University of Alberta

Small Area Rates:

Methods of Estimation and Inference

By

Eshetu Tesfaye Wondmagegnehu

A thesis

submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirement for the degree of a Master of Science in Statistics.

Department of Mathematical Sciences

Edmonton, Alberta, Canada.

Fall 1998.

0-612-34435-5

University of Alberta

Library Release Form

Name of Author: *Eshetu Tesfaye Wondmagegnehu*

Title of Thesis: *Small Area Rates: Methods of Estimation and Inference.*

Degree: *Master of Science.*

Year this Degree Granted: *1998.*

Eshetu Tesfaye Wondmagegnehu
Addis Ababa Commercial College
P.O.Box 3131
Addis Ababa
ETHIOPIA.

Date: August 07, 1998

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled *Small Area Rates: Methods of Estimation and Inference* submitted by *Eshetu Tesfaye Wondmagegnehu* in partial fulfillment of the requirements for the degree of *Master of Science* in *Statistics*.

K C Carrière

Dr. K.C. Carrière (Supervisor)
Department of Mathematical Sciences
University of Alberta.

Dr. N.G.N. Prasad (Committee Chair & Examiner)
Department of Mathematical Sciences
University of Alberta.

Stephen Newman

Dr. Stephen Newman (Examiner)
Department of Public Health Sciences
University of Alberta.

Date: August 6, 1998

Abstract

There are various methods for measuring the extent of poverty and needs across geographical areas. Some of these measures are simple and applicable to any data, while others are based on rigorous theoretical foundations where their robustness depends on the underlying assumptions. Some require small aggregate data measured at an area level, while others require extensive data measured at an individual level which are often difficult to obtain in practice. First, we review and discuss several methods to construct poverty and need indices across several geographical areas. Then we discuss their limitations and develop strategies that can accommodate some of these limitations to improve the current state of estimation methods of small areas. Specifically we suggest data reduction methods for multivariate and correlated data prevalent in most empirical situations. We also consider parametric and non-parametric estimation to obtain these indices including small area estimation techniques. Simulation study on some selected procedures will validate these estimation and making inferences across small areas using these indices. The discrepancies or similarities of the empirical interpretations of employing various approaches will be demonstrated using data from the province of Manitoba to study variation in needs and poverty across small areas.

To

My Wife, **Eyerusalem Yematawork**

and

My Parents.

Acknowledgement

I am very grateful to my supervisor, Dr. K.C. Carriere, for leading me to the research field for this thesis, and for her many helpful and thoughtful comments, discussions and suggestions throughout the preparation of this thesis. I would also like to thank her for her support in obtaining a scholarship from the Alberta Heritage Foundation for Medical Research and for her patience and kindness in spending her time on this work, which has made my study at this university very rewarding and fulfilling. Without her guidance, this thesis would have been impossible.

Thanks also go to the Department of Mathematical Sciences, University of Alberta, for providing me the opportunity to study here and the financial support. I would like to express my appreciation to the faculty members of the department who provided us with many interesting courses in statistics.

Special thanks go to Dr. N.G.N. Prasad for his kind support in providing materials and inspiring me with ideas during the preparation of this thesis and to Dr. Stephen Newman for reading my thesis and for his valuable suggestions.

Contents

Table of Notations

g (G)	- Gini's coefficient from the observed (random) sample.
g_{poor} (G_{poor})	- Gini's coefficient from the observed (random) sample of households whose income is below the poverty line.
g_{dist} (G_{dist})	- Gini's distance from the observed (random) sample of households.
n	- Sample size of the entire area under consideration.
N	- Population size of the entire area under consideration.
T	- Total number of small areas contained in a given large area.
n_i	- Sample size of ith small area, where $i=1, \dots, T$.
N_i	- Population size of the ith small area, $i=1, \dots, T$.
s_i	- Set of samples in area i, $i=1, \dots, T$.
s	- The entire sample.
PG	- Poverty gap.
Z	- Poverty line.
q	- Number of households whose income below the poverty line from the sample.
H	- Head count ratio.
P	- Sen's index of poverty measure.
P_α	- Foster, Greer and Thorbecke (FGT) measures of poverty.
$\bar{x}$ $(\bar{X})$	- Mean of an auxiliary variable X from the observed (random) sample.
$\bar{y}$ $(\bar{Y})$	- Mean of a characteristic Y from the observed (random) sample.
$\bar{x}_i$ $(\bar{X}_i)$	- Mean of an auxiliary variable X from the observed (random) sample in the ith small area.
$\bar{y}_i$ $(\bar{Y}_i)$	- Mean of a characteristic Y from the observed (random) sample in the ith small area.
$\bar{x}_h$ $(\bar{X}_h)$	- Mean of the auxiliary variable X from the observed (random) sample in the hth stratum.

$\bar{y}_h$ ($\bar{Y}_h$) - Mean for a variable of interest Y from the observed (random) sample in the hth stratum.

x_{ij} - Observed vector of k auxiliary variables from unit j in area i.

y - Observed vector of a characteristic Y.

