This yields

\[
E[(\text{Term 2})] = \frac{\sum_{i<j}^{n(n-1)} \left(I_{g_i} I_{g_j} + I_{g_j} I_{g_i}^\prime\right) M_{g_i} M_{g_j} (\bar{Y}_i - \bar{Y}_g) (\bar{Y}_j - \bar{Y}_g)}{N \cdot (N - 1) \cdot n^2 \left( \frac{M}{N} \right) \left( \frac{M^\prime}{N} \right)}
\]

\[
= \frac{(n-1)N}{n(N-1)M_{g}M_{g}^\prime} \left[ \sum_{g}^{N} M_{g} (\bar{Y}_1 - \bar{Y}_g^\prime) \right] \left[ \sum_{g^\prime}^{N} M_{g^\prime} (\bar{Y}_1 - \bar{Y}_g) \right]
\]

(summed only over the base units in strata \( g \) or \( g^\prime \))

\[
= \frac{(n-1)N}{n(N-1)M_{g}M_{g}^\prime} \left[ M_{g} \bar{Y}_g - M_{g^\prime} \bar{Y}_g^\prime \right] \left[ M_{g^\prime} \bar{Y}_g - M_{g} \bar{Y}_g^\prime \right]
\]

\[
= 0 . \tag{2.4.11}
\]

Hence, the covariance of \( \hat{y}_g \) and \( \hat{y}_g^\prime \) is approximately equal to 0 for large \( n \).

When a single-stage cluster sampling design is considered, the result follows more readily. Here,

\[
\text{Cov}(\hat{y}_g, \hat{y}_g^\prime) = [1 - \left( \frac{n}{N} \right)] \frac{\sum_{i=1}^{N} I_{g_i} I_{g_i}^\prime M_{g_i}^2 (\bar{Y}_i - \bar{Y}_g) (\bar{Y}_i - \bar{Y}_g^\prime)}{M_{g}M_{g}^\prime(N-1)}
\]

\[
= 0 . \tag{2.4.12}
\]

for \( g = 1, 2, \ldots, K \), \( g^\prime = 1, 2, \ldots, K \), and \( g \neq g^\prime \). By the definition of the \( I_{g_i}^\prime \)'s, clearly the product of \( I_{g_i} \) and \( I_{g_i}^\prime \) is 0. Hence, the result \( \text{Cov}(\hat{y}_g, \hat{y}_g^\prime) = 0 \) follows immediately.
CHAPTER III

A REFORMULATION OF THE KALSBEEK MODEL:
SOME ANALYTICAL AND EMPIRICAL INVESTIGATIONS

3.1. Introduction

An analytical expression for the mean squared error of our local area estimator has been derived in the previous chapter. Yet, the inherent bias of the model does not allow for tests of its precision unless another unbiased estimate or the true value of the criterion variable is obtained at the local level. In practice, this is usually unavailable and is the motivating reason that alternative strategies are necessary to consider.

In order to determine the accuracy of the small area estimator and allow for comparisons of precision with respect to other strategies, we attempt to express the relationship between criterion and symptomatic variables by means of a probabilistic model. The model enables one to determine the true value of the criterion variable for target areas of interest and to approximate the bias and mean squared error of the respective local area estimators, and provides a framework for comparisons.

3.2. Determination of Stratum Boundaries

As noted, our small area estimator of a criterion variable for the $l^{th}$ local area using the Kalsbeek model takes the form
To avoid unnecessary complications which would occur with the multi-stage sampling design, we reconsider the single stage cluster sample design, adding the restriction that all target base units consist of the same number of elements. As described in the first chapter, strata (groups) are to be formed which are optimally homogeneous within, while simultaneously dissimilar between themselves. When the underlying relationship between the criterion and symptomatic variables is unknown, the strategy that has been entertained consists of forming groups by minimizing their within sum of squares while maximizing their between sum of squares, using only the sample data on the symptomatic variables. However, when a certain probabilistic model is entertained, another reasonable strategy is to determine those boundaries on the predictor variables which minimize the mean square error of \( \hat{y}_l \). Here, each local area estimator usually consists of a different weighted linear combination of the respective stratum estimators. Consequently, the boundaries which are optimal for small area \( l \) would not necessarily be so for small area \( l' \).

One solution would be to determine the optimal strata boundaries on the symptomatic variables which minimize the mean squared error of the criterion variable estimator for the over-all population. This estimator is actually the weighted average of all small area estimators, weighted by the respective proportion of elements belonging to the particular small area.

As before,
\[ \hat{y} = \frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} I_{gi} M_i} \]  

(3.2.2)

where \( M_i = M \) for \( i = 1, 2, \ldots, N \).

and because we are now considering a single-stage cluster design,

\[ \bar{y}_i = \frac{\sum_{j=1}^{M} y_{ij}}{M} = \bar{y}_i, \]

and therefore,

\[ \hat{y} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{n} I_{gi} \bar{y}_i}{\sum_{i=1}^{M} \sum_{j=1}^{n} I_{gi}} = \frac{\sum_{g=1}^{n} \bar{y}_i}{\sum_{g=1}^{n}}. \]  

(3.2.3)

Consequently,

\[ \hat{y} = \sum_{l=1}^{L} \frac{M_{l} \hat{y}_{l}}{\sum_{l=1}^{L} M_{l}} = \sum_{l=1}^{L} \frac{M_{l} \hat{y}_{l}}{M} \sum_{g=1}^{K} \frac{M_{lg} \hat{y}_{lg}}{M_{lg}}. \]

\[ = \sum_{l=1}^{L} \sum_{g=1}^{K} \frac{M_{lg} \hat{y}_{lg}}{M_{lg}} = \sum_{g=1}^{K} \frac{M_{lg} \hat{y}_{lg}}{M_{lg}} \sum_{l=1}^{L} \frac{M_{lg} \hat{y}_{lg}}{M_{lg}}. \]

\[ = \sum_{g=1}^{K} \frac{M_{lg} \hat{y}_{lg}}{M_{lg}} = \sum_{g=1}^{K} \frac{M_{lg} \hat{y}_{lg}}{M_{lg}}. \]  

(3.2.4)

since \( M_{lg} = N_{lg} \) and \( M_{lg} = N_{lg} \).

We also note that this linear combination of local area estimators is an approximately unbiased estimator of the criterion variable for the over-all population:
\[
E(\hat{y}_{..}) = \sum_{g=1}^{K} \frac{N}{g} \cdot E(\hat{y}_g) = \sum_{g=1}^{K} \frac{N}{g} \cdot \frac{\sum_{i=1}^{N} I_{g_i} M \bar{y}_i}{N \cdot M} \\
= \sum_{g=1}^{K} \frac{N}{g} \cdot \frac{\sum_{i=1}^{N} \bar{y}_i}{N \cdot M} = \sum_{i=1}^{N} \frac{\sum_{g=1}^{K} \sum_{i=1}^{N} y_{ij}}{N \cdot M}.
\] (3.2.5)

Since the estimator is approximately unbiased, our mean squared error term is actually the variance of the overall population estimator. We must determine the boundaries on the symptomatic variables which will minimize \(Var(\hat{y}_{..})\). Here, we are faced with the additional problem of working with a linear combination of post-stratified estimators.

For any fixed sample size \(n\) out of \(N\) base units, the \(n_g\), \(g = 1, 2, \ldots, K\) (\(K\) fixed) are random, subject only to the restriction
\[
\sum_{g=1}^{K} n_g = n.
\]
Because the variance of a post-stratified estimator is most similar to that of a stratified estimator with proportional allocation, it would be reasonable to use those boundaries on the symptomatic variables which are optimal here. The strategy is most appropriate when
\[
\sum_{g=1}^{K} n_g /K
\]
is reasonably large, since the post-stratified estimator's variance approaches that of the stratified estimator's variance (considering proportional allocation) when this occurs. One may wonder whether the ideal allocation of the sample is not the Neyman allocation. The noted resemblance of the post-stratified scheme to proportional allocation is one reason. The other is that gains from proportional allocation are nearly as great as those from Neyman allocation. Optimal strata boundaries (for minimum variance stratification) often yield strata variances which are homoscedastic, where the two
allocations are nearly identical. By adopting this strategy, we can not only determine the precision of the criterion variable estimator under a stratified sampling scheme with proportional allocation, but also ascertain the loss in precision that occurs when using the linear combination of post-stratified estimators first suggested by Kalsbeek.

Daleniус (1957) and Singh and Sukhatme (1972) have considered the case of minimum variance stratification when a single auxiliary variable was used as the stratification variable. They showed that for a particular allocation (i.e., Neyman, proportional), the boundaries on the auxiliary variable must satisfy a set of minimal equations. Since these equations are ill adapted to practical computation, a quick approximate method has been developed by Daleniус and Hodges (1959), known as the $\text{CUM}\sqrt{f}$ rule, and has been shown to be quite efficient. Thomsen (1975) has found that by taking equal intervals using the $\text{CUM}^3\sqrt{f}$ rule, approximately optimum stratum boundaries are determined which compare favorably with those derived by the $\text{CUM}\sqrt{f}$ rule.

Often, the stratification scheme will depend on more than one variable. Here as well, several methods have been developed which consider the problem of determining those stratum boundaries which are optimal in the sense of minimum variance stratification. Daleniус (1957) has proposed as a criteria the minimization of a weighted average of the relative difference between $\sigma_i^{2(M)}$, the variance of the criterion variable estimator using the multivariate stratification boundaries, and $\sigma_i^{2(0)}$, the variance using the univariate optimum stratification points for each stratification variable (weighted by their relative importance). Similarly, Murthy (1967) proposed maximizing
where the sum is taken over all the stratification variables. Their respective solutions, however, are rather unfeasible from a computational point of view. Anderson (1976) suggests a method which uses the CUMV/\sqrt{f} rule (or CUM3V/\sqrt{f} rule) along each marginal stratifier such that the product of the number of strata for each variable equals \( \prod_{i=1}^{p} K(K_i = K) \). The method is not optimal, but is practical. It has been shown to yield estimators that are more precise, then, when only one strong stratifier is used. Another practical method, suggested by Kalsbeek (1973), allows for the determination of boundaries at successive stages of stratification. Approximately optimum boundaries are obtained for the most significant stratifier, then for the second conditioned on the stratum means of the first, and so forth until all the stratification variables have been included. In the research that follows, both the methods advanced by Anderson and Kalsbeek are considered.

3.3. A Reformulation of the Kalsbeek Model

We wish to consider the case of sampling from super populations with specified continuous multivariate distributions. To use such an approach requires rather strong underlying assumptions regarding the nature of relationships between the criterion and symptomatic variables. To be consistent in getting the finite population results to conform to the new scheme, we disregard the finite population correction factors. The work here resembles the analytic formulation considered by
Kalsbeek. His model, however, was limited to the consideration of
the bivariate case (one criterion and one symptomatic variable), whereas
the intent here is to consider the more general multivariate setting.
Another major difference concerns the method used to determine the respec-
tive stratum boundaries. Kalsbeek attempted to estimate the optimal
boundaries by using only sample information and determined them by
using the order statistics of the sample (on the symptomatic variable).
To apply such a scheme to multivariate stratification would prove
intractable. Contrarily, we consider the methods advanced by Anderson
and Kalsbeek in Section (3.2) directly in the determination of multivariate
stratum boundaries. Finally, the model considered by Kalsbeek at this
stage took the form

$$\hat{\gamma}(K) = \sum_{g=1}^{K} \pi_{lg} \hat{\gamma}_g$$

(3.3.1)

where $\pi_{lg}$ was defined as the probability a target base unit from the
$\ell$th local area would be classified in the $g$th strata. Once again,
Kalsbeek used the order statistics to determine the approximate proba-
bilities of classification. Since we are considering a single-stage
cluster sampling design with the restriction that all target base units
consist of the same number of elements, our small area estimator re-
mains

$$\hat{Y}_{\ell} = \sum_{g=1}^{K} \frac{M_{lg}}{M_{\ell}} \hat{y}_g$$

$$= \sum_{g=1}^{K} \frac{N_{lg} M_{lg}}{N_{\ell} M_{\ell}} \hat{y}_g = \sum_{g=1}^{K} \frac{N_{lg}}{N_{\ell}} \hat{y}_g$$

(3.3.2)

where $\frac{N_{lg}}{N_{\ell}} = \frac{\text{# of target base units falling in } g \text{th strata for } \ell \text{th local area}}{\text{Total # of target base units in the } \ell \text{th local area}}$
and

\[
\hat{y}_g = \frac{M \sum_{i=1}^{n} I_{gi} \bar{y}_i}{M \sum_{i=1}^{n} I_{gi}} = \frac{\sum_{g}^{n_g} \bar{y}_i}{n_g},
\]

(3.3.3)

where \( n_g \) (the number of sample base units falling in the \( g \)th stratum) is random. Consequently,

\[
E(\hat{y}_g) = \sum_{g=1}^{K} \frac{N_g n_g}{N} E(\hat{y}_g)
\]

(3.3.4)

where

\[
E(\hat{y}_g) = E_{n_g} E_{n_g}(\bar{y}_g) = E_{n_g} \left( \frac{\sum_{g}^{n_g} \bar{y}_i}{n_g} \right) = E_{n_g} \left( \frac{n_g}{n_g} E_{n_g}(\bar{y}_i) \right)
\]

and, if we assume \( n_g \neq 0 \) for \( g = 1, 2, \ldots, K \),

\[
E_{n_g}(\bar{y}_i) = E_{g^{th} \text{ strata}}(\bar{y}_i)_{n_g \text{ fixed}}.
\]

(3.3.5)

Similarly, we have shown

\[
\text{Var}(\hat{y}_g) = \sum_{g=1}^{K} \left( \frac{N_g n_g}{N} \right)^2 \text{Var}(\hat{y}_g)
\]

(3.3.6)

where, considering the underlying sampling design, the variance of \( \hat{y}_g \) can be re-expressed as

\[
\text{Var}(\hat{y}_g) = \text{Var}(n_g) E_{n_g}(\hat{y}_g) + E(n_g) \text{Var}_{n_g}(\hat{y}_g)
\]

(3.3.7)

The first term is zero since \( E_{n_g}(\hat{y}_g) \) is a constant, and

\[
E_{n_g} \left( \frac{\sum_{g}^{n_g} \bar{y}_i}{n_g} \right) = E(n_g) \left[ \frac{\text{Var}_{n_g}(\bar{y}_i)}{n_g} \right]
\]

(ignoring the finite population correction), and again when \( n_g \neq 0 \).
for \( g = 1, 2, \ldots, K \). Stephan (1947) has shown that to terms of order \( n^{-2} \)

\[
E_n(n_g) \left( \frac{1}{n_g} \right) \leq \left[ \frac{1}{n W_g} + \frac{1 - W_g}{n^2 W_g^2} \right]
\]  (3.3.9)

where \( W_g, g = 1, 2, \ldots, K \) are the respective stratum weights.

Therefore,

\[
\text{Var}(\bar{y}_g) \leq \left[ \frac{1}{n W_g} + \frac{1 - W_g}{n^2 W_g^2} \right] \text{Var}_{\text{g th stratum}}(\bar{y}_i) \text{, n fixed} \]  (3.3.10)

and

\[
\text{Var}(\hat{y}_g) \leq \sum_{g=1}^{K} \left( \frac{N_g}{N} \right)^2 \left[ \frac{1}{n W_g} + \frac{1 - W_g}{n^2 W_g^2} \right] \text{Var}_{\text{g th stratum}}(\bar{y}_i) \text{, n fixed} \]  (3.3.11)

3.4. The Theoretical Framework

Assume a simple random sample of size \( n \) is drawn from an infinite \( p+1 \) dimensional multivariate population (with continuous distribution) whose observations take the form of the \(((p+1)\times 1)\) random vector \((y, x_1, x_2, \ldots, x_p)\). Here, the \( y \) element conforms to the \( \bar{y}_1 \) cluster mean, while the \((x_1, x_2, \ldots, x_p)\) are symptomatic indicators which conform to those for each target base unit. The joint density of the multivariate super population is \( f(y, x_1, x_2, \ldots, x_p) \) with marginal probability density functions \( f_1(y), f_2(x_1), \ldots, f_{p+1}(x_p) \).

\[
E_{((p+1)\times 1)}(y) = \begin{bmatrix} u_y \\ u_{x_1} \end{bmatrix} \text{ are the respective means of the criterion and symptomatic variables while } \text{Var}_{(p+1)\times(p+1)} \left[ \begin{array}{c@{\quad}c@{\quad}c} \sigma^2_y & \sigma_{yX_1} & \vdots \\ \sigma_{yX_1} & \sigma^2_{x_1} & \vdots \\ \vdots & \vdots & \sigma^2_{x_j} \end{array} \right] = \Sigma
\]
is the respective variance-covariance matrix assumed to be positive definite.

Once the underlying multivariate distribution has been specified, we are able to construct target areas of interest for fixed values of \( N_l \). Here, the respective target base units are represented by \( N_l \) \((1 \times p)\) vectors of symptomatic information taking the form \((x_{11}, x_{12}, \ldots, x_{1p})\). These are determined by taking the values of equally spaced percentiles on the respective marginal distributions of the symptomatic variables over different ranges of interest such that their product is equal to \( N_l \). To be more explicit, consider the bivariate case with \( N_l = 49 \) and the 20\(^{th}\) to 80\(^{th}\) percentile as the range of interest on each marginal stratifier. The values of the equally spaced, cross classified percentiles observed in the following diagram determine the target area's symptomatic information configuration.

![Diagram](image)

With the number of strata \((K)\) fixed, the multivariate stratum
boundaries are of form

\[(a_1 < x_1 < b_1, a_2 < x_2 < b_2, \ldots, a_p < x_p < b_p)\]

which are rectilinear, non-overlapping and exhaustive. Consequently, the expected value of \(\overline{y}_i\) for the \(g^{th}\) strata (\(n_g\) fixed), \(E|_{n_g}(\overline{y}_i)\) in Section 3.3 is equivalent to

\[E(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}, \ldots, a_{pg} < x_p < b_{pg}) = \]

\[\int_{a_{1g}}^{b_{1g}} \ldots \int_{a_{pg}}^{b_{pg}} \frac{E(y|x) g(x) \, d\, x}{w_g} \]  \hfill (3.4.1)

assuming the underlying multivariate distribution. Anderson (1976) has shown that

\[E(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}) = \]

\[\int_{a_{1g}}^{b_{1g}} \ldots \int_{a_{pg}}^{b_{pg}} \frac{E(y|x) \, d\, x}{w_g} \]

where \(E(y|x)\) is the conditional expectation of \(y\) given \(x\);

\[g(x) = \int_{-\infty}^{\infty} f(y, x_1, x_2, \ldots, x_p) \, d\, y \]  \hfill (3.4.2)

is the respective joint density function of the symptomatic variables; and

\[w_g = \int_{a_{1g}}^{b_{1g}} \ldots \int_{a_{pg}}^{b_{pg}} g(x) \, d\, x = \Pr\{a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}\} \]  \hfill (3.4.3)

is the probability of being in the \(g^{th}\) strata. Therefore,

\[\hat{E}(\overline{y}_g) = \sum_{g=1}^{K} \frac{N_g}{N} E(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}) \]  \hfill (3.4.4)

Similarly, the variance of \(\overline{y}_i\) for the \(g^{th}\) strata (\(n_g\) fixed), \(\text{Var}_{n_g}(\overline{y}_i)\) in Section 3.3 is equivalent to
\[ \text{Var}(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}, \ldots, a_{pg} < x_p < b_{pg}) , \]

for which Anderson has derived the expression

\[
\int_{a_{1g}}^{b_{1g}} \ldots \int_{a_{pg}}^{b_{pg}} \frac{\text{Var}(y|x) g(x) d(x)}{W_g} + \int_{a_{1g}}^{b_{1g}} \ldots \int_{a_{pg}}^{b_{pg}} [E(y|x) - E(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg})]^2 g(x) d x
\]

(3.4.5)

where \( \text{Var}(y|x) \) is the conditional variance of \( y \) given \( x \). Consequently,

\[
\hat{\text{Var}}(\hat{\mu}_{\ell}) =
\sum_{g=1}^{K} \left( \frac{N_{\ell g}}{N_{\ell}} \right)^2 \left[ \frac{1}{n_{\ell g}} W_g + \frac{1 - W_g}{n_{\ell g} W_g / 2} \right] \text{Var}(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}) .
\]

(3.4.6)

Were the allocation of the sample fixed before it was drawn using proportional allocation (\( n_g = n_{\ell g} W_g \)) and the same multivariate strata considered, one could determine how much precision is lost when using the post-stratification scheme. Here, the variance of the stratified small area estimator is \( \hat{\text{Var}}(\hat{\mu}_{\ell}) \) (stratified)

\[
\sum_{g=1}^{K} \left( \frac{N_{\ell g}}{N_{\ell}} \right)^2 \left( \frac{1}{n_{\ell g}} \right) \text{Var}(y|a_{1g} < x_1 < b_{1g}, \ldots, a_{pg} < x_p < b_{pg}) .
\]

(3.4.7)

3.4.1. Determination of the Bias

In Section 2.3, the true value of a criterion variable of interest for local area \( \ell \) was defined as

\[
\bar{Y}_{\ell} = \sum_{\ell} \frac{N_{\ell}}{M_1 \bar{Y}_1 / M_2} .
\]

(3.4.8)
for the two-stage sampling design. Similarly,

\[ \bar{y}_{\ell} = \frac{\sum_{i \in \ell} \frac{N_i}{n_i} \bar{y}_i}{\sum_{i \in \ell} \frac{N_i}{n_i}} = \frac{\sum_{i \in \ell} \frac{N_i}{n_i} \bar{y}_i}{N_{\ell}}. \]  

(3.4.9)

for the single-stage cluster design with target base units having the same number of elements. In the theoretical framework considered, \( \bar{y}_{\ell} \) has been defined as a function of the vector of symptomatic information \( (x_1, x_2, \ldots, x_p) \) for different target areas of interest. Here,

\[ \bar{y}_{\ell} = \frac{\sum_{i \in \ell} \frac{N_i}{n_i} E(y|x)}{N_{\ell}}. \]  

(3.4.10)

for \( x = (x_1, x_2, \ldots, x_p) \) fixed. Consequently, the bias of our post-stratified local area estimator \( \hat{\bar{y}}_{\ell} \) can be approximated by:

\[ \text{Bias}_{(p.s.)} \approx \sum_{g=1}^{K} \frac{N_{lg}}{N_{\ell}} E(y|a_{lg} < x_1 < b_{lg}, \ldots, a_{pg} < x_p < b_{pg}) - \frac{\sum_{g} \frac{N_{lg}}{N_{\ell}} E(y|x)}{N_{\ell}}. \]  

(3.4.11)

Also, the mean squared error of \( \hat{\bar{y}}_{\ell} \) can be approximated by

\[ \text{M.S.E.}(\hat{\bar{y}}_{\ell}) = \sum_{g=1}^{K} \left( \frac{N_{lg}}{N_{\ell}} \right)^2 \left[ \frac{1}{n_{wg}} + \frac{1}{n_{wg}^2} \right] \text{Var}(y|a_{lg} < x_1 < b_{lg}, \ldots, a_{pg} < x_p < b_{pg}) + \text{Bias}^2_{(p.s.)}. \]  

(3.4.12)

3.5. Estimation Using the Ericsken Model

To allow for a comparison of the method's accuracy, we reconsider the Ericsken model which is applicable in the same general setting. Here,
the least squares regression estimator is determined using data obtained from the sample base units. Estimates of the criterion variable for the respective target area base units are then derived by substituting their vectors of symptomatic information into the resulting equation. The model of Ericksen is represented by:

\[
\hat{\mathbf{y}}_{\ell} = \frac{\sum_{s=1}^{N_{\ell}} \mathbf{x}_{s} \hat{B}(E)}{N_{\ell}} = \hat{B}_0 + \bar{x}_{\ell 1} \hat{B}_1 + \bar{x}_{\ell 2} \hat{B}_2 + \ldots + \bar{x}_{\ell p} \hat{B}_p \tag{3.5.1}
\]

where \( \mathbf{x}^{\sim}(E) = (1, x_{11}, x_{12}, x_{13}, \ldots, x_{ip}) \) is a \((1 \times (p+1))\) vector of symptomatic information from the \(i^{th}\) base unit in the \(\ell^{th}\) target area;

\[
\hat{B}(E) = \begin{bmatrix} \hat{B}_0 \\ \hat{B}_1 \\ \vdots \\ \hat{B}_p \end{bmatrix}
\]

is the \((p+1) \times 1\) vector of the least squares regression coefficients determined by the criterion and symptomatic variable information for the \(n\) sample base units;

\(N_{\ell}\) is the number of base units in the \(\ell^{th}\) target area;

and \(\bar{x}_{\ell s}, s = 1, 2, \ldots, p\) is the \(s^{th}\) symptomatic variable's mean for the \(\ell^{th}\) target area.

For \(\mathbf{x}^{\sim} = (x_1, x_2, \ldots, x_p)\) fixed, the expectation of the Ericksen estimator can be expressed as

\[
E(\hat{\mathbf{y}}_{\ell}(E)) = \frac{\sum_{s=1}^{N_{\ell}} \mathbf{x}^{\sim}(E) E(|x(\hat{B}^{(E)}))}{N_{\ell}} = E(|x(\hat{B}_0)) + \bar{x}_{\ell 1} E(|x(\hat{B}_1)) + \bar{x}_{\ell 2} E(|x(\hat{B}_2)) + \ldots + \bar{x}_{\ell p} E(|x(\hat{B}_p)) \tag{3.5.2}
\]

Similarly, its variance is of form
\[
\text{Var}(\hat{y}_{k^*}) = \text{Var}(\hat{b}_0) + \bar{x}_{1}^2 \text{Var}(\hat{b}_1) + \bar{x}_{2}^2 \text{Var}(\hat{b}_2) + \ldots + \\
+ \bar{x}_{p}^2 \text{Var}(\hat{b}_p) + 2\bar{x}_{1} \bar{x}_{p} \text{Cov}(\hat{b}_0, \hat{b}_1) \\
+ 2\bar{x}_{2} \text{Cov}(\hat{b}_0, \hat{b}_2) + \ldots + 2\bar{x}_{p} \text{Cov}(\hat{b}_0, \hat{b}_p) \\
+ 2\bar{x}_{1} \bar{x}_{2} \text{Cov}(\hat{b}_1, \hat{b}_2) + \ldots + 2\bar{x}_{p-1} \bar{x}_{p} \text{Cov}(\hat{b}_{p-1}, \hat{b}_p),
\]

(3.5.3)

again for \(x\) fixed. Consequently, the bias of the Ericksen estimator for the \(k^{th}\) target area is

\[
\frac{N_{k^*}}{N_{k^*}} \frac{E(y|x)}{E(y | x)} = \text{Bias}(\hat{y}_{k^*})
\]

(3.5.4)

Finally, the mean squared error of \(\hat{y}_{k^*}(E)\) is

\[
\text{M.S.E.}(\hat{y}_{k^*}(E)) = \text{Var}(\hat{y}_{k^*}(E)) + \text{Bias}^2(\hat{y}_{k^*}(E)).
\]

(3.5.5)

3.6. Target Area Estimation for the Trivariate Normal Distribution

To give our findings a degree of validity beyond the scope of the theoretical framework, the relationship between criterion and symptomatic variables must be characterized by those distributions most relevant to the practical setting. Since the vector \((y, x_1, x_2, \ldots, x_p)\) of criterion and symptomatic variables has been defined to represent a vector of cluster means, their distributions approach the normal when the underlying distributions are not markedly skewed. Consequently, the first distribution we have chosen to consider is the multivariate normal. To facilitate the presentation, we examine the trivariate case.
where the random vector \( \mathbf{y}' = (y, x_1, x_2) \) has a three dimensional multivariate normal distribution with joint density function

\[
f(y) = \frac{\exp\left(\frac{-1}{2}(y - \mu_y)' \Sigma_y^{-1}(y - \mu_y)\right)}{(2\pi)^{3/2} |\Sigma_y|^{1/2}}
\]

(3.6.1)

\( -\infty < y, x_1, x_2 < \infty \)

with

\[
E(y) = \mu_y = \begin{bmatrix} \mu_y \\ \mu_x_1 \\ \mu_x_2 \end{bmatrix}
\]

and

\[
\Sigma_y = \begin{bmatrix} \sigma_y^2 & \sigma_y x_1 & \sigma_y x_2 \\ \sigma_y x_1 & \sigma_x^2 & \sigma_{x_1 x_2} \\ \sigma_y x_2 & \sigma_{x_1 x_2} & \sigma_{x_2}^2 \end{bmatrix}
\]  

(assumed to be positive definite)

Here,

\[
E(y|x) = \mu_y - \Sigma_y^{-1} \Sigma_x (\mu_x - x)
\]

(3.6.2)

and

\[
\text{Var}(y|x) = \sigma_y^2 - \Sigma_y^{-1} \Sigma_x \Sigma_y^{-1} \Sigma_x
\]

(3.6.3)

where \( \Sigma_y = (B_1, B_2) \) is a vector of regression coefficients. To determine the expected value of the post-stratified target area estimator when the underlying distribution is trivariate normal, we must consider

\[
E(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g})
\]

for the respective multivariate strata. In this setting,
\[ E(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}) = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} E(y|x) \frac{g(x)}{w_g} \, dx \]

\[ = \mu_y - B^\prime \frac{g(x)}{\approx} + \frac{2}{i=1} B_1 \mu_{x_{1g}} \]

\[ = \mu_y - B^\prime (\mu_x - \mu_{x_{g}}) \quad (3.6.4) \]

where \( g(x) = g(x_1, x_2) = \)

\[ \exp\left\{ \frac{-1}{2(1 - \rho^2)} \left[ \frac{(\frac{x_1 - \mu_{x_1}}{\sigma_{x_1}})^2 - 2\rho \left( \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}} \right) \left( \frac{x_2 - \mu_{x_2}}{\sigma_{x_2}} \right) + \left( \frac{x_2 - \mu_{x_2}}{\sigma_{x_2}} \right)^2}{2\pi \sigma_{x_1} \sigma_{x_2}^{/1 - \rho}} \right] \right\} \]