X - n x k matrix of the auxiliary variables.

M - Population mean of the variable X.

μ - Population mean for the variable Y.

M_i - Population mean of the auxiliary variable X in the ith small area.

μ_i - Population mean for the variable of interest Y in the ith small area.

M_{hi} - Population mean of the auxiliary variable X in the hth stratum and ith small area.

μ_{hi} - Population mean for the variable of interest Y in the hth stratum and ith small area.

M_h - Population mean of the auxiliary variable X in the hth stratum.

μ_h - Population mean for the variable of interest Y in the hth stratum.

$\hat{\vartheta}_J$ - Jackknife estimator for a parameter ϑ.

$\hat{\vartheta}_b$ - Bootstrap estimator for a parameter ϑ.

$\hat{\vartheta}^{(-i)}$ - Estimator of ϑ after the ith sample point is deleted.

$\hat{\vartheta}_{(i)}$ - ith jackknife pseudo-value.

$\hat{\vartheta}_L$ - Lower confidence limit of a parameter ϑ.

$\hat{\vartheta}_U$ - Upper confidence limit of a parameter ϑ.

RB - Per cent gain in relative bias.

RHA - Regional Health Authority.

MLM - Multiple Linear Model.

VCM - Variance components model,

EL - Model based on Ecological level data.

PC - Principal component.

FA - Factor analysis.

Chapter 1

Introduction

Estimating poverty and need levels has become a major issue throughout the world as planners base policy initiatives on need. Various methodological techniques of measuring the extent of poverty and needs for "small areas" are reviewed and discussed in this thesis. This chapter will give background and definitions of the terms used throughout this paper.

1.1 Background

By all conventional measures, a large number of developing countries fared poorly in their economic and social conditions (Wondmagegnehu, 1995). The nature of poverty, needs and other related factors have not received due consideration in understanding the reasons for the variation in poverty and needs level. Most attempts made to alleviate needs and poverty have been treating in symptoms rather than removing the causes.

We define "needs" to refer to needs of any kind broadly applicable to education, income, health care, pharma care, societal care and others. Also, we define "small area" to refer to any area contained within a larger jurisdiction. For example, in the study of small area variation across the areas of Manitoba, a small area can be defined as each of the 12 regional health authorities (RHA). The issue may be to make resource allocations based on needs varying across the RHAs. The methodological problem is how to measure "needs"?

The first step is to measure the small area rates that will be used as indicators of needs. Then we consider which rate is appropriate and how it should be measured to best evaluate them. Estimating poverty or individual needs levels has become a major issue throughout the world because the planning of policy initiatives is ideally based on needs. There are various methodological techniques for measuring the extent of poverty and needs; some are simple to apply to any data while others build on rigorous theoretical foundations that depend on the type of data available.

1.2 Objectives

Wondmagegnehu (1995) reviewed several techniques to evaluate the level of poverty and individual needs across the geographical areas of Ethiopia. However, most of these techniques are simplistic and limited in the sense that they cannot deal with practical situations where there are many factors influencing needs, all of which are inter-correlated. The causes of poverty are also the causes of inequality in the distribution of income. The poor are poor because of their limited amount of human capital characterised by malnutrition, disease and illiteracy, impairing their capacity to earn a sufficient income. It is known that poverty and needs are defined in terms of many broad indicators of economic resources and basic needs required to survive, which include education, health care, income level and others (Wondmagegnehu and Carriere, 1997). For example, it is known that income level is closely related to education level. Taking only one of these factors is insufficient in attempting to summarise the areas' overall well-being. This thesis will review and discuss commonly used techniques of measuring the needs. We will then expand these to accommodate multivariate and inter-correlated data and we consider estimating small area rates while incorporating various levels of education, employment, income and resource availability.

1.3 Definition

1.3.1 Poverty and Needs

There is no clear-cut definition of poverty or needs and defining it is not an easy task. There is always some disagreement over what factors constitute poverty, poor health or poor living conditions. The poverty level measured by income, as well as the minimum basic needs required, differs from one country to another or from one place to another even within the same country.

Poverty can be specifically defined as a situational syndrome in which the following are combined: malnutrition, lack of resources, precarious living conditions, unemployment and others. Needs on the other hand have been defined in a more general sense as a value criterion of minimum adequate levels of welfare to keep the absolutely essential basic requirements of a household. Such criteria consequently imply a reference to some norm of basic needs and their satisfaction that makes it possible to distinguish between those who are poor and those who are not.

There are two different approaches to defining poverty and needs: absolute and relative measures. If poverty is defined in terms of the amount of household income and in terms of expenditure on the minimum essentials of food, clothes and other basic needs, such an approach to defining poverty is called an absolute measure. A household, under this definition, is then said to be poor if its total consumption is below a specified amount. Poverty is most often defined in relative terms when one's level of poverty depends upon the income of others in the community. If the income of a certain family is 10 percent lower than the average income of a given community, the family may consider itself poor even though their income level is above the poverty line. Therefore, in a relative sense of defining poverty, one might categorize a household as poor if the household's income is below a certain percent of average household income.

If an absolute standard is considered, rising real living standards will push more and more families above the poverty level. If, however, the relative standard is used, the war on poverty would be unwinable. If both rich and poor families receive equal percentage increases in income, the poor will not have improved their relative position. Therefore, poverty and inequality can be eliminated only by equalizing the distribution of household income. It is the relative measure of needs that we seek to estimate and make inferences about in this thesis.

1.3.1 Small Area

In recent years, the demand for small area statistics has greatly increased. This is due, among other things, to the growing use of statistical data in formulating policies and programs, to allocations of government funds and other services to small areas based on data obtained from large areas. These have increasingly created a need for small area statistics in recent decades. As a result, the construction and inference of reliable small area statistics has emerged in recent years as a major new research area. The term small area and local area are commonly used to denote a geographically contained area, for example, a county, a municipality, a census division or any subgroup sharing common characteristics. However, this term is broadly used to describe a small domain or small sub-population such as those defined by income quintiles, education level or a specific age-sex-race combination within a large geographical area (Ghosh and Rao, 1994).

1.4 Thesis Overview

Chapter 2 reviews the literature that determines inequality and poverty measures based on a single characteristic of interest for small areas. Small area estimation of a parameter using design and model based approaches is also discussed. In chapter 3, we discuss variance estimation and inference techniques of the measures summarised in chapter 2. Three different approaches to variance estimation are considered. These include the standard non-parametric, jackknife and bootstrap methods. Chapter 4 validates the inference procedures of the methods discussed in earlier chapters through a simulation study. Socio-economic data from the province of Manitoba will be used to compare these various approaches to estimation of small area rates.

Chapter 2

Review of Literature

There are various ways to measure poverty and needs indices. We first review currently available measures for small area rates to investigate the extent of their poverty and income inequality. Then, we review small area estimation methods for small area rates from design and model based approaches.

2.1 Methods of Estimation for Small Area Rates

In this section we review methods of estimating the extent of poverty and needs. These are mainly based on non-parametric techniques, which are simple and intuitive.

2.1.1 Gini's coefficient

Gini's coefficient is a very convenient summary measure of the relative degree of inequality of a characteristic such as income in an area. For the characteristic X in the random sample of size n, Gini's coefficient can be determined from the formula given by:

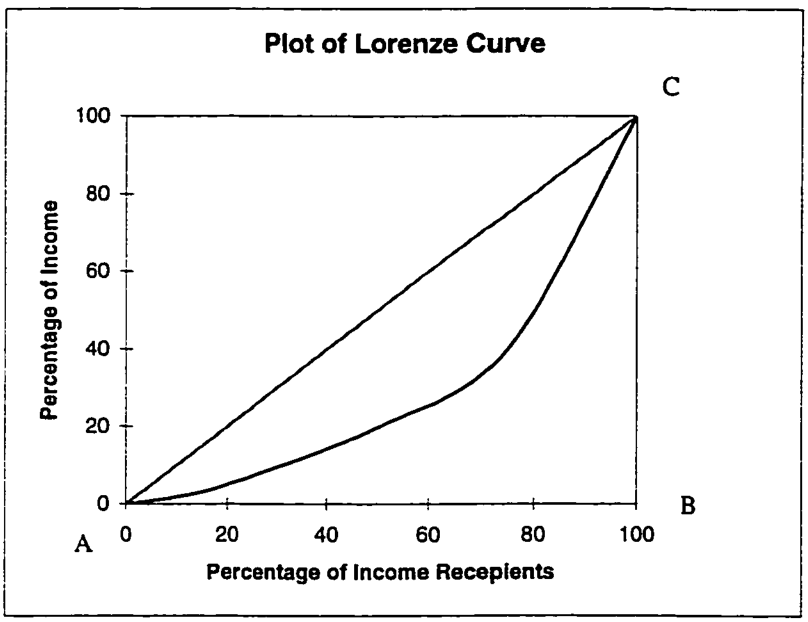
$$G = \frac{1}{n(n-1)\overline{X}} \sum_{i=1}^{n} \sum_{j>i}^{n} \left| X_i - X_j \right| \tag{2.1}$$

where $\overline{X} = \sum_{i=1}^{n} X_i \Big/ n$ is the mean of the random sample.

This coefficient measures the degree of variation of a characteristic of interest, such as income, which is perceived as the influencing factor of need. The Gini's coefficients are aggregate inequality measures and can vary anywhere from 0 (perfect equality) to 1 (gross inequality). In fact, the Gini's coefficient for areas with highly unequal income distributions typically lies between 0.50 and 0.70, while for areas with relatively equitable distributions it is on the order of 0.20 to 0.35 (Todaro, 1990).

Another method of computing Gini's coefficient is using the Lorenze curve. The Lorenze curve is a graphical method used to analyse the distribution of income for a group of individuals. It shows the actual quantitative relationship between the percentage of income recipients and the percentage of total income they received during a given year. In Figure 1, the bottom 10% of the population receive only 1.8% of the total income, the bottom 20% is receiving 5% of the total income and middle 50% is receiving 19.8% of the total income and so on.