Consequently,

\[ E(\hat{y}_g) = \sum_{g=1}^{K} \frac{N_{lg}}{N_{g}} E(y|x) \]

\[ = \mu_y - B^\prime \frac{g(x)}{\approx} + \sum_{g=1}^{K} \frac{N_{lg}}{N_{g}} (B^\prime \frac{x_g}{\approx}) \quad (3.6.5) \]

with

\[ \text{Bias (p.s.)} = E(\hat{y}_g) - \sum_{g=1}^{K} \frac{N_{lg}}{N_{g}} (\mu_y - B^\prime \frac{x_g}{\approx}) \]

\[ = B^\prime \frac{\mu_x}{\approx} + \sum_{g=1}^{K} \frac{N_{lg}}{N_{g}} (B^\prime \frac{\mu_{x_g}}{\approx}) - \frac{N_{g}}{N_{g}} (\mu_y - B^\prime \frac{x_g}{\approx}) \]

\[ = \frac{K}{g=1} \frac{N_{lg}}{N_{g}} (B^\prime \frac{\mu_{x_g}}{\approx} - B^\prime \frac{x_{lg}}{\approx}) = \sum_{g=1}^{K} \frac{N_{lg}}{N_{g}} B^\prime \frac{\mu_{x_g}}{\approx} - \frac{N_{lg}}{N_{g}} \right) \quad (3.6.6) \]
where \( \bar{X}_g \) is the vector:

\[
\begin{bmatrix}
\bar{x}_{1g} \\
\bar{x}_{2g}
\end{bmatrix}
\]

Similarly, we must consider \( V(y | a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}) \) to determine the variance of \( \hat{Y}_g \). From Anderson (1976) we note that

\[
\text{Var}(y | a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}) = \\
\int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{\text{Var}(y | x) g(x) \, dx}{W_g} + \\
\int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} (E(y | x) - E(y | a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}))^2 \frac{g(x) \, dx}{W_g}
\]

\[
= \sigma^2_{yg} - B^\prime \Sigma B + \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \left[ B^\prime (x - \bar{x}_g) \right]^2 \frac{g(x) \, dx}{W_g}
\]

\[
= \sigma^2_{yg} - B^\prime \Sigma B + \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{2}{B_{i,j}} \left[ \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{(x_i - \mu_{x_{ig}})(x_j - \mu_{x_{jg}}) \, g(x) \, dx}{W_g} \right]
\]

\[
= \sigma^2_{yg} - B^\prime \left( \Sigma - \Sigma_g \right) B
\]

Thus,

\[
\text{Var}(\hat{Y}_g) = \sum_{g=1}^{K} \left( \frac{N_g}{N_r} \right)^2 \left[ \frac{1}{n_g W_g} + \frac{1 - W_g}{n \frac{W_g}{w^2}} \right] \left[ \sigma^2_{yg} - B^\prime \left( \Sigma - \Sigma_g \right) B \right]
\]

(3.6.8)

and the mean squared error of \( \hat{Y}_g \) can be determined by adding \( (\text{Bias}_{(p.s.)})^2 \) to its variance. Explicit expressions for \( W_g, \bar{X}_g, \), and \( \Sigma_g \) are given in Appendix A, in addition to the approximations used for their numerical evaluation.

When the underlying distribution is normal and the model of
Ericksen is considered, the target area estimator is unbiased. Here,

\[
E(y_{\xi}, (E)) = \frac{N_{\xi}}{L} \left( \frac{X_{(E)}^\prime E | X_{(E)}^\prime (\hat{B}_{(E)})}{N_{\xi}} \right)
= B_0 + \frac{N_{\xi} B_1 B_2}{N_{\xi}} \frac{X}{X_{\xi}}
= \mu_y + B_\xi (\bar{y} - \bar{X}_{\xi})
= \frac{N_{\xi}}{L} \frac{E(y | x)}{N_{\xi}}
\]

(3.6.9)

since \( \hat{B}_{(E)} \) is an unbiased estimator of \((B_0, B_1, B_2)\) for fixed \(x\).

Similarly, its variance is of form

\[
(1, \bar{X}_{\xi1}, \bar{X}_{\xi2}) \Sigma_{\hat{B}_{(E)} | X} (1, \bar{X}_{\xi1}, \bar{X}_{\xi2})^\prime
\]

(3.6.10)

where

\[
\Sigma_{\hat{B}_{(E)} | X} = (X_{(E)}^\prime X_{(E)})^{-1} \text{Var}(y | x).
\]

Here,

\[
\text{Var}(y | x) = \sigma^2_y - B_\xi X \Sigma_{\xi} B_\xi
\]

and \( (X_{(E)}^\prime X_{(E)}) = \)

\[
\frac{1}{|X_{(E)}^\prime X_{(E)}|} \begin{bmatrix}
  c_{11} & c_{12} & c_{13} \\
  c_{21} & c_{22} & c_{23} \\
  c_{31} & c_{32} & c_{33}
\end{bmatrix}
\]

(3.6.11)
where

\[ c_{11} = \sum_{i=1}^{n} x_{1i}^2 \left( \sum_{i=1}^{n} x_{i2} - \left( \sum_{i=1}^{n} x_{1i} x_{i2} \right)^2 \right) \]

\[ c_{22} = n \sum_{i=1}^{n} x_{i2}^2 - n^2 \bar{x}_2^2 \]

\[ c_{33} = n \sum_{i=1}^{n} x_{1i}^2 - n \bar{x}_1^2 \]

\[ c_{12} = c_{21} = -(n \bar{x}_1 \sum_{i=1}^{n} x_{i2}^2 - n \bar{x}_2 \sum_{i=1}^{n} x_{1i} x_{i2}) \]

\[ c_{13} = c_{31} = n \bar{x}_1 \sum_{i=1}^{n} x_{1i} x_{i2} - n \bar{x}_2 \sum_{i=1}^{n} x_{1i}^2 \]

\[ c_{23} = c_{32} = -(n \sum_{i=1}^{n} x_{1i} x_{i2} - n^2 \bar{x}_1 \bar{x}_2) \]

The respective elements of \((X^\top (E)^{-1} X (E)^{-1})\) have been shown to approach the matrix of terms:

\[
\begin{bmatrix}
\frac{1}{n} & a_{12} & a_{13} \\
\frac{a_{11}}{3} & a_{22} & a_{23} \\
a_{31} & a_{32} & a_{33}
\end{bmatrix}
\]

where

\[ a_{11} = n^2 \left( \frac{\sigma_x^2}{x_1} + \frac{\mu_x}{x_2} \right) \left( \frac{\sigma_y^2}{x_1} + \frac{\mu_y}{x_2} \right) - \left( \frac{\sigma_x}{x_1} \frac{\mu_x}{x_2} \right)^2 \]

\[ a_{22} = n^2 \frac{\sigma_x^2}{x_2} \]

\[ a_{33} = n^2 \frac{\sigma_x^2}{x_1} \]
\[ a_{12} = a_{21} = n^2 \left( \mu_{x_1} \sigma_{x_2} x_1 x_2 - \mu_{x_1} \sigma_{x_2}^2 \right) \]
\[ a_{13} = a_{31} = n^2 \left( \mu_{x_1} \sigma_{x_2} x_1 x_2 - \mu_{x_2} \sigma_{x_1}^2 \right) \]
\[ a_{23} = a_{32} = -n^2 \sigma_{x_1 x_2} \]

by application of a theorem attributed to Slutsky for large \( n \). In the work that follows, we use the large sample approximation.

3.7. Target Area Estimation for the Trivariate Logistic Distribution

Another continuous distribution of major interest to our research is the multivariate logistic distribution. The logistic curve has long been a valuable tool to demographers as a model for estimating population growth in respective geographical areas. Also, the marginal distributions of the multivariate logistic are quite similar to the normal. More importantly, since its curve of regression is non-linear in \( x \), we have a setting for which the Ericksen estimator is biased.

As before, we shall consider the trivariate case where the random vector \( y = (v_1, v_2, v_3) = (y_1, x_1, x_2) \) has the density function described by Gumbel (1961),

\[
f(y) = \frac{3! \left[ 1 + \sum_{i=1}^{3} \exp\left\{ -(v_i - \mu_{v_i}) / \xi_{v_i} \right\} \right]^{-4} \exp\left\{ -\sum_{i=1}^{3} \frac{(v_i - \mu_{v_i})}{\xi_{v_i}} \right\}}{\prod_{i=1}^{3} \xi_{v_i}}\]

\[-\infty < y < \infty \]

(3.7.1)

with

\[
E(y) = \begin{bmatrix} \mu_{v_1} \\ \mu_{v_2} \\ \mu_{v_3} \end{bmatrix} = \begin{bmatrix} \mu_y \\ \mu_{x_1} \\ \mu_{x_2} \end{bmatrix}
\]
and $\xi_i = \sigma_i \sqrt{3}$ such that the cumulative distribution function of $v_i$ is

$$F_{v_i}(v_i) = \left[1 + \exp\left(\frac{-(v_i - \mu v_i)}{\xi v_i}\right)\right]^{-1} \quad (3.7.2)$$

Here,

$$E(y|x) = \xi_y E(y^{(s)}|x^{(s)}) + \mu_y \quad (3.7.3)$$

can be expressed in terms of the standard form of the distribution

$$f^{(s)}(y^{(s)}) = 3!(1 + \sum_{i=1}^{3} e^{-v_i^{(s)}})^{-4} \exp\{-\sum_{i=1}^{3} v_i^{(s)}\} \quad (3.7.4)$$

where

$$v_i^{(s)} = \frac{v_i - \mu v_i}{\xi v_i} \quad \text{for } i = 1, 2, 3$$

and

$$E(y^{(s)}|x^{(s)}) = \frac{3}{2} - \ln(1 + \sum_{i=1}^{2} e^{-x_i^{(s)}}) \quad (3.7.5)$$

Similarly,

$$\text{Var}(y|x) = \xi_y^2 \text{Var}(y^{(s)}|x^{(s)}) \quad (3.7.6)$$

and Johnson and Kotz (1972) have shown

$$\text{Var}(y^{(s)}|x^{(s)}) = \psi^{(1)} + \psi^{(3)} = \psi^{(1)} + \psi^{(2)} - \frac{1}{4} = \frac{\pi^2}{3} - \frac{5}{4} \quad (3.7.7)$$

where $\psi(a)$ is the digamma function. Also,

$$\text{Cov}(v_i v_j) = \sigma_{v_i v_j} \quad \text{for all } i \neq j$$

$$= E(v_i v_j) - E(v_i) E(v_j) = \xi_{v_i} \xi_{v_j} \text{Cov}(v^{(s)}_i, v^{(s)}_j) = \frac{\sigma_{v_i} \sigma_{v_j}}{2} \quad (3.7.8)$$
To determine the expected value of the proposed target area estimator, we must consider

\[
E(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}) = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{E(y|x) g(x) \, dx}{W_g} \tag{3.7.9}
\]

where \(g(x)\) is the density function of the bivariate logistic distribution

\[
g(x) = \frac{2!}{\zeta_{x_1} \zeta_{x_2}} \left[ 1 + \sum_{i=1}^{2} \exp\{- (x_i - \mu_{x_i})/\zeta_{x_i}\} \right]^{-3} \exp\{- \frac{2}{\zeta_{x_1}} (x_1 - \mu_{x_1}) \} \tag{3.7.10}
\]

and

\[
W_g = \Pr\{a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}\} = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} g(x) \, dx \tag{3.7.11}
\]

By considering the change of variables \(x_1 = \zeta_{x_1} x_1^{(s)} - \mu_{x_1}\),

\[
E(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}) \text{ can be expressed as:}
\]

\[
\zeta_y \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{E(y^{(s)}|x^{(s)}) g^{(s)}(x^{(s)}) \, dx^{(s)}}{W^{(s)} g} + \mu_y, \tag{3.7.12}
\]

which further reduces to

\[
\zeta_y \left(\frac{3}{2}\right) + \mu_y - \zeta_y \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{\ln(1 + \sum_{i=1}^{2} e^{-x_i^{(s)}})}{W^{(s)} g} \tag{3.7.13}
\]

where

\[
g^{(s)}(x^{(s)}) = 2! \left[ 1 + \sum_{i=1}^{2} e^{-x_i^{(s)}} \right]^{-3} \exp\{- \sum_{i=1}^{2} x_i^{(s)} \},
\]
\[ a_{ig}^{(s)} = \frac{a_{ig} - \mu x_i}{\zeta_{x_i}}, \quad b_{ig}^{(s)} = \frac{b_{ig} - \mu x_i}{\zeta_{x_i}} \quad \text{for } i = 1, 2, \]

and \( W_g^{(s)} = W_g \). To evaluate \( W_g^{(s)} \), we consider the integral

\[
\int_{h_1}^{k_1} \int_{h_2}^{k_2} 2\{1 + e^{-z_1} + e^{-z_2}\}^{-3} e^{-z_1} e^{-z_2} \, d\, z_2 \, d\, z_1
\]

\[
= \int_{h_1}^{k_1} e^{-z_1} \left[ \int_{h_2}^{k_2} (1 + e^{-z_1} + e^{-z_2})^{-3} e^{-z_2} \, d\, z_2 \right] \, d\, z_1 \quad (3.7.14)
\]

Letting \( u = (1 + e^{-z_1} + e^{-z_2}) \), \( d\, u = -e^{-z_2} \, d\, z_2 \), the term in brackets is equivalent to

\[
(1 + e^{-z_1} + e^{-k_2})^{-2} - (1 + e^{-z_1} + e^{-h_2})^{-2} \quad (3.7.15)
\]

and for \( c \) constant,

\[
\int_{h_1}^{k_1} e^{-z_1} (1 + e^{-z_1} + e^{-c})^{-2} \, d\, z_1
\]

\[
= (1 + e^{-k_1} + e^{-c})^{-1} - (1 + e^{-h_1} + e^{-c})^{-1} \quad (3.7.16)
\]

Consequently,

\[
W_g^{(s)} = (1 + e^{-b_{1g}} + e^{-b_{2g}})^{-1} - (1 + e^{-a_{1g}} + e^{-a_{2g}})^{-1}
\]

\[
- (1 + e^{-b_{1g}} + e^{-2g})^{-1} + (1 + e^{-a_{1g}} + e^{-2g})^{-1} \quad (3.7.17)
\]

In order to evaluate \( E(y | a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}) \), we must reconsider the integral.
\[
\int_{h_1}^{k_1} \int_{h_2}^{k_2} \ln(1 + e^{-z_1} + e^{-z_2}) \cdot 2(1 + e^{-z_1} + e^{-z_2})^{-3} e^{-z_1} e^{-z_2} \, dz_2 \, dz_1 \\
= \int_{h_1}^{k_1} e^{-z_1} \cdot 2[\int_{h_2}^{k_2} \ln(1 + e^{-z_1} + e^{-z_2})(1 + e^{-z_1} + e^{-z_2})^{-3} e^{-z_2} \, dz_2] \, dz_1
\]

(3.7.18)

Letting \( u = \ln(1 + e^{-z_1} + e^{-z_2}) \),
\[
\begin{align*}
    du &= -(1 + e^{-z_1} + e^{-z_2})^{-1} e^{-z_2} \, dz_2, \\
    dv &= 2(1 + e^{-z_1} + e^{-z_2})^{-3} e^{-z_2} \, dz_2, \text{ and} \\
    v &= (1 + e^{-z_1} + e^{-z_2})^{-2},
\end{align*}
\]

and integrating by parts, the term in brackets reduces to
\[
{1 + e^{-z_1} + e^{-z_2}}^{-2} \left[\log(1 + e^{-z_1} + e^{-z_2}) + \frac{1}{2}\right] \bigg|_{h_2}^{k_2}
\]

(3.7.19)

For \( c \) constant,
\[
\int_{h_1}^{k_1} {1 + e^{-z_1} + e^{-c}}^{-2} \ln(1 + e^{-z_1} + e^{-c}) e^{-z_1} \, dz_1 \\
= \left[1 + e^{-z_1} + e^{-c}\right]^{-1} [\ln(1 + e^{-z_1} + e^{-c}) + 1] \bigg|_{h_1}^{k_1},
\]

(3.7.20)

letting \( u = \ln(1 + e^{-z_1} + e^{-c}) \), \( dv = (1 + e^{-z_1} + e^{-c})^{-2} e^{-z_1} \, dz_1 \), and integrating by parts.