Figure 1: Lorenze curve of per cent income by per cent population



The greater the degree of inequality, the more "bend" and the closer to the bottom horizontal axis the Lorenze curve will be (Todaro, 1990). The extreme case of inequality would be represented by the Lorenze curve bounded by the bottom horizontal and the right hand vertical axis. Since most countries couldn't exhibit perfect equality or extreme inequality of income distribution, the Lorenze curve will lie below the diagonal of the square figure.

Gini's coefficient, in this case, is obtained by calculating the ratio of the area between the diagonal and the Lorenze curve by the total area of the bottom triangle in which the curve lies. In Figure 1, Gini's coefficient is the ratio of an area between the line and curve to the total area of the triangle formed by connecting the points A, B and C.

$$G = \frac{Area \;\; between \;\; the \;\; line \;\; and \;\; curve}{Area \;\; of \;\; triangle \;\; ABC}$$

One major drawback of Gini's coefficient is that we cannot use it to measure inequality if the characteristic of interest has mean near zero. It is primarily constructed for positive values such as income. Addition or subtraction of a constant number to all values to make them positive would affect this measure of inequality (Dagum, 1983).

Alternately, Gini's distance can overcome this drawback of the Gini coefficient. For the X_i's , we described earlier, Gini's distance is given by

$$G_{dist} = \sum_{i=1}^{n} \sum_{j>i}^{n} \left| X_i - X_j \right| \Big/ \binom{n}{2} \qquad (2.2)$$

It measures the average absolute distance between any two data points. Gini's distance no longer has the property of being contained within a (0,1) boundary. However, it has the property of being more broadly applicable to all data types.

2.1.2 Poverty Line

Poverty lines are those cut-off lines in the economic welfare dimension of populations, across any geographical region, below which households or individuals are considered poor. Therefore, these are used to identify the poor using absolute terms. Income and total expenditure on consumables may be considered as indicators that combine different dimensions of the standard of living and are used to draw a poverty line. However, poverty lines do not provide any information on how far below the line most poor people are. It is important to know about the distribution of the poor who fall below the line to have a full understanding of poverty for various strategic purposes. Therefore, drawing poverty lines in terms of income or total expenditure implies simply the setting of norms for the minimum quantum of resources required.

Using poverty lines computed at the national level is not recommended when trying to distinguish poor households in different small areas or domains. It is necessary to draw sets of lines for different groups or areas, when the standards are not directly comparable from area to area. The most common and simplest method of obtaining a low income cut off for a population in a given area is by computing half of the median family income (Shao and Rao, 1993).

2.1.3 Sen's Index

When a needs index is designed to measure the extent and severity of poverty in any geographical area, the following two problems must be considered:

i. Identifying the poor in the total population, and
ii. Constructing an index of poverty using the available information on the poor.

The first problem involves the choice of a criterion for selecting or setting a "poverty line" in terms of the level of income, and then selecting those who fall below the poverty line. In the literature on poverty and needs measures, many are concerned with (i), but relatively little work has been done to resolve (ii) with which Sen (1976) was concerned.

Thus, most widely used measures of poverty to satisfy (ii) are developed to have the following properties:

1. **(Monotonocity)** Other things remaining the same, a reduction in income of a person below a poverty line must increase the poverty measures,

2. **(Transfer)** Other things remaining the same, a pure transfer of income from a person below the poverty line to any one who is above the poverty line must increase the poverty measure.

The first and perhaps the most simple measure of poverty is the head-count ratio (H). It simply takes the ratio of the poor, q, who are defined in a certain way and the total sample in a community of size n. While the head count ratio identifies the number of poor, it ignores how poor the poor really are, and therefore has the absurd property that it remains unchanged when a previously poor individual becomes poorer. It has been noted in the literature on poverty profiles that this measure is insensitive to decreases in the income of a household below the poverty line, i.e. to the deepening of poverty, and to the transfer of income among the poor, as well as to the transfers from the poor to the non-poor.

The head-count ratio is a crude measure and has as its main drawback that it does not satisfy the properties of monotonocity and transfer. Similar concerns apply to study persons hospitalized or persons discharged to investigate, for example, quality of care between hospitals.

The other common measure of poverty is called the "Poverty-Gap". It is defined as the aggregate short-fall of the income of those whose income level is below the poverty line. This measure gives a description of the depth of poverty, because it depends on the distance of the poor below the poverty line. The Poverty-Gap ratio is mathematically defined as

$$PG = \frac{1}{n}\sum_{i=1}^{q}\left(\frac{Z - X_i}{Z}\right)$$
$$= \frac{q}{n} - \sum_{i=1}^{q} X_i / nZ \tag{2.3}$$

where Z represents the poverty line, X_i total annual income of a household and n the total number of persons or households considered and q the number of persons or households whose income level is below the poverty line.

The main motivation of Sen (1976) in developing a new poverty measure is by recognising the violation of the monotonocity and transfer conditions in the poverty measures discussed above. For a large q, Sen derived a new poverty measure which is called Sen's Index given by

$$P = H\left[M + (1 - M)G_{poor}\right] \tag{2.4}$$

where G_{poor} is the Gini's coefficient of the income distribution of the poor, H is the head-count ratio and M is the income-gap ratio which tells us the percentage of their mean short-fall from the poverty line and is given by $M = (1/q)\sum_{i=1}^{q}(Z - X_i)/Z = \frac{n}{q}PG$.

2.1. 4 Foster-Greer-Thorbecke (FGT) Measures of Poverty

This measure of poverty is proposed by Foster, Greer and Thorbecke (1984). The FGT class of poverty measures include both, the head-count ratio and the poverty-gap index. This measure is sensitive to the distribution of the poor through the choice of a non negative parameter α, $0 \leq \alpha \leq 2$; the greater the weight of α given to the index the greater the sensitivity of the distribution of the poor. If the total population size is given by n, and q is the number of poor individuals, then the FGT class poverty measures may be written as:

$$P_{\alpha} = \frac{1}{n} \sum_{i=1}^{q} \left(1 - \frac{X_i}{Z} \right)^{\alpha} \quad . \tag{2.5}$$

There are three cases of the FGT measure considered here:

1. When $\alpha = 0$, the FGT measure is reduced to the head count ratio H = q/n given as the proportion of the population whose standard of living is below the poverty line.

2. When $\alpha = 1$, the FGT measure is reduced to the poverty-gap ratio (PG) in the population, expressed as a proportion of the poverty line.

3. When $\alpha = 2$, unlike the above two cases the measure is sensitive to the distribution of income among the poor. It also satisfies the main axioms of a desirable poverty measure in Sen (1976), monotonocity and transfer.

2.2 Small Area Estimation Methods for Small Area Rates

In this section we discuss and summarize techniques for estimating small area rates. Two different small area estimation approaches for small area means are considered: the survey mean estimator and the model based approach.

2.2.1 Survey Mean Estimator for Small Areas

Consider a survey where a characteristic is observed from a sample in a small area. Survey estimates for data from a small area of interest contained in a large domain is one method of estimating small area parameters. When using small sample from a large domain, the standard method of survey estimation breaks down because the small sample size yields very large standard error. Even though sample size in the small area is large enough, the coverage of the total population of the given area is inadequate to apply the survey estimates (Johnson, 1993). Rao (1986) and Ghosh and Rao (1994) proposed an estimation strategy by " borrowing strength" from other related small areas to increase the sample size and proper coverage for the purpose of improving the efficiency of the estimator.

Suppose N_i is the population size of the ith small area and μ_i is the parameter of interest, and the characteristic of interest is measured in Y_{ij} from person j in area i. The usual survey estimates, assuming simple random sampling , are given by

$$\overline{Y_i} = \frac{1}{n_i} \sum_{j \in s_i} Y_{ij} \tag{2.6}$$

where s_i is the set of samples from small area i and n_i is the size of s_i. The variance of $\overline{Y_i}$ is of order n_i^{-1} which gives a large standard error for small sample size. The simplest synthetic estimator of μ_i proposed by Rao (1986) with variance of order n^{-1} $\left(\sum n_i = n\right)$ is

$$\overline{Y_i} = \frac{1}{n} \sum_{j \in s} Y_{ij} \tag{2.7}$$

where s is the entire sample. This estimator becomes an efficient and unbiased estimator if all of the small areas are similar. If an auxiliary or concomitant data, $X = (X_{i1}, X_{i2}, ..., X_{in_i})$, is available with known mean M_i of X, the best synthetic estimator, if the assumption $R_i = \mu_i / M_i \cong R = \mu / M$ satisfied, is the ratio estimator given by

$$\overline{Y}_{i(ratio)} = \left(\frac{\overline{Y}}{\overline{X}} \right) M_i \tag{2.8}$$

where $\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in s_i} Y_{ij}$ and $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in s_i} X_{ij}$ are the means from the total combined sample, and n is the size of the combined sample.

The two estimators given in (2.7) and (2.8) are obtained by assuming homogeneity within small areas. This assumption can be relaxed by considering homogeneity within post-strata, i.e. $\mu_{hi} = \mu_h$ for $h = 1, 2, ..., L$ and $i = 1, 2, ..., T$. Let Y_{hij} be the observation from person j in stratum h and area i, the usual post-stratified estimator can be written as

$$\overline{Y}_i = \sum_{h=1}^{L} w_{hi} \overline{Y}_h \quad \text{where} \quad w_{hi} = \frac{n_{hi}}{N_{hi}} \tag{2.9}$$

where $\overline{Y}_h = \sum_{i=1}^{T} \sum_{j \in s_i} Y_{hij} \Big/ \sum_{i=1}^{T} n_{hi}$ and n_{hi} is the sample size of area i in stratum h. An alternative approach to synthetic estimators for small area means proposed by Ghosh and Rao (1994) is given by