Also,
\[
\int_{h_1}^{k_1} (1 + e^{-z_1} + e^{-c})^{-2} \frac{e^{-z_1}}{2} \, dz_1 = \frac{1}{2}(1 + e^{-z_1} + e^{-c})^{-1} \bigg|_{h_1}^{k_1}
\]

(3.7.21)
Therefore,

\[
\int_{b_{1g}}^{b_{2g}} \int_{a_{1g}}^{a_{2g}} \ln(1 + \sum_{i=1}^{2} e^{-x_i}) g(s(x)) \, d(x) \, d(s(x))
\]

\[
= (1 + e^{-x_1} + e^{-x_2})^{-1} \left[ \ln(1 + e^{-x_1} + e^{-x_2}) + \frac{3}{2} \right] \int_{b_{1g}}^{b_{2g}} \int_{a_{1g}}^{a_{2g}} \left( s(x) \right) \left( s(x) \right) \, d(s(x)) \, d(s(x))
\]

\[
- (1 + e^{-x_1} + e^{-x_2})^{-1} \left[ \ln(1 + e^{-x_1} + e^{-x_2}) + \frac{3}{2} \right] \int_{b_{1g}}^{b_{2g}} \int_{a_{1g}}^{a_{2g}} \left( s(x) \right) \left( s(x) \right) \, d(s(x)) \, d(s(x))
\]

(3.7.22)

Similarly, \( \text{Var}(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}) \) must be evaluated to determine the variance of \( \frac{\hat{y}}{\hat{\lambda}} \). As before, we can express the results in terms of the standardized logistic distribution. Here,

\[
\text{Var}(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g})
\]

\[
= \int_{b_{1g}}^{b_{2g}} \int_{a_{1g}}^{a_{2g}} \text{Var}(y|x) g(x) \, d(x) \, d(s(x))
\]

\[
+ \int_{a_{1g}}^{a_{2g}} \int_{b_{1g}}^{b_{2g}} \left[ \text{E}(y|x) - \text{E}(y|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}) \right]^2 g(x) \, d(x) \, d(s(x))
\]

\[
= \xi^2 \int_{b_{1g}}^{b_{2g}} \int_{a_{1g}}^{a_{2g}} \text{Var}(y|x) g(x) \, d(x) \, d(s(x))
\]

\[
+ \xi^2 \int_{a_{1g}}^{a_{2g}} \int_{b_{1g}}^{b_{2g}} \frac{\text{E}(y|x)}{g(x)} \, d(x) \, d(s(x))
\]

\[
\frac{\text{E}(y(s)|x(s)) - \text{E}(y(s)|a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g})^2}{g(x)} \, d(x)
\]

\[
\frac{(s)}{W(s)}
\]
\[ \zeta_y^2 (\psi^-(1) + \psi^-(3)) + \zeta_y^2 \int_{b_1g}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{E^2(y(s)|x(s)) g(s) x(s)}{W(s)} d\frac{d}{d s} \]

\[ - \zeta_y^2 (E(y(s)|a_{1g}) < x_1(s) < b_{1g}, a_{2g} < x_2(s) < b_{2g})^2 \]

\[ = \zeta_y^2 \left( \frac{\pi}{3} - \frac{5}{4} \right) + \zeta_y^2 \int_{b_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{(\ln(1 + \sum_{i=1}^{2} e^{-x_i(s)})^2 g(s) x(s)}{W(s)} d\frac{d}{d s} \]

\[ - \zeta_y^2 \int_{1g}^{b(s)} \int_{a_{2g}}^{b(s)} (\ln(1 + \sum_{i=1}^{2} e^{-x_i(s)}) g(s) x(s)}{W(s)} d\frac{d}{d s} \]

(3.7.23)

Here,

\[ \int_{h_1}^{k_1} \int_{h_2}^{k_2} (\ln(1 + \sum_{i=1}^{2} e^{-z_i}))^2 2(1 + e^{-z_1} + e^{-z_2})^{-3} e^{-z_1} e^{-z_2} dz_2 dz_1 \]

\[ = \int_{h_1}^{k_1} e^{-z_1} \int_{h_2}^{k_2} (\ln(1 + \sum_{i=1}^{2} e^{-z_i}))^2 2(1 + e^{-z_1} + e^{-z_2})^{-3} e^{-z_2} dz_2 \]

(3.7.24)

Letting \( u = (\ln(1 + \sum_{i=1}^{2} e^{-z_i}))^2 \), \( dv = 2(1 + e^{-z_1} + e^{-z_2})^{-3} e^{-z_2} dz_2 \)

and integrating by parts twice, the term in brackets reduces to

\[ \left( 1 + e^{-z_1} + e^{-z_2} \right)^{-2} [\ln^2(1 + e^{-z_1} + e^{-z_2}) + \ln(1 + e^{-z_1} + e^{-z_2}) + \frac{1}{2}] \]

(3.7.25)
For \( c \) constant,

\[
\int_{h_1}^{k_1} \left( 1 + e^{-z_1 + c} \right)^{-2} \ln^2(1 + e^{-z_1 + c}) e^{-z_1} dz_1
\]

\[
= \{1 + e^{-z_1 + c}\}^{-1} \left[ \ln^2(1 + e^{-z_1 + c}) + 2\ln(1 + e^{-z_1 + c}) + 2 \right] \bigg|_{h_1}^{k_1},
\]

\[
(3.7.26)
\]

\[
\int_{h_1}^{k_1} \left( 1 + e^{-z_1 + c}\right)^{-2} \ln(1 + e^{-z_1 + c}) e^{-z_1} dz_1
\]

\[
= \{1 + e^{-z_1 + c}\}^{-1} \left[ \ln(1 + e^{-z_1 + c}) + 1 \right] \bigg|_{h_1}^{k_1},
\]

\[
(3.7.27)
\]

and

\[
\int_{h_1}^{k_1} \left( 1 + e^{-z_1 + c}\right)^{-2} \frac{e^{-z_1}}{2} dz_1 = \left\{ \frac{1 + e^{-z_1 + c}}{2} \right\}^{-1} \bigg|_{h_1}^{k_1}.
\]

\[
(3.7.28)
\]

Therefore,

\[
\int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \ln^2(1 + \sum_{i=1}^{2} e^{-x_i(s)}) g(s) x(s) d x(s)
\]

\[
= \{1 + e^{-x_1(s)} + e^{-b_{2g}(s)}\}^{-1}
\]

\[
\cdot \left[ \ln^2(1 + e^{-x_1(s)} + e^{-b_{2g}(s)}) + 3\ln(1 + e^{-x_1(s)} + e^{-b_{2g}(s)}) + \frac{7}{2} \right] \bigg|_{a_{1g}}^{b_{1g}}
\]

\[
- \{1 + e^{-x_1(s)} + e^{-a_{2g}(s)}\}^{-1}
\]

\[
\cdot \left[ \ln^2(1 + e^{-x_1(s)} + e^{-a_{2g}(s)}) + 3\ln(1 + e^{-x_1(s)} + e^{-a_{2g}(s)}) + \frac{7}{2} \right] \bigg|_{a_{1g}}^{b_{1g}}
\]

\[
(3.7.29)
\]
In this setting the true value of the criterion variable for local area \( l \) is
\[
\bar{y}_l = \frac{\sum_{l \in L} E(y_{l} | x_l)}{N_L} = \mu_y + \frac{3}{2} \zeta_y - \frac{\zeta_y}{N_L} \ln \left( 1 + e^{\left( -\frac{x_1 - \mu_{x_1}}{\zeta_{x_1}} \right)} + e^{\left( -\frac{x_2 - \mu_{x_2}}{\zeta_{x_2}} \right)} \right)
\tag{3.7.30}
\]

Since the regression curve \( E(y_{l} | x_l) \) is nonlinear in \( x_l \), we have a setting for which the Ericksen model is biased. Here,
\[
\hat{E}(\bar{y}_l) = \frac{\sum_{l \in L} \hat{X}_{l}^\prime (E) E(\hat{B}(E))}{N_L}, \tag{3.7.31}
\]
which for large \( n \) (using Slutsky's Theorem), can be approximated by
\[
B_0 + \frac{\sum_{l \in L} \hat{X}_{l} \hat{B}(E)}{N_L} = B_0 + B_1 \bar{X}_{l_1} + B_2 \bar{X}_{l_2}, \tag{3.7.32}
\]
where \( B = (\sigma_{y_{x_1}}, \sigma_{y_{x_2}}) \Sigma^{-1} \) and \( B_0 = \mu_y - \frac{\zeta_y}{\zeta} x \). Similarly, its variance is of form
\[
(1, \bar{X}_{l_1}, \bar{X}_{l_2}) \Sigma_{\hat{B}(E)} \times (1, \bar{X}_{l_1}, \bar{X}_{l_2})^\prime \tag{3.7.33}
\]
where \( \Sigma_{\hat{B}(E)} = (\hat{X}_{l_1}^\prime, \hat{X}_{l_2}^\prime)^{-1} \text{Var}(y_{l} | x), \text{Var}(y_{l} | x) = \zeta_y^2 \frac{H^2}{3} - \frac{5}{4} \), and the elements of \( \frac{\hat{X}_{l_1}^\prime, \hat{X}_{l_2}^\prime}{(E)}^{-1} \) approach the same terms as in (3.6.12) for large \( n \).

3.8. Distribution Specific Results

To determine the accuracy of our post-stratified target area estimator and compare its precision with respect to the Ericksen model, the following settings are specified:
(1) Underlying Trivariate Normal Distribution with a high association level \((R = .95)\)

\[
\begin{pmatrix}
y \\ x_1 \\ x_2
\end{pmatrix} =
\begin{pmatrix}
60 \\ 50 \\ 50
\end{pmatrix}
\text{ and } \Sigma =
\begin{bmatrix}
100 & 42.5 & 42.5 \\
42.5 & 25 & 15 \\
42.5 & 15 & 25
\end{bmatrix}
\]

(2) Underlying Trivariate Normal Distribution with a low association level \((R = .58)\)

\[
\begin{pmatrix}
y \\ x_1 \\ x_2
\end{pmatrix} =
\begin{pmatrix}
60 \\ 50 \\ 50
\end{pmatrix}
\text{ and } \Sigma =
\begin{bmatrix}
100 & 25 & 25 \\
25 & 25 & 12.5 \\
25 & 12.5 & 25
\end{bmatrix}
\]

(3) Underlying Trivariate Logistic Distribution with level of association corresponding to \((R = .58)\)

\[
\begin{pmatrix}
y \\ x_1 \\ x_2
\end{pmatrix} =
\begin{pmatrix}
60 \\ 50 \\ 50
\end{pmatrix}
\text{ and } \Sigma =
\begin{bmatrix}
100 & 25 & 25 \\
25 & 25 & 12.5 \\
25 & 12.5 & 25
\end{bmatrix}
\]

The target areas we consider consist of \(N_{x} = 49\) target base units whose representation is given in Section 3.4 with (.2, .8), (.05, .95) and (.35, .95) as the ranges of interest. These coincide with the symptomatic variable configuration for many of the local areas encountered in practice. Two values of \(n\), the number of base units in the design, are given: \(n=120\), \(n=480\). These values are similar to the number of sample P.S.U.'s (counties or contiguous counties) used to provide respective regional and national estimates of parameters in the Current Population Survey.

The number of strata we consider varies as \(K = b^2\), where \(b = 2, 3, 4\) is the number of boundaries on each marginal stratifier. When the underlying distribution is trivariate normal, two alternative strategies are used in the determination of the stratum boundaries. The first method
is attributed to Anderson (1976) whereby marginally optimum stratum boundaries for proportional allocation are selected. These are given by Sethi for the standardized normal variate:

<table>
<thead>
<tr>
<th>b</th>
<th>Optimum Boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>-.61, .61</td>
</tr>
<tr>
<td>4</td>
<td>-.99, 0.0, .99</td>
</tr>
</tbody>
</table>

The other, attributed to Kalsbeek, is a hierarchical scheme described in Section 3.2. The respective stratum boundaries are shown in Figure (3.8.1) for the standardized normal variates when \( \rho_{x_1x_2} = .6 \). The same boundaries are used for \( \rho_{x_1x_2} = .5 \) to improve the target area estimator's precision.

When the underlying distribution is trivariate logistic, we use Anderson's approach with the CUM/\( F \) rule. To implement this procedure on each marginal stratifier, we consider the theoretically infinite population to be finite and of size 10,000. Selecting the 0.5th and 99.5th percentiles as endpoints, we construct 100 equally spaced intervals on the range of the distribution, determine their respective frequencies, and apply the CUM/\( F \) rule. Here,

<table>
<thead>
<tr>
<th>b</th>
<th>Stratum Boundaries Using CUM/( F ) Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>-0.99, 0.99</td>
</tr>
<tr>
<td>4</td>
<td>-1.57, 0.0, 1.57</td>
</tr>
</tbody>
</table>

We knew a priori that the Ericksen model was most appropriate to the linear setting, being an unbiased target area estimator when the underlying continuous distribution in multivariate normal. This is confirmed in the high association model under study (\( R \approx .95 \)). When the level of association is seriously reduced (\( R \approx .58 \)), the superiority of the linear estimator is nowhere as clear. At the same time, we note gains in precision for the post-stratified estimator when the hierarchical scheme of stratum boundary determination is employed. This is reflected in both the variance and mean squared error terms. Similarly, we note
Figure (3.8.1)

Stratum Boundaries Using Hierarchical Scheme

\[ b=2 \]
\[
\begin{align*}
X_2 & \leq -.479 \\
X_1 & \leq .0 \\
X_2 & > -.479 \\
\text{(Stratum Mean } &= -.798) \\
X_1 & > 0.0 \\
X_2 & \leq .479 \\
\text{(S.M. } &= .798)
\end{align*}
\]

\[ b=3 \]
\[
\begin{align*}
X_2 & \leq -1.222 \\
X_1 & \leq -.61 \\
-1.22 < X_2 & \leq -2.46 \\
X_2 & > -.246 \\
\text{(S.M. } &= -1.223) \\
X_2 & \leq -.488 \\
-.61 & < X_1 \leq .61 \\
-.488 & < X_2 \leq .488 \\
X_2 & > .488 \\
\text{(S.M. } &= 0.0)
\end{align*}
\]

\[ b=4 \]
\[
\begin{align*}
X_2 & \leq -1.702 \\
-1.702 < X_2 & \leq -.910 \\
X_1 & \leq -.99 \\
-.910 & < X_2 \leq -.118 \\
X_2 & > -.118 \\
\text{(S.M. } &= -1.517) \\
X_2 & \leq -1.066 \\
-1.066 & < X_2 \leq -.274 \\
-.99 & < X_1 \leq 0.0 \\
-.274 & < X_2 \leq .518 \\
X_2 & > .518 \\
\text{(S.M. } &= -.456)
\end{align*}
\]

\[ b=5 \]
\[
\begin{align*}
X_2 & \leq -.518 \\
-.518 & < X_2 \leq .274 \\
.0 & < X_1 \leq .99 \\
.274 & < X_2 \leq 1.066 \\
X_2 & > 1.066 \\
\text{(S.M. } &= .456)
\end{align*}
\]

\[ b=6 \]
\[
\begin{align*}
X_2 & \leq .118 \\
.118 & < X_2 \leq .910 \\
X_1 & > .99 \\
.910 & < X_2 \leq 1.702 \\
X_2 & > 1.702 \\
\text{(S.M. } &= 1.517)
\end{align*}
\]
gains in precision for both estimators with an increase in sample size. Consequently, when the sample size is large and a hierarchical scheme is employed, the post-stratified estimator does reasonably well for the linear setting.

When attention is directed to the non-linear setting of the trivariate logistic distribution, the merits of the proposed approach become more obvious. As before, we also note gains in precision for both estimators as reflected in the variance and mean squared error terms with increased sample size. Here, the inherent bias in the linear estimator generally dominates that of the post-stratified estimator. This is primarily a function of the lack of fit of the Ericksen model. Had we considered a trivariate setting with an even more striking non-linear curve of regression, the relative bias of the linear estimator would be greater. For each target area under consideration, there is at least one stratification scheme for n=120, and at least two for n=480, which demonstrate the post-stratified estimators' superiority using the mean squared error as the measure of precision. Had a more optimal scheme for the determination of strata boundaries been available, further increases in the precision of our post-stratified estimator could have been observed.

Generally, when stratification is the strategy used to yield an estimator of a criterion variable for a particular target population, an increase in the number of strata, K, is followed by an increase in the estimator's precision (as measured by a decrease in the variance) for relatively small values of K. Subsequent increases in K coincide with diminishing returns with respect to further proportional reductions in the estimator's variance. Since each target area estimator under
consideration consists of a different weighted linear combination of
stratum estimators, and the sampled population does not completely
coincide with the target population, we did not expect to find strong
evidence of a consistent relationship between the proposed method's
precision and the number of strata to be specified. Nevertheless, an
empirical examination of our post-stratified estimator's performance
for the target areas under investigation appears to suggest the optimal
number of strata lies closest to 9, using the mean squared error as our
measure of precision.
<table>
<thead>
<tr>
<th>Stratification Scheme</th>
<th>Model</th>
<th>Strata (n=120)</th>
<th>Approximate Values for Criterion Parameters</th>
<th>True Value of Criterion Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Expectation</td>
<td>Variance</td>
</tr>
<tr>
<td>Optimal Boundaries</td>
<td>Ericksen</td>
<td>4</td>
<td>60.000</td>
<td>0.081</td>
</tr>
<tr>
<td>on Marginals</td>
<td>(Modified)</td>
<td>9</td>
<td>58.625</td>
<td>0.292</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>16</td>
<td>60.000</td>
<td>0.228</td>
</tr>
<tr>
<td></td>
<td>(Model)</td>
<td></td>
<td>59.288</td>
<td>0.263</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Ericksen</td>
<td>4</td>
<td>60.000</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>(Modified)</td>
<td>9</td>
<td>59.316</td>
<td>0.285</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>16</td>
<td>60.000</td>
<td>0.211</td>
</tr>
<tr>
<td></td>
<td>(Model)</td>
<td></td>
<td>59.773</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td>(n=480)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal Boundaries</td>
<td>Ericksen</td>
<td>4</td>
<td>60.000</td>
<td>0.020</td>
</tr>
<tr>
<td>on Marginals</td>
<td>(Modified)</td>
<td>9</td>
<td>58.625</td>
<td>0.071</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>16</td>
<td>60.000</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>(Model)</td>
<td></td>
<td>59.288</td>
<td>0.063</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Ericksen</td>
<td>4</td>
<td>60.000</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>(Modified)</td>
<td>9</td>
<td>59.316</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>16</td>
<td>60.000</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>(Model)</td>
<td></td>
<td>59.773</td>
<td>0.059</td>
</tr>
<tr>
<td>Stratification Scheme</td>
<td>Model</td>
<td>Strata (n=120)</td>
<td>Range (.2,.8)</td>
<td>True Value of Criterion Variable</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------</td>
<td>----------------</td>
<td>---------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approximate Values for Criterion Parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Expectation</td>
<td>Variance</td>
<td>Bias</td>
</tr>
<tr>
<td>Optimal Boundaries</td>
<td>Ericksen</td>
<td>60.00</td>
<td>0.556</td>
<td>0.000</td>
</tr>
<tr>
<td>on Marginals</td>
<td>(Modified)</td>
<td>59.145</td>
<td>0.731</td>
<td>-0.855</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>60.000</td>
<td>0.925</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>59.554</td>
<td>1.292</td>
<td>-0.446</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Ericksen</td>
<td>60.000</td>
<td>0.556</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(Modified)</td>
<td>59.580</td>
<td>0.720</td>
<td>-0.420</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>60.000</td>
<td>0.813</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>59.862</td>
<td>1.162</td>
<td>-0.138</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(n=480)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal Boundaries</td>
<td>Ericksen</td>
<td>60.000</td>
<td>0.139</td>
<td>0.000</td>
</tr>
<tr>
<td>on Marginals</td>
<td>(Modified)</td>
<td>59.145</td>
<td>0.178</td>
<td>-0.855</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>60.000</td>
<td>0.224</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>59.554</td>
<td>0.309</td>
<td>-0.446</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Ericksen</td>
<td>60.000</td>
<td>0.139</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(Modified)</td>
<td>59.580</td>
<td>0.180</td>
<td>-0.420</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>60.000</td>
<td>0.196</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>59.862</td>
<td>0.274</td>
<td>-0.138</td>
</tr>
<tr>
<td>Stratification Scheme</td>
<td>Model</td>
<td>Strata (n=120)</td>
<td>Strata (n=480)</td>
<td>Approximate Values for Criterion Parameters</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------</td>
<td>----------------</td>
<td>----------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Range (.2, .8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Expectation</td>
<td></td>
<td>Bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximate Optimal</td>
<td>Ericksen</td>
<td>60.000</td>
<td>60.000</td>
<td>0.517</td>
</tr>
<tr>
<td>Boundaries on</td>
<td>(Modified)</td>
<td>59.334</td>
<td>59.334</td>
<td>0.763</td>
</tr>
<tr>
<td>Marginals Using</td>
<td>Kalsbeek</td>
<td>60.956</td>
<td>60.956</td>
<td>0.869</td>
</tr>
<tr>
<td>Cum $\sqrt{T}$ Rule</td>
<td>(Model)</td>
<td>61.108</td>
<td>61.108</td>
<td>1.276</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |
|                       |            |                |                |      |        |                               |                                 |
|                       |            |                |                |      |        |                               |                                 |
|                       | Ericksen   | 60.000         | 60.000         | 0.129 | -1.296 | 1.808 | 61.295                     |
| (Modified)            | 4          | 59.334         | 59.334         | 0.186 | -1.962 | 4.036 | 61.295                     |
| Kalsbeek              | 9          | 60.956         | 60.956         | 0.210 | -0.340 | 0.325 | 61.295                     |
| (Model)               | 16         | 61.108         | 61.108         | 0.304 | -0.188 | 0.340 | 61.296                     |</p>
<table>
<thead>
<tr>
<th>Stratification Scheme</th>
<th>Model</th>
<th>Strata (n=120)</th>
<th>Approximate Values for Criterion Parameters</th>
<th>True Value of Criterion Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Boundaries on Marginals</td>
<td>Ericksen</td>
<td>60.000</td>
<td>0.081</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Modified</td>
<td>58.625</td>
<td>0.292</td>
<td>-1.375</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>60.000</td>
<td>0.311</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>59.278</td>
<td>0.476</td>
<td>-0.722</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Ericksen</td>
<td>60.000</td>
<td>0.081</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Modified</td>
<td>59.316</td>
<td>0.285</td>
<td>-0.684</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>60.000</td>
<td>0.215</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>59.588</td>
<td>0.218</td>
<td>-0.412</td>
</tr>
<tr>
<td>(n=480)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal Boundaries on Marginals</td>
<td>Ericksen</td>
<td>60.000</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Modified</td>
<td>58.625</td>
<td>0.071</td>
<td>-1.375</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>60.000</td>
<td>0.065</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>59.278</td>
<td>0.069</td>
<td>-0.722</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>Ericksen</td>
<td>60.000</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Modified</td>
<td>59.316</td>
<td>0.070</td>
<td>-0.684</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>60.000</td>
<td>0.051</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>59.588</td>
<td>0.048</td>
<td>-0.412</td>
</tr>
</tbody>
</table>
### TABLE 3.5
TARGET AREA ESTIMATION FOR TRIVARIATE NORMAL DISTRIBUTION WITH $R = .58$

<table>
<thead>
<tr>
<th>Stratification Scheme</th>
<th>Model</th>
<th>Strata (n=120)</th>
<th>Expected Values for Criterion Parameters</th>
<th>True Value of Criterion Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Expectation</td>
<td>Variance</td>
</tr>
<tr>
<td>Optimal Boundaries</td>
<td></td>
<td></td>
<td>--------------</td>
<td>----------</td>
</tr>
<tr>
<td>on Marginals</td>
<td></td>
<td></td>
<td>60.000</td>
<td>0.556</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.145</td>
<td>0.731</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60.000</td>
<td>0.926</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.548</td>
<td>1.118</td>
</tr>
<tr>
<td></td>
<td>Ericksen</td>
<td></td>
<td>60.000</td>
<td>0.556</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.580</td>
<td>0.731</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60.000</td>
<td>0.722</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.745</td>
<td>0.818</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(n=480)</td>
<td></td>
</tr>
<tr>
<td>Hierarchical</td>
<td></td>
<td></td>
<td>60.000</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.145</td>
<td>0.178</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60.000</td>
<td>0.205</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.548</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>Ericksen</td>
<td></td>
<td>60.000</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.580</td>
<td>0.179</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60.000</td>
<td>0.172</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.745</td>
<td>0.187</td>
</tr>
<tr>
<td>Stratification Scheme</td>
<td>Model</td>
<td>Strata (n=120)</td>
<td>Approximate Values for Criterion Parameters</td>
<td>True Value of Criterion Variable</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------</td>
<td>----------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Range (.05, .95)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Expectation</td>
<td>Variance</td>
</tr>
<tr>
<td>Approximate Optimal</td>
<td></td>
<td></td>
<td>60.000</td>
<td>0.517</td>
</tr>
<tr>
<td>Boundaries on</td>
<td></td>
<td></td>
<td>59.334</td>
<td>0.763</td>
</tr>
<tr>
<td>Marginals Using</td>
<td></td>
<td></td>
<td>59.925</td>
<td>0.903</td>
</tr>
<tr>
<td>Cum F Rule</td>
<td></td>
<td></td>
<td>59.796</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(n=480)</td>
<td>60.000</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.334</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.925</td>
<td>0.201</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>59.796</td>
<td>0.191</td>
</tr>
<tr>
<td>Stratification Scheme</td>
<td>Model</td>
<td>Strata (n=120)</td>
<td>Approximate Values for Criterion Parameters</td>
<td>True Value of Criterion Variable</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------</td>
<td>----------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Expectation</td>
<td>Variance</td>
</tr>
<tr>
<td>Optimal Boundaries on Marginals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ericksen</td>
<td>Modified</td>
<td>4</td>
<td>65.094</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64.124</td>
<td>0.366</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>9</td>
<td>65.751</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>16</td>
<td>65.369</td>
<td>0.274</td>
</tr>
<tr>
<td>Hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ericksen</td>
<td>Modified</td>
<td>4</td>
<td>65.094</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>63.799</td>
<td>0.360</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>9</td>
<td>65.102</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>16</td>
<td>64.562</td>
<td>0.246</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal Boundaries on Marginals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ericksen</td>
<td>Modified</td>
<td>4</td>
<td>65.094</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64.124</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>9</td>
<td>65.751</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>16</td>
<td>65.369</td>
<td>0.060</td>
</tr>
<tr>
<td>Hierarchical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ericksen</td>
<td>Modified</td>
<td>4</td>
<td>65.094</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>63.799</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>Kalsbeek</td>
<td>9</td>
<td>65.102</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>16</td>
<td>64.962</td>
<td>0.056</td>
</tr>
<tr>
<td>Stratification Scheme</td>
<td>Model</td>
<td>Strata (n=120)</td>
<td>Expectation</td>
<td>Variance</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------</td>
<td>---------------</td>
<td>-------------</td>
<td>----------</td>
</tr>
<tr>
<td><strong>Optimal Boundaries on Marginals</strong></td>
<td>Ericksen</td>
<td>4</td>
<td>63.196</td>
<td>0.726</td>
</tr>
<tr>
<td></td>
<td>(Modified)</td>
<td>9</td>
<td>63.622</td>
<td>1.103</td>
</tr>
<tr>
<td></td>
<td>(Kalsbeek)</td>
<td>16</td>
<td>63.385</td>
<td>1.069</td>
</tr>
<tr>
<td><strong>Hierarchical</strong></td>
<td>Ericksen</td>
<td>4</td>
<td>63.196</td>
<td>0.726</td>
</tr>
<tr>
<td></td>
<td>(Modified)</td>
<td>9</td>
<td>63.202</td>
<td>1.072</td>
</tr>
<tr>
<td></td>
<td>(Kalsbeek)</td>
<td>16</td>
<td>63.130</td>
<td>1.071</td>
</tr>
<tr>
<td>(n=480)</td>
<td><strong>Optimal Boundaries on Marginals</strong></td>
<td>Ericksen</td>
<td>4</td>
<td>63.196</td>
</tr>
<tr>
<td></td>
<td>(Modified)</td>
<td>9</td>
<td>63.622</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>(Kalsbeek)</td>
<td>16</td>
<td>63.385</td>
<td>0.241</td>
</tr>
<tr>
<td><strong>Hierarchical</strong></td>
<td>Ericksen</td>
<td>4</td>
<td>63.196</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td>(Modified)</td>
<td>9</td>
<td>63.202</td>
<td>0.257</td>
</tr>
<tr>
<td></td>
<td>(Kalsbeek)</td>
<td>16</td>
<td>63.130</td>
<td>0.247</td>
</tr>
<tr>
<td>Stratification Scheme</td>
<td>Model</td>
<td>Strata (n=120)</td>
<td>Approximate Values for Criterion Parameters</td>
<td>True Value of Criterion Variable</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------</td>
<td>----------------</td>
<td>---------------------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Approximate Optimal Boundaries on Marginals Using Cum ( \Phi ) Rule</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ericksen</td>
<td></td>
<td>63.034</td>
<td>0.659</td>
<td>-0.680</td>
</tr>
<tr>
<td>(Modified)</td>
<td>4</td>
<td>62.530</td>
<td>0.742</td>
<td>-1.184</td>
</tr>
<tr>
<td>Kalsbeek Model</td>
<td>9</td>
<td>63.865</td>
<td>0.924</td>
<td>0.151</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>63.340</td>
<td>0.856</td>
<td>-0.314</td>
</tr>
<tr>
<td>(n=480)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER IV

AN APPLICATION OF THE PROPOSED METHODOLOGY:
SMALL AREA ESTIMATION OF UNEMPLOYMENT

4.1. Introduction

To complete our investigation of the proposed strategy for estimating the parameters of local areas, we wish to observe its performance when applied to a setting more common to federal agencies, the commercial sector, planners and candidates alike. Again, we wish to demonstrate that when underlying assumptions regarding the relationship between criterion and symptomatic variables are too strict or unrealistic, the need for a more flexible approach is clear. In particular, we wish to confirm the suitability of the proposed method over the use of a regression estimator (the Ericksen model) when there is a definite non-linear trend in the underlying population. We shall also examine the effect of the number of specified strata upon the relative precision of our post-stratified estimator.