$$\overline{Y}_i^{syn} = \sum_h \left(\frac{M_{hi}}{M_h} \right) \hat{\overline{Y}}_h' \tag{2.10}$$

where M_{hi} is the mean of the auxiliary information in the hth stratum and ith small area, M_h the mean of the auxiliary information in the hth stratum, and $\hat{\overline{Y}}_h'$ is a direct ratio estimate of the mean for stratum h in the population given by

$$\hat{\overline{Y}}_h' = \left(\frac{\overline{Y}_h}{\overline{X}_h} \right) M_h \tag{2.11}$$

where $\overline{Y}_h$ and $\overline{X}_h$ are the usual design estimators of small area means and M_h is the mean of the auxiliary information in stratum h. A composite estimator obtained from design and synthetic estimators can be considered as an alternative small area means estimator. This estimator is written as

$$\overline{Y}_i^{comp} = w_i \overline{Y}_i^{des} + (1 - w_i) \overline{Y}_i^{syn} \quad \text{where } 0 \le w_i \le 1. \tag{2.12}$$

Several methods have been proposed to estimate w_i. The optimal weight suggested by Ghosh and Rao (1994) assuming $\text{cov}\left(\overline{Y}_i^{des}, \overline{Y}_i^{syn}\right) \cong 0$ is

$$w_i(opt) = \frac{MSE(\overline{Y}_i^{syn})}{MSE(\overline{Y}_i^{syn}) + MSE(\overline{Y}_i^{des})} . \tag{2.13}$$

The small area mean estimators discussed above are designed to increase the efficiency of the estimate by merging samples from similar small areas to increase the sample size and the coverage of the sample.

2.2.2 Model Based Approach to Estimation of Small Area Rates

Consider a general mixed effects model for $\mathbf{y}=(\mathbf{y}_1',\mathbf{y}_2',\dots,\mathbf{y}_I')'$ with $\mathbf{y}_i=(Y_{i1},\dots,Y_{in_i})'$ given by

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{v}+\boldsymbol{\varepsilon} \tag{2.14}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices linking to a k vector of unknown parameters for fixed effects $\boldsymbol{\beta}$ and a q vector parameters of random effects $\boldsymbol{v}$. The $\boldsymbol{v}$ and ε are mutually independent random vectors with zero means and covariance matrices $\mathbf{G}$ and $\mathbf{R}$ respectively. A general BLUP method shown by Henderson (1975) for the overall mean of the form

$$\mu=\boldsymbol{k}'\boldsymbol{\beta}+\boldsymbol{m}'\boldsymbol{v} \tag{2.15}$$

is obtained by $\hat{\mu}=\boldsymbol{k}'\tilde{\boldsymbol{\beta}}+\boldsymbol{m}'\mathbf{G}\mathbf{Z}'\mathbf{D}^{-1}(\mathbf{y}-\boldsymbol{x}'\tilde{\boldsymbol{\beta}})$, where $\mathbf{D}=\mathbf{R}+\mathbf{Z}\mathbf{G}\mathbf{Z}'$ represents the covariance matrix of $\mathbf{y}$, $\tilde{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{D}^{-1}\mathbf{y})$, and $\boldsymbol{k}$ and $\boldsymbol{m}$ are known vectors. Prasad and Rao (1990), based on Henderson's method, discussed a small area BLUP of (2.15) by considering three special cases of a general mixed effects model. The three special cases discussed are the variance component model (or nested error regression model), the random regression coefficients model and the model based on ecological level data.

2.2.2.1. Variance Components Model

This model is originally proposed by Battese and Fuller (1981) for the purpose of predicting mean corn per hectare for T=12 counties (small areas), and is given by

$$Y_{ij} = x'_{ij}\beta + v_i + e_{ij} \;;\; i=1, 2, \ldots, T \text{ and } j=1, 2, \ldots, n_i \tag{2.16}$$

where Y_{ij} is the characteristic of interest for the jth sampled unit in the ith small area, $x'_{ij}=(X_{ij1}, X_{ij2}, \ldots, X_{ijk})$ is a vector corresponding to k auxiliary variables, $\beta=(\beta_1, \ldots, \beta_k)'$ is a vector of k unknown parameters and n_i is the number of sampled units observed in the ith small area. The random error v_i's are assumed to be iid $N(0,\sigma_v^2)$ and independent of e_{ij}'s which are assumed to be iid $N(0,\sigma_e^2)$. The main interest here is to find an estimate for the ith small area mean given by

$$\mu_i = M'_i\beta + v_i \qquad i= 1, 2, \ldots, T \tag{2.17}$$

assuming that $M'_i=(M_{i1}, \ldots, M_{ik})$ is known. The BLUP of (2.17) derived by Prasad and Rao (1990) is given by

$$\hat{\mu}_i = M'_i\tilde{\beta} + \gamma_i\left(\overline{Y}_i - \overline{X}'_i\tilde{\beta}\right) \tag{2.18}$$

where $\overline{X}'_i = (\overline{X}_{i1}, \overline{X}_{i2}, \ldots, \overline{X}_{ik})$ represents the sample mean of X_{ij}'s for the ith small area, with $\sigma^2 = (\sigma_v^2, \sigma_e^2)$ and $\gamma_i = \sigma_v^2\left(\sigma_v^2 + \sigma_e^2 n_i^{-1}\right)^{-1}$.