In the previous chapter, we noted that the inherent bias of the post-stratified estimator did not allow for comparisons of precision unless another unbiased estimate or the true value of the criterion variable was obtained at the local level. Consequently, our choice of criterion variables is restricted to those for which direct measurements are available at the local level. To satisfy this constraint, we have referred to available data from the 1970 census. In the analysis that follows, we have selected unemployment as the criterion
variable to be studied. The rationale for this choice is the potential for constructing non-linear settings for the underlying population by selection of suitable auxiliary variables for which local level information exists. In addition, unemployment statistics for local areas such as states, Standard Metropolitan Statistical Areas (SMSA's) and counties are required by various federal agencies to determine the distribution of funds for special manpower and revenue sharing programs. The Department of Labor is directed to prepare annual estimates of unemployment for states and SMSA's specifically for this purpose by the Comprehensive Employment and Training Act of 1973. Consequently, unemployment serves as a good representative from a large class of criterion variables for which post-censal small area estimates are in demand. Here, it is important to note that other methods for generating annual state and SMSA estimates of the unemployment rate have been developed. These methods include the "70-step" estimate of the employment rate as published in the 1974 Manpower Report of the President, and application of the Ericksen model using the following auxiliary variables which exhibit a linear trend with the unemployment rate: the percent of insured unemployment to total unemployment, the final annual 70-step estimate of the unemployment rate, and a synthetic estimate of the unemployment rate based on marital status-sex-age-race composition. Often, however, the relationship between criterion and available symptomatic information is more appropriately characterized by a non-linear model. Since we have stated our intent to concentrate upon settings for which non-linear trends exist, we consider only those auxiliary variables that satisfy this condition.
4.2. A Description of the Sampling Framework and the General Methodology


To implement our procedure for generating post-censal estimates of the desired parameters of local areas, we follow the path of Ericksen and Kalsbeek by referring to the Census Bureau's monthly Current Population Survey (C.P.S.) as our source of sample information. From its inception in 1940, the survey has served as the major source of current information on the labor force in the United States and provides post-censal estimates for several other demographic parameters of interest (i.e., per capita income). Consequently, it is perhaps the most natural choice for sample data when the criterion variable to be entertained is unemployment, as measured by the unemployment rate.

As mentioned in the literature review, the Current Population Survey is based on a national multistage probability sample of the population. Basically, the total land area of the United States is divided into 1,913 primary sampling units (P.S.U.'s), consisting of a single county or group of contiguous counties combined to achieve the greatest degree of heterogeneity with respect to income, urbanization, etc. In addition, each SMSA is considered as a separate P.S.U. Homogeneous P.S.U.'s in characteristics such as geographic region, population density, rate of growth in previous decade, percent nonwhite, principal industry, and so forth, are combined into 449 strata and a single P.S.U. is drawn from each with probability proportional to their 1960 population. Here, the largest 107 SMSA's constitute individual strata (self-representing) by themselves and are automatically part of the sample. This is true for five other areas which are not SMSA's. As of 1970, the size of the C.P.S. had expanded to approximately 58,500 housing units.
in a 449-area sample in 863 counties and independent cities.

The second stage consists of a further division of the sample P.S.U.'s into enumeration districts and a sample is selected systematically with respect to geographical location to insure the sample enumeration districts are distributed over the entire P.S.U. Again, the probability of selection of the sample enumeration districts is proportionate to its 1960 population. The final stage requires the selection of a cluster of approximately six households, referred to as Ultimate Sampling Units, belonging to the sample enumeration districts. For the majority of areas, generally urban centers, selection is achieved by reference to address lists compiled from the 1960 census. The remainder of the ultimate sampling units are selected by means of an area sample.

Due to the nature of the C.P.S. design, the level for which local area estimates of unemployment are to be generated is the state. This must be the case if the sample base units we have described in Chapters I, II, and III are to coincide with the primary sampling units of the C.P.S. Although it is preferred that the sample base units consist of the same type of constituent geographical sectors (i.e., individual counties), the effect of the change in using the P.S.U.'s is most likely to be negligible. The same approach has been taken by both Ericksen and Kalsbeek, who argued that since both counties and P.S.U.'s are arbitrary geographic areas, a county could be regarded as a small P.S.U. without losing the principle of the method.

4.2.2. Determination of the Stratum Boundaries

Our stratification scheme is closely modeled after the approach
taken in Chapter III, that being the determination of stratum boundaries which are approximately optimal in the sense of minimum variance stratification. In contrast to the strategy entertained by Kalsbeek which consisted of group formation via clustering algorithms using only sample data, our approach considers symptomatic information from the entire population of sample and target base units. As before, we are faced with the same set of constraints encountered when a multivariate stratification scheme is used. Once more, we turn to the method advanced by Anderson (1976) which calls for application of the \( \text{CUM}_p\sqrt{F} \) rule along each marginal stratifier such that the product of the number of strata for each variable equals \( \prod_{i=1}^{p} K_i = K \) (where \( K \) is specified in advance).

Here, however, a modification in the method is made. From the analytical results of Chapter III, we noted consistent gains in precision for the post-stratified estimator when the hierarchical scheme of stratum boundary determination was employed. Consequently, this plan is incorporated into the boundary determination design in the following manner: boundaries are obtained using Anderson's approach for the most significant stratifier, then for the second using the \( \text{CUM}_p\sqrt{F} \) rule only on those base units falling within the respective stratum boundaries of the first stratifier, and so forth until all the stratification variables have been considered. Once the strata have been formed, classification of the base units (here counties) with respect to their symptomatic configuration presents no difficulties. This is due to the partition of strata which are rectilinearly defined, non-overlapping and exhaustive.

4.2.3. The Estimation Procedure

With the strata defined and the sample base units appropriately
classified, we obtain an estimate of the $g^{th}$ stratum mean, $y_g$, in the following manner:

$$y_g = \frac{\sum_{i=1}^{n} I_{gi} M_i y_i}{\sum_{i=1}^{n} I_{gi} M_i} = \frac{\sum_{g=1}^{g} M_i y_i}{\sum_{g=1}^{g} M_i} \quad (4.2.1)$$

where $I_{gi}$ is an indicator variable defined in Section 2.3;

$n$ equals the 441 PSU's considered from the 1970 C.P.S.;

$y_i$ corresponds to the respective 1970 P.S.U. estimate of the unemployment rate; and

$M_i$ corresponds to the 1970 census population of the $i^{th}$ P.S.U.

Moreover, to obtain an estimate of the 1970 unemployment rate, $y_{\ell}$, for the $\ell^{th}$ state using the proposed strategy, we consider:

$$y_{\ell} = \frac{\sum_{g=1}^{K} M_{\ell g} y_g}{M_{\ell}} \quad (4.2.2)$$

where $K$ is the number of strata;

$M_{\ell g}$ is the 1970 population of all counties in state $\ell$ whose symptomatic configuration is aligned with stratum $g$; and

$M_{\ell}$ is the 1970 population of all counties in the $\ell^{th}$ state (the state population).

Since our primary concern is to determine how closely the post-stratified estimates of the respective unemployment rates "hit their mark," the measure of reliability to be considered is the relative absolute deviation, defined as:

$$\left| \frac{y_{\ell} - \bar{y}_{\ell}}{\bar{y}_{\ell}} \right| \quad (4.2.3)$$

where $\bar{y}_{\ell}$ is the predicted estimate and $\bar{y}_{\ell}$ is the actual value. This
criterion for measuring the accuracy of small area estimates is also
to be found in the work of Levy (1971), Gonzalez and Waksberg (1973),
Gonzalez (1973), Kalsbeek (1973), and Namekater et al. (1975).

To implement the Ericksen procedure, a linear least squares
regression estimate is obtained using criterion and symptomatic data
only for the P.S.U.'s. County estimates are then determined by sub-
stituting their respective symptomatic configuration into the resulting
model. To derive an estimate of the employment rate for the $i^{th}$ state,
the respective county estimates are weighted proportionately by their
1970 census populations.

4.3. State Estimates of Unemployment: An Application of the Proposed
Strategy and the Ericksen Model

In the analysis that follows, 1970 state estimates of unemploy-
ment, measured as a percent of the total civilian work force, are
generated for the continental United States. The symptomatic variables
selected for this study include the percent of families with female
head, $X_1$, and the percent of households with 1.01 or more persons per
room, $X_2$. It was our conviction that these auxiliary variables would
more accurately be characterized by a non-linear relationship with the
unemployment rate. A further investigation revealed a higher order
polynomial model to be more appropriate.

The Current Population Survey estimates that were used are an
average of five monthly estimates, spaced at quarterly intervals and
centered around April 1970. This was done to increase the accuracy
of the respective P.S.U. estimates. To determine the relative pre-
cision of the proposed method and the Ericksen model, the 1970 census
estimate of the unemployment rate (based on a 15% national sample) was
treated as the true value of the criterion variable.

To examine the effect of the number of specified strata upon
the relative precision of the post-stratified estimator, we initially
attempted to generate state estimates of unemployment for \( K = 6, 9, 
16, 25 \). The respective stratum boundaries which were determined using
a hierarchical scheme with the CUM/\( \sqrt{r} \) rule can be observed in Figure 4.3.1.
Here, no difficulties were encountered for \( K = 6, 9 \). When the primary
sampling units were allocated to their respective strata for \( K = 16, 25, 
\) however, some of the strata were characterized by a very limited number
of sample observations (\( n_k < 4 \)). To increase the precision of the res-
pective stratum means for the proposed method, then particular strata
were collapsed as indicated in Figure 4.3.1 such that \( K \) was reduced to
14 from 16 and to 21 from 25.

State estimates of unemployment which were derived using the
proposed strategy with different stratification schemes are to be found
in Table 4.1. The same is true for estimates obtained by the method of
Erickson. Here, the fitted regression coefficients using sample data
are observed in the following linear model:

\[
\hat{Y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2
\]

where \( \hat{B}_0 = 1.958, \hat{B}_1 = 0.205, \hat{B}_2 = 0.084; \) and

\( x_1, x_2 \) are the symptomatic variables described above.

4.4. An Interpretation of the Results

The data in Table 4.1 show the post-stratified state estimates
of unemployment to be generally more accurate than those determined by
Figure (4.3.1)

Determination of Stratum Boundaries Using
A Hierarchical Scheme with the Cum/F Rule

**K = 6 Strata**

\[ \begin{align*}
    &x_2 < 7. \\
    &7. \leq x_2 < 11. \\
    &x_2 \geq 11.
\end{align*} \]

\[ \begin{align*}
    &x_1 < 9. \\
    &9. \leq x_1 < 17.
\end{align*} \]

**K = 9 Strata**

\[ \begin{align*}
    &x_2 < 6. \\
    &6. \leq x_2 < 10. \\
    &x_2 \geq 10.
\end{align*} \]

\[ \begin{align*}
    &x_1 < 8. \\
    &8. \leq x_1 < 12.
\end{align*} \]

\[ \begin{align*}
    &x_1 \geq 12. \\
    &12. \leq x_2 < 19. \\
    &x_2 \geq 19.
\end{align*} \]

**K = 16 Strata (K = 16)**

\[ \begin{align*}
    &x_2 < 5. \\
    &5. \leq x_2 < 8. \\
    &8. \leq x_2 < 12. \\
    &x_2 \geq 12. \text{ collapsed}
\end{align*} \]

\[ \begin{align*}
    &x_1 < 7. \\
    &7. \leq x_1 < 11. \\
    &11. \leq x_1 < 16.
\end{align*} \]

\[ \begin{align*}
    &x_1 \geq 13. \\
    &16. \leq x_2 < 23. \text{ collapsed}
\end{align*} \]

\[ \begin{align*}
    &x_1 \geq 13. \\
\end{align*} \]
Figure 4.3.1 CONT.

$K = 25$ Strata ($K = 21$)

\[
\begin{align*}
X_2 &< 5. \\
5. &\leq X_2 < 7. \\
7. &\leq X_2 < 10. \\
-10. &\leq X_2 < 13. \text{ collapsed} \\
X_2 &> 13. \text{ collapsed}
\end{align*}
\]

\[
\begin{align*}
X_2 &< 6. \\
-6. &\leq X_2 < 8. \\
8. &\leq X_2 < 11. \\
-11. &\leq X_2 < 15. \\
X_2 &> 15.
\end{align*}
\]

\[
\begin{align*}
X_2 &< 5. \\
5. &\leq X_2 < 9. \\
6. &\leq X_2 < 8. \\
7. &\leq X_2 < 9. \\
-9. &\leq X_2 < 13. \text{ collapsed} \\
X_2 &> 13. \text{ collapsed}
\end{align*}
\]

\[
\begin{align*}
X_2 &< 8. \\
-8. &\leq X_2 < 11. \\
8. &\leq X_2 < 11. \\
-11. &\leq X_2 < 14. \\
11. &\leq X_2 < 14. \\
-14. &\leq X_2 < 20. \text{ collapsed} \\
X_2 &> 20.
\end{align*}
\]
the Ericksen model. This is most readily observed for the stratification scheme which specified $K=14$. Here, the proposed strategy yielded estimates which were equally precise or better in 32 of the 48 states under consideration, as measured by the relative absolute deviation described in Section 4.2.3. Variations in the degree of accuracy do exist, however, with respect to the different stratification schemes. When 6 strata were specified, the proposed method did not match its performance with $K=14$, yielding estimates equivalent (4 ties) or superior to the regression estimates in 29 states. A similar distribution was noticeable between the stratification schemes which considered $K=9$ and $K=21$ respectively. Estimates equal to or better than those derived using the Ericksen method were observed in 31 states for $K=21$ (2 ties), as opposed to 30 states for $K=9$ (9 ties).

A general pattern appeared to emerge in our examination of the effect of the number of strata upon the accuracy of the proposed strategy. From the data in Table 4.1, we find that gains in precision could be achieved with a controlled increase in the number of strata. State estimates of unemployment generated for $K=14$ were equivalent (14 ties) or superior to those for $K=6$ in 38 states. Similarly, estimates derived for $K=21$ were equally precise (11 ties) or more accurate in 35 states. Here, however, there appeared to be a value between $K=14$ and $K=21$ beyond which any subsequent increase would not be advantageous. As noted in Section 4.3, a major constraint to be encountered was the allocation of a limited number of sample observations to their respective strata, thereby reducing the level of precision of the sample estimates of the stratum means.

Perhaps the greatest limitation of the proposed method could be
### Table 4.1

Relative Absolute Deviation of the 1970 Unemployment Rate Estimates for States

<table>
<thead>
<tr>
<th>State</th>
<th>Number of Counties</th>
<th>Proposed Method (K=6)</th>
<th>Proposed Method (K=9)</th>
<th>Proposed Method (K=14)</th>
<th>Proposed Method (K=21)</th>
<th>Ericksen Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>67</td>
<td>0.16</td>
<td>0.13</td>
<td>0.16</td>
<td>0.16</td>
<td>0.22</td>
</tr>
<tr>
<td>Arizona</td>
<td>14</td>
<td>0.24</td>
<td>0.24</td>
<td>0.17</td>
<td>0.26</td>
<td>0.19</td>
</tr>
<tr>
<td>Arkansas</td>
<td>75</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>California</td>
<td>58</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.21</td>
</tr>
<tr>
<td>Colorado</td>
<td>63</td>
<td>0.12</td>
<td>0.10</td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>Connecticut</td>
<td>8</td>
<td>0.37</td>
<td>0.43</td>
<td>0.31</td>
<td>0.37</td>
<td>0.31</td>
</tr>
<tr>
<td>Delaware</td>
<td>3</td>
<td>0.29</td>
<td>0.32</td>
<td>0.26</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td>Florida</td>
<td>66</td>
<td>0.29</td>
<td>0.32</td>
<td>0.26</td>
<td>0.26</td>
<td>0.32</td>
</tr>
<tr>
<td>Georgia</td>
<td>159</td>
<td>0.63</td>
<td>0.56</td>
<td>0.63</td>
<td>0.56</td>
<td>0.75</td>
</tr>
<tr>
<td>Idaho</td>
<td>44</td>
<td>0.10</td>
<td>0.08</td>
<td>0.12</td>
<td>0.12</td>
<td>0.19</td>
</tr>
<tr>
<td>Illinois</td>
<td>102</td>
<td>0.27</td>
<td>0.24</td>
<td>0.27</td>
<td>0.27</td>
<td>0.30</td>
</tr>
<tr>
<td>Indiana</td>
<td>92</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.07</td>
</tr>
<tr>
<td>Iowa</td>
<td>99</td>
<td>0.29</td>
<td>0.29</td>
<td>0.14</td>
<td>0.17</td>
<td>0.14</td>
</tr>
<tr>
<td>Kansas</td>
<td>105</td>
<td>0.13</td>
<td>0.09</td>
<td>0.11</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>Kentucky</td>
<td>120</td>
<td>0.13</td>
<td>0.13</td>
<td>0.11</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>Louisiana</td>
<td>64</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
</tr>
<tr>
<td>Maine</td>
<td>16</td>
<td>0.17</td>
<td>0.19</td>
<td>0.14</td>
<td>0.17</td>
<td>0.10</td>
</tr>
<tr>
<td>Maryland</td>
<td>24</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.25</td>
<td>0.53</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>14</td>
<td>0.29</td>
<td>0.29</td>
<td>0.26</td>
<td>0.26</td>
<td>0.29</td>
</tr>
<tr>
<td>Michigan</td>
<td>83</td>
<td>0.20</td>
<td>0.22</td>
<td>0.19</td>
<td>0.20</td>
<td>0.22</td>
</tr>
<tr>
<td>Minnesota</td>
<td>87</td>
<td>0.20</td>
<td>0.12</td>
<td>0.10</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>Mississippi</td>
<td>82</td>
<td>0.16</td>
<td>0.20</td>
<td>0.12</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td>Missouri</td>
<td>115</td>
<td>0.12</td>
<td>0.12</td>
<td>0.14</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Montana</td>
<td>57</td>
<td>0.21</td>
<td>0.21</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>Nebraska</td>
<td>93</td>
<td>0.67</td>
<td>0.70</td>
<td>0.56</td>
<td>0.56</td>
<td>0.52</td>
</tr>
<tr>
<td>Nevada</td>
<td>17</td>
<td>0.09</td>
<td>0.07</td>
<td>0.09</td>
<td>0.06</td>
<td>0.15</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>10</td>
<td>0.34</td>
<td>0.37</td>
<td>0.34</td>
<td>0.40</td>
<td>0.23</td>
</tr>
<tr>
<td>New Jersey</td>
<td>21</td>
<td>0.21</td>
<td>0.24</td>
<td>0.16</td>
<td>0.16</td>
<td>0.24</td>
</tr>
<tr>
<td>New Mexico</td>
<td>32</td>
<td>0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>New York</td>
<td>62</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
<td>0.18</td>
<td>0.30</td>
</tr>
<tr>
<td>North Carolina</td>
<td>100</td>
<td>0.50</td>
<td>0.47</td>
<td>0.50</td>
<td>0.50</td>
<td>0.56</td>
</tr>
<tr>
<td>North Dakota</td>
<td>53</td>
<td>0.04</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.11</td>
</tr>
<tr>
<td>Ohio</td>
<td>88</td>
<td>0.18</td>
<td>0.18</td>
<td>0.15</td>
<td>0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>77</td>
<td>0.14</td>
<td>0.14</td>
<td>0.12</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td>Oregon</td>
<td>36</td>
<td>0.36</td>
<td>0.37</td>
<td>0.40</td>
<td>0.40</td>
<td>0.41</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>67</td>
<td>0.27</td>
<td>0.27</td>
<td>0.24</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>5</td>
<td>0.18</td>
<td>0.20</td>
<td>0.25</td>
<td>0.20</td>
<td>0.23</td>
</tr>
<tr>
<td>South Carolina</td>
<td>46</td>
<td>0.39</td>
<td>0.37</td>
<td>0.39</td>
<td>0.37</td>
<td>0.50</td>
</tr>
<tr>
<td>South Dakota</td>
<td>67</td>
<td>0.30</td>
<td>0.27</td>
<td>0.22</td>
<td>0.24</td>
<td>0.16</td>
</tr>
<tr>
<td>Tennessee</td>
<td>95</td>
<td>0.14</td>
<td>0.11</td>
<td>0.16</td>
<td>0.16</td>
<td>0.18</td>
</tr>
<tr>
<td>Texas</td>
<td>254</td>
<td>0.44</td>
<td>0.44</td>
<td>0.42</td>
<td>0.44</td>
<td>0.39</td>
</tr>
<tr>
<td>Utah</td>
<td>29</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>0.15</td>
</tr>
<tr>
<td>Vermont</td>
<td>14</td>
<td>0.15</td>
<td>0.20</td>
<td>0.15</td>
<td>0.17</td>
<td>0.07</td>
</tr>
<tr>
<td>Virginia</td>
<td>134</td>
<td>0.63</td>
<td>0.57</td>
<td>0.53</td>
<td>0.43</td>
<td>0.63</td>
</tr>
<tr>
<td>Washington</td>
<td>39</td>
<td>0.42</td>
<td>0.39</td>
<td>0.43</td>
<td>0.42</td>
<td>0.47</td>
</tr>
<tr>
<td>West Virginia</td>
<td>55</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>72</td>
<td>0.13</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>Wyoming</td>
<td>23</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Number of times the Proposed Strategy does as well as or better than Ericksen Method

<table>
<thead>
<tr>
<th></th>
<th>29 (4 Ties)</th>
<th>30 (9 Ties)</th>
<th>32 (5 Ties)</th>
<th>31 (2 Ties)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>60.4%</td>
<td>62.5%</td>
<td>66.7%</td>
<td>64.6%</td>
</tr>
</tbody>
</table>
observed when estimates were generated for those states whose 1970 unemployment experience most notably deviated from the national average. Due to the nature of their derivation, the estimates tend to cluster nearest the stratum means most representative of the underlying population. Consequently, the states which exhibit the more extreme values do not receive the most appropriate characterization. The data in Table 4.2 serve to illustrate this pattern. Here, the post-stratified estimates were most reliable for states with unemployment rates in the range \(3.9 \leq \bar{y}_k < 5.9\). As noted, the least accurate results occurred for the most extreme values: \(\bar{y}_k < 3.9\) and \(\bar{y}_k \geq 5.9\).
### Table 4.2
**Relative Absolute Deviation of the 1970 Unemployment Rate Estimates for States**

<table>
<thead>
<tr>
<th>Unemployment Rate</th>
<th>State</th>
<th>Number of Counties</th>
<th>Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>K=6</td>
<td>K=9</td>
</tr>
<tr>
<td>$\bar{y}_k &lt; 3.5$</td>
<td>Georgia</td>
<td>159</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>Maryland</td>
<td>24</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Nebraska</td>
<td>93</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>North Carolina</td>
<td>100</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Virginia</td>
<td>134</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.57</td>
</tr>
<tr>
<td>Mean R.A.D.</td>
<td></td>
<td></td>
<td>0.37</td>
</tr>
<tr>
<td>$3.5 \leq \bar{y}_k &lt; 3.9$</td>
<td>Connecticut</td>
<td>8</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Delaware</td>
<td>3</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Florida</td>
<td>67</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Illinois</td>
<td>102</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Iowa</td>
<td>99</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Massachusetts</td>
<td>14</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>New Hampshire</td>
<td>10</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>New Jersey</td>
<td>21</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Pennsylvania</td>
<td>67</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>South Carolina</td>
<td>46</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>South Dakota</td>
<td>67</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Texas</td>
<td>254</td>
<td>0.44</td>
</tr>
<tr>
<td>Mean R.A.D.</td>
<td></td>
<td></td>
<td>0.31</td>
</tr>
<tr>
<td>$3.9 \leq \bar{y}_k &lt; 4.4$</td>
<td>Arizona</td>
<td>14</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>Colorado</td>
<td>63</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Indiana</td>
<td>92</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Kansas</td>
<td>105</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>Maine</td>
<td>16</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>Minnesota</td>
<td>87</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Missouri</td>
<td>115</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>New York</td>
<td>62</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Ohio</td>
<td>88</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Oklahoma</td>
<td>77</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Rhode Island</td>
<td>5</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Vermont</td>
<td>14</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Wisconsin</td>
<td>72</td>
<td>0.13</td>
</tr>
<tr>
<td>Mean R.A.D.</td>
<td></td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>$4.4 \leq \bar{y}_k &lt; 4.9$</td>
<td>Alabama</td>
<td>67</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Kentucky</td>
<td>120</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>North Dakota</td>
<td>53</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Tennessee</td>
<td>95</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Wyoming</td>
<td>23</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean R.A.D.</td>
<td></td>
<td></td>
<td>0.09</td>
</tr>
<tr>
<td>$4.9 \leq \bar{y}_k &lt; 5.4$</td>
<td>Idaho</td>
<td>44</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Mississippi</td>
<td>82</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Utah</td>
<td>29</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>West Virginia</td>
<td>55</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean R.A.D.</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>$5.4 \leq \bar{y}_k &lt; 5.9$</td>
<td>Arkansas</td>
<td>75</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>Louisiana</td>
<td>64</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Nevada</td>
<td>17</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>New Mexico</td>
<td>32</td>
<td>0.02</td>
</tr>
<tr>
<td>Mean R.A.D.</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>Unemployment Rate</td>
<td>State</td>
<td>Number of Counties</td>
<td>Proposed Method</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------</td>
<td>--------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>( \bar{y} \geq 5.9 )</td>
<td>California</td>
<td>58</td>
<td>0.22 0.22 0.22 0.22</td>
</tr>
<tr>
<td></td>
<td>Michigan</td>
<td>83</td>
<td>0.20 0.22 0.19 0.20</td>
</tr>
<tr>
<td></td>
<td>Montana</td>
<td>57</td>
<td>0.21 0.21 0.23 0.23</td>
</tr>
<tr>
<td></td>
<td>Oregon</td>
<td>36</td>
<td>0.36 0.37 0.40 0.40</td>
</tr>
<tr>
<td></td>
<td>Washington</td>
<td>39</td>
<td>0.42 0.39 0.43 0.42</td>
</tr>
<tr>
<td>Mean R.A.D.</td>
<td></td>
<td></td>
<td>0.28 0.28 0.29 0.29</td>
</tr>
</tbody>
</table>
CHAPTER V
SUMMARY AND SUGGESTIONS FOR FURTHER RESEARCH

5.1. Introduction
In this final chapter, our first objective will be to summarize the direction our research has taken. This will be followed by a discussion and re-examination of some general conclusions we have reached. Finally, suggestions for future research will be given.

5.2. Summary
Chapter I is devoted to a thorough investigation of the present state of the "art" of local area estimation. The orientation of this presentation is both descriptive and evaluative in nature. Two major constraints are observed to resurface continually in several of the most "popular" strategies entertained for estimating the parameters of local areas. Perhaps the most common is their restrictive underlying assumptions. The other equally serious drawback is their inability to account properly for related local factors. Included in this chapter is a discussion of the work of Kalsbeek, which has served as the motivation of this dissertation.

A more formal representation of the Kalsbeek model is developed in Chapter II. Proceeding more rigorously along the lines of standard sample survey theory, we consider the method of small area estimation for a single stage cluster sample design and a two-stage sampling design
with equal probabilities of selection. Here, "groups" are redefined as strata, and their determination can also be achieved by methods more along the lines of minimum variance stratification (i.e., use of $\text{CUM}_F$ rule). The methodology involves ratio estimation of the respective stratum means via indicator variables which serve to classify the sample base units. In addition, a general analytic expression for the mean square error of the resultant estimators is derived.

In Chapter III, we have reformulated the model of Kalsbeek to allow for further theoretical investigations of the local (target) area estimator's behavior when more than one symptomatic (stratification) variable is considered. Our extension to the multivariate setting has enabled us to derive approximations for the expectation, variance, bias and mean squared error of the proposed target area estimator without having to resort to Monte Carlo simulations. Here, the relationship between criterion and symptomatic variables is to be specified by relevant continuous multivariate distributions. Our research focuses on both the trivariate normal and logistic settings. Because the theoretical framework considers multivariate stratification, each target area estimator is expressed in terms of a weighted linear combination of post-stratified estimators. We attempt to form strata which are optimal in the sense of minimum variance stratification. This is reflected in the choice of strategies considered to determine approximately optimum boundaries. Consequently, the strata we construct are rectilinearly defined, non-overlapping, and exhaustive. Comparisons of precision are then made with respect to the Ericksen model, which is applicable to the same general setting. We also examine the effect of the number of specified strata on the performance of the proposed target area estimator.
An application of the proposed strategy to real data is presented in Chapter IV. Using the Current Population Survey as our source of sample information, we generate state estimates of the 1970 unemployment rate. As before, comparisons of accuracy are made with respect to the Ericksen method. The choice of symptomatic information (taken from the 1972 County City Data Book) is conditioned by our intent to demonstrate the suitability of the post-stratification scheme when the relationship between criterion and symptomatic variables is more appropriately characterized by a non-linear model. In contrast to the strategy entertained by Kalsbeek which consists of group formation via clustering algorithms (i.e., A.I.D. II) using only sample data, our stratification scheme is more closely modelled after the approach taken in Chapter III. Here, the determination of stratum boundaries is achieved by application of the CUM/2 rule in a hierarchical framework which considers the symptomatic configuration of all base units in the population.