2.2.2.2. Random Regression Coefficients Model

The random regression coefficient model, in the context of small area estimation, is given by

$$y_{ij} = x'_{ij}\beta_i + e_{ij} \tag{2.19}$$

where $\beta_i = \beta + v_i$, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, T$ with ith small area mean given by

$$\mu_i = M'_i \beta_i = M'_i(\beta + v_i) \tag{2.20}$$

The BLUP of (2.20) for a small area population mean for the case of one covariate is the same as (2.18) with

$$\gamma_i = \sigma_v^2 \left(\sigma_v^2 + \sigma_e^2 \left(\sum x_{ij}^2 \right)^{-1} \right)^{-1},$$

and $\tilde{\beta}$ is simplified to

$$\tilde{\beta} = \left\{ \sum_i \gamma_i \left(\sum_j x_{ij} y_{ij} \right) \left(\sum_j x_{ij}^2 \right)^{-1} \right\} \left(\sum_i \gamma_i \right)^{-1}$$

2.2.2.3. Model using Ecological Level Data

The model based on ecological level data, first considered by Fay and Herriot (1979), is given by

$$\overline{Y}_i = \mu_i + e_i \quad \text{and} \quad \mu_i = \overline{X}'_i \beta + v_i \tag{2.21}$$

where the e_i's and v_i's are independent with $N(0, B_i)$ and $N(0, A)$ respectively, and $\overline{X}'_i = (\overline{X}_{i1}, \overline{X}_{i2}, \ldots, \overline{X}_{ik})$ is a mean vector of k auxiliary variables. The BLUP of μ_i derived by Prasad and Rao (1990) is written as

$$\hat{\mu}_i = \overline{X}_i'\tilde{\beta} + \frac{A}{A+B_i}\left(\overline{Y}_i - \overline{X}_i'\tilde{\beta}\right) \tag{2.22}$$

where assuming B_i is known, $\tilde{\beta} = \left(\overline{\mathbf{X}}'\mathbf{D}^{-1}\overline{\mathbf{X}}\right)^{-1}\left(\overline{\mathbf{X}}'\mathbf{D}^{-1}\overline{Y}\right)$, $\mathbf{D} = \underset{1 \le i \le T}{diag}\ (A + B_i)$,

$\overline{X} = (\overline{X}_1', \ldots, \overline{X}_T')'$ and $\overline{Y} = (\overline{Y}_1, \ldots, \overline{Y}_T)'$.

One of the fundamental problems of the BLUPs given above is that they depend on the variance components, which are unknown. The common practice is to determine the BLUP of the *i*th small area means by replacing these variance components by asymptotically consistent estimators which are briefly discussed by Prasad and Rao (1990). They also discussed the second order mean square estimation of small area mean estimators from the models considered above.

For the variance components model, the MSE approximation given by Prasad and Rao (1990), for $i=1, 2, \ldots, T$, is expressed as

$$MSE(\hat{\mu}_i) = g_{1i}(\sigma^2) + g_{2i}(\sigma^2) + 2g_{3i}(\sigma^2). \tag{2.23}$$

Here, $\sigma^2 = (\sigma_e^2, \sigma_v^2)$ and the estimators of the two variance components are given by

$\hat{\sigma}_e^2 = (n - T - k + \lambda)\sum_{i=1}^{T}\sum_{j=1}^{n_i}\hat{e}_{ij}^2$ and $\tilde{\sigma}_v^2 = (1/n_*)\left[\sum_{i=1}^{T}\sum_{j=1}^{n_i}\hat{u}_{ij}^2 - (n-k)\hat{\sigma}_e^2\right]$ for T

areas and k auxiliary variables, $\lambda = 0$ if the variance component model has no intercept

and $\lambda = 1$ otherwise, and $n_* = n - tr\left[(\mathbf{X}'\mathbf{X})^{-1}\sum_{i=1}^{T} n_i \overline{X}_i \overline{X}_i'\right]$. Furthermore, $\{\hat{e}_{ij}\}$ and

$\{\hat{u}_{ij}\}$ are residuals from the ordinary least square regression of

$\{Y_{ij} - \overline{Y}_i\}$ on $\{X_{ij1} - \overline{X}_{i1}, ..., X_{ijk} - \overline{X}_{ik}\}$ and $\{Y_{ij}\}$ on $\{X_{ij1}, ..., X_{ijk}\}$ respectively. Note that we define $\hat{\theta}_i = \max(\tilde{\theta}_i, 0)$ for a variance component θ_i, which is required to be non-negative.