5.3. Conclusions

Reliable estimates of parameters at the local level are generally difficult, if not impossible, to obtain directly from sample surveys, primarily due to the constraints of sample size and design. Yet, the very nature of the problem has served as the motivating force in the development of several alternative procedures. When underlying assumptions are too strict or unrealistic, the need for a more flexible approach is obvious. The method considered in our research is particularly attractive in that no functional model between criterion and symptomatic variables must be specified. Here, the most limiting as-
sumption is the availability of good symptomatic information.

From the distribution specific results of our analytical work, we demonstrate the proposed method's comparative superiority over the Ericksen model when the non-linear setting of the trivariate logistic distribution is considered. The post-stratified estimator performs reasonably well for the linear setting of the trivariate normal distribution with a moderate association level ($R = .58$) when the sample size is large and a hierarchical scheme of stratum boundary determination is employed. As expected, the method of Ericksen is found to be most appropriate to the high association model ($R = .95$). In each case, our measure of precision is the mean squared error. Reductions in the variance and mean squared error terms are consistently achieved for both strategies with an increase in sample size. Gains in precision are also noted for the post-stratified estimator when the hierarchical stratification scheme is selected over the method of Anderson. Here, there is no conclusive evidence of any consistent relationship between the number of strata and the method's accuracy.

The suitability of the post-stratification scheme over the Ericksen method when the relationship between criterion and symptomatic variables is more appropriately characterized by a non-linear model is also illustrated in the unemployment example. Here, estimates which are equally precise or more accurate are obtained for 32 of the 48 states under consideration ($K=14$). An empirical examination of the results suggests that gains in precision can be achieved with a controlled increase in the number of strata. Finally, the proposed strategy does not perform as well for those states whose unemployment experience most notably deviate from the national average.
5.4. Suggestions for Future Research

We conclude by making the following suggestions for future research:

1. New developments in methodology are needed to determine those multivariate stratification boundaries which are optimal, or approximately so, in the sense of minimum variance stratification. Strategies which are well adapted to practical settings are particularly relevant to the method under study, as potential gains in precision may be achieved. Here, a reinvestigation of the effects of alternative procedures in cluster analysis may prove valuable.

2. In the theoretical work of the dissertation, we have made the implicit assumption of no errors of measurement. An adaptation of the proposed model which takes into account measurement and/or response errors could be useful. The subsequent effect upon the local area estimator's precision would need to be examined.

3. Another extension, if possible, would be the development of a new technique which combines the attributes of our local area estimator with those of the Ericksen model and NCHS synthetic estimator to yield an estimator superior to each individually. This might be achieved by some appropriate weighting scheme or iterative procedure.

Finally, we encourage the undertaking of additional empirical studies which further examine the behavior of the proposed local area estimator and allow for comparisons of precision with respect to other strategies.
APPENDIX A

EXPLICIT EXPRESSIONS FOR FIRST AND SECOND ORDER MOMENTS
OF THE DOUBLY TRUNCATED BIVARIATE NORMAL DISTRIBUTION

In Chapter III, section 6, we have assumed the underlying distribution of the random vector \((y, x_1, x_2)\) is trivariate normal. In order to determine the approximate expected value of the proposed target area estimator \(\hat{y}_{\hat{x}}\), we need to evaluate the respective conditional first order moments

\[
\mu_{x_i g} = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} x_i \frac{g(x)}{w_g} d x \quad i = 1, 2
\]  

(A.1)

where \(g(x) = g(x_1, x_2)\) is the bivariate normal density defined in Section 3.6 and

\[
w_g = \Pr\{a_{1g} < x_1 < b_{1g}, a_{2g} < x_2 < b_{2g}\} = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} g(x) d x
\]

(A.2)

For the singly truncated standardized bivariate normal distribution, Rosenbaum (1961) has shown that

\[
T_1(h_1, h_2) = \int_{h_1}^{\infty} \int_{h_2}^{\infty} x_1(s) g(s) (\bar{x}(s)) d \bar{x}(s)
\]

\[
= Z(h_1)(1 - Q\left(\frac{h_2 - \rho h_1}{\sqrt{1 - \rho^2}}\right)) + \rho Z(h_2)(1 - Q\left(\frac{h_1 - \rho h_2}{\sqrt{1 - \rho^2}}\right))
\]

(A.3)
and
\[ T_2(h_1, h_2) = \int_{h_1}^{h_2} \int_{h_1}^{h_2} x_1(s) g(s) (x(s)) d x(s) \]
\[ = \rho Z(h_1)(1 - Q\left( \frac{h_2 - \rho h_1}{\sqrt{1 - \rho^2}} \right)) + Z(h_2)(1 - Q\left( \frac{h_1 - \rho h_2}{\sqrt{1 - \rho^2}} \right)) \]  
(A.4)

where \( Z \) and \( Q \) are the standard frequency and distribution functions of the univariate normal distribution and

\[
g(s)(x_1(s), x_2(s)) = \begin{vmatrix}
\frac{\partial x_1}{\partial x_1(s)} & \frac{\partial x_1}{\partial x_2(s)} \\
\frac{\partial x_2}{\partial x_1(s)} & \frac{\partial x_2}{\partial x_2(s)}
\end{vmatrix} g(x_1, x_2)
\]

\[
= \frac{- (x_1(s) - \mu_{x_1})^2 - 2\rho x_1(s) x_2(s) + x_2(s)^2}{2(1 - \rho^2)}
\]
\[ \exp[\frac{- (x_1(s) - \mu_{x_1})^2}{2(1 - \rho^2)}] \]
\[ \frac{2\pi(1 - \rho^2)}{2\pi\sqrt{1 - \rho^2}} \]

for \( x_1(s) = \frac{x_1 - \mu_{x_1}}{G_{x_1}} \), \( i = 1, 2 \). Consequently,

\[
E(x_i(s) | a_{1g} < x_i(s) < b_{1g}, a_{2g} < x_i(s) < b_{2g})
\]
\[ = (W(s))^{-1}[T_i(b_{1g}, b_{2g}) - T_i(b_{1g}, a_{2g}) - T_i(a_{1g}, b_{2g}) + T_i(a_{1g}, a_{2g})] \]
\[ = \mu_{x_{1g}}(s) \]  
(A.6)
where \( a_{ig}^{(s)} = \frac{a_{ig}}{\sigma_{x_{i}}} \) and \( b_{ig}^{(s)} = \frac{b_{ig}}{\sigma_{x_{i}}} \) are the respective strata boundaries and

\[
W_{g}^{(s)} = \Pr\{a_{ig}^{(s)} < x_{i}^{(s)} < b_{ig}^{(s)}, a_{2g}^{(s)} < x_{2}^{(s)} < b_{2g}^{(s)}\}.
\]

To determine \( \mu_{x_{ig}} \), we consider the transformation

\[
x_{i} = \sigma_{x_{i}} x_{i}^{(s)} + \mu_{x_{i}} \quad \text{for } 1, 2,
\]

such that

\[
\mu_{x_{ig}} = \int_{a_{lg}}^{b_{lg}} \int_{a_{2g}}^{b_{2g}} \frac{x_{i} g(x) d x}{W_{g}}
= \int_{a_{lg}}^{b_{lg}} \int_{a_{2g}}^{b_{2g}} \frac{(\sigma_{x_{i}} x_{i}^{(s)} + \mu_{x_{i}}) g^{(s)}(x^{(s)}) d x^{(s)}}{W_{g}}
= \int_{a_{lg}}^{b_{lg}} \int_{a_{2g}}^{b_{2g}} \sigma_{x_{i}} x_{i}^{(s)} g^{(s)}(x^{(s)}) d x^{(s)}
\]

\[
+ \mu_{x_{i}} \int_{a_{lg}}^{b_{lg}} \int_{a_{2g}}^{b_{2g}} \frac{g^{(s)}(x^{(s)}) d x^{(s)}}{W_{g}}
= \sigma_{x_{i}} \mu_{x_{ig}}^{(s)} + \mu_{x_{i}}^{(s)}, \quad (A.7)
\]

since \( W_{g} = \Pr\{a_{lg} < x_{1} < b_{lg}, a_{2g} < x_{2} < b_{2g}\} \)

\[= \Pr\{a_{lg}^{(s)} < x_{1}^{(s)} < b_{lg}^{(s)}, a_{2g}^{(s)} < x_{2}^{(s)} < b_{2g}^{(s)}\}
= W_{g}^{(s)}.
\]
To determine the approximate variance of the proposed target area estimator \( \hat{Y}_{\lambda} \), we must evaluate the respective conditional second order and joint moments

\[
E(x_{1}^{2} | a_{1g} < x_{1} < b_{1g}, a_{2g} < x_{2} < b_{2g}) = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} x_{1}^{2} g(x) \frac{d \theta}{w_{g}}
\]

for \( i = 1, 2 \) \hspace{1cm} (A.8)

and

\[
E(x_{1}x_{2} | a_{1g} < x_{1} < b_{1g}, a_{2g} < x_{2} < b_{2g}) = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} x_{1}x_{2} g(x) \frac{d \theta}{w_{g}}
\]

\hspace{1cm} (A.9)

Rosenbaum has also shown that

\[
T_{11}(h_{1}, h_{2}) = \int_{h_{1}}^{\infty} \int_{h_{2}}^{\infty} x_{1}(s)^{2} g(s)(x(s)) \frac{d x(s)}{x(s)} = L(h_{1}, h_{2}) + h_{1} Z(h_{1}) \left(1 - Q\left(\frac{h_{2} - \rho h_{1}}{\sqrt{1 - \rho^{2}}}\right)\right)
\]

\[
+ \rho^{2} h_{2} Z(h_{2}) \left(1 - Q\left(\frac{h_{1} - \rho h_{2}}{\sqrt{1 - \rho^{2}}}\right)\right) + \rho \sqrt{(1 - \rho^{2}) / (\pi \sqrt{2\pi})} Z\left(\frac{\sqrt{h_{1} - 2\rho h_{1}h_{2} + h_{2}^{2}}}{\sqrt{1 - \rho^{2}}}\right)
\]

\hspace{1cm} (A.10)

where

\[
L(h_{1}, h_{2}) = \int_{h_{1}}^{\infty} \int_{h_{2}}^{\infty} g(s)(x(s)) \frac{d x(s)}{x(s)}
\]
\[ T_{22}(h_1, h_2) = \int_{h_1}^{\infty} \int_{h_2}^{\infty} x_2(s)^2 g(s) x_1(s) d x_1(s) \]

\[ = L(h_1, h_2) + \rho h_1 Z(h_1) Q\left( \frac{h_2 - \rho h_1}{\sqrt{1 - \rho^2}} \right) \]

\[ + h_2 Z(h_2) (1 - Q\left( \frac{h_2 - \rho h_1}{\sqrt{1 - \rho^2}} \right)) + \frac{\rho \sqrt{(1 - \rho^2)}}{\sqrt{2\pi}} Z\left( \frac{\sqrt{h_1^2 - 2\rho h_1 h_2 + h_2^2}}{\sqrt{1 - \rho^2}} \right) \]

and

\[ T_{21}(h_1, h_2) = T_{12}(h_1, h_2) = \int_{h_1}^{\infty} \int_{h_2}^{\infty} x_1(s) x_2(s)^2 g(s) x_1(s) d x_1(s) \]

\[ = \rho L(h_1, h_2) + \rho h_1 Z(h_1) Q\left( \frac{h_2 - \rho h_1}{\sqrt{1 - \rho^2}} \right) \]

\[ + \rho h_2 Z(h_2) (1 - Q\left( \frac{h_1 - \rho h_2}{\sqrt{1 - \rho^2}} \right)) + \frac{\sqrt{(1 - \rho^2)}}{\sqrt{2\pi}} Z\left( \frac{\sqrt{h_1^2 - 2\rho h_1 h_2 + h_2^2}}{\sqrt{1 - \rho^2}} \right) \]

\[ (A.12) \]

Consequently, \( E(x_1(s)x_j(s) | a_1(s) < x_1(s) < b_1(s), a_2(s) < x_2(s) < b_2(s)) \)

\[ = (W_g^{(s)})^{-1} [T_{ij}(b_1, b_2) - T_{ij}(b_1, a_2) - T_{ij}(a_1, b_2) + T_{ij}(a_1, a_2)] \]

\[ \text{for } i = 1, 2 \text{ and } j = 1, 2. \]  

(A.13)

We can use these results directly to determine the respective elements of \( \Sigma_{x_g} \).
\[ \sigma_{1jg} = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{(x_i - \mu_{x_{ig}})(x_j - \mu_{x_{jg}})}{W_g} g(x) \, d\,x \]

\[ = \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{\sigma_i \sigma_j}{W_g} (x_i(s) - \mu_i(s))(x_j(s) - \mu_j(s)) g(s)(\xi(s)) \, d\,\xi(s) \]

\[ = \sigma_i \sigma_j \int_{a_{1g}}^{b_{1g}} \int_{a_{2g}}^{b_{2g}} \frac{x_i(s) x_j(s)}{W_g} g(s)(\xi(s)) \, d\,\xi(s) - \mu_i(s) \mu_j(s) \]

for \( i = 1, 2 \) and \( j = 1, 2 \). \hspace{1cm} (A.14)

We also note that \( W_g = W_g(s) \)

\[ = L(b_{1g}, b_{2g}) - L(b_{1g}, a_{2g}) - L(a_{1g}, b_{2g}) + L(a_{1g}, a_{2g}) \]

(A.15)

and that Curnow and Dunnett have shown \( L(h_1, h_2) \) can be reduced to a single integration

\[ L(h_1, h_2) = \int_{-\infty}^{\infty} \prod_{i=1}^{2} Q\left\{ (h_i - \alpha_i y)/(1 - \alpha^2) \right\}^{1/2} Z(y) \, d\,y \] \hspace{1cm} (A.16)

where \( \alpha_1 = \alpha_2 = \sqrt{\rho} \) for \( \rho \geq 0 \) and \( \alpha_1 = -\alpha_2 = -\sqrt{|\rho|} \) for \( \rho < 0 \).

Anderson (1976) has found that application of the trapezoidal rule over the interval \((-10, 10)\) on \( y \) (where the width of each trapezoid is 1/10) yields a numerical approximation of \( L(h_1, h_2) \) accurate to five decimal places. This is the approach we have adopted to approx-
imate numerically the respective bivariate probabilities,

\[ w_g = w_g(s) = \Pr\{a_{1g} < x_1(s) < b_{1g}, a_{2g} < x_2(s) < b_{2g}\} \]
REFERENCES