The three terms given in the above MSE approximation of (2.23) are respectively given by,

$$g_{1i}(\sigma^2) = (1 - \gamma_i)\sigma_v^2$$
$$g_{2i}(\sigma^2) = (M_i - \gamma_i \overline{X}_i)' (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} (M_i - \gamma_i \overline{X}_i), \text{ and}$$
$$g_{3i}(\sigma^2) = n_i^{-1}(\sigma_v^2 + \sigma_e^2 n_i^{-1})^{-3} \left[\sigma_e^4 \operatorname{var}(\tilde{\sigma}_v^2) + \sigma_v^4 \operatorname{var}(\tilde{\sigma}_e^2) - 2\sigma_e^2\sigma_v^2 \operatorname{cov}(\tilde{\sigma}_v^2, \hat{\sigma}_e^2)\right]$$

where, assuming normality for the two residuals, $\operatorname{var}(\tilde{\sigma}_e^2) = 2(n-t-k+\lambda)^{-1}\hat{\sigma}_e^2$,

$\operatorname{var}(\tilde{\sigma}_v^2) = 2n_\cdot^{-2}\left[(n-T-k+\lambda)(T-\lambda)(n-k)\sigma_e^2 + 2n_\cdot\sigma_e^2\sigma_v^2 + n_{\cdot\cdot}\sigma_v^4\right]$,

$\operatorname{cov}(\tilde{\sigma}_e^2, \tilde{\sigma}_v^2) = -(T-\lambda)n_\cdot^{-1}\operatorname{var}(\tilde{\sigma}_e^2)$, and $n_{\cdot\cdot} = tr(\mathbf{MZZ}')^2$ with $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

For the random regression coefficients model, Prasad and Rao (1990) gave the MSE expression for the case of $k = 1$ as in (2.23)with

$$g_{1i}(\sigma^2) = \overline{X}_i(1 - \gamma_i)\sigma_v^2$$
$$g_{2i}(\sigma^2) = \overline{X}_i(1 - \gamma_i)^2\left(\sum \gamma_i\right)^{-1}\sigma_v^2 \quad \text{and}$$
$$g_{3i}(\sigma^2) = \overline{X}_i^2\left(\sum_j X_{ij}^2\right)^{-2}\left[\sigma_v^2 + \sigma_e^2\left(\sum_j X_{ij}^2\right)^{-1}\right]^{-3}\left(\sigma_e^4 \operatorname{var}(\tilde{\sigma}_v^2) + \sigma_v^2 \operatorname{var}(\tilde{\sigma}_e^2) - 2\sigma_e^2\sigma_v^2 \operatorname{cov}(\tilde{\sigma}_v^2, \tilde{\sigma}_e^2)\right)$$

Furthermore, under normality of $\{v_i\}$ and $\{e_{ij}\}$

$$\operatorname{var}(\tilde{\sigma}_e^2) = 2(n-T)^{-1}\sigma_e^2$$
$$\operatorname{var}(\tilde{\sigma}_v^2) = 2\tilde{n}_\cdot^{-2}\left[(n-1)(T-1)(n-T)^{-1}\sigma_e^4 + 2\tilde{n}_\cdot\sigma_e^2\sigma_v^2 + \tilde{n}_{\cdot\cdot}\sigma_v^4\right] \quad \text{and}$$
$$\operatorname{cov}(\tilde{\sigma}_e^2, \tilde{\sigma}_v^2) = -(T-1)\tilde{n}_\cdot^{-1}\operatorname{var}(\tilde{\sigma}_e^2)$$

where $\tilde{n}_{..} = tr(MZZ')$ and $\tilde{n}_{.} = \sum\sum X_{ij}^2 - \left[\sum_i \left(\sum_j X_{ij}^2\right)^2 \right]\left(\sum\sum X_{ij}^2\right)^{-1}$.

For the case of a model using ecological level data, the proposed MSE estimator is given by,

$$MSE\ (\hat{\mu}_i) = g_{1i}(A) + g_{2i}(A) + 2\,g_{3i}(A) \tag{2.24}$$

where

$g_{1i}(A) = A\,B_i\,(A + B_i)^{-1}$

$g_{2i}(A) = B_i^2\,(A + B_i)^{-2}\,\overline{X}_i'(\overline{X}'V^{-1}\overline{X})^{-1}\,\overline{X}_i$, and

$g_{3i}(A) = B_i^2\,(A + B_i)^{-3}\ \mathrm{var}(\tilde{A})$

and under normality of $\{v_i\}$ and $\{e_i\}$, $\tilde{A} = (T-k)^{-1}\left[\sum_{i=1}^{T}\hat{u}_i^2 - \sum_{i=}^{T} B_i\left(1 - \overline{X}_i'(\overline{X}'\overline{X})^{-1}\,\overline{X}_i\right)\right]$,

$\mathrm{var}(\tilde{A}) = 2T^{-1}\left(A^2 + 2A\sum_{i=1}^{T} B_i/T + \sum_{i=1}^{T} B_i^2/T\right)$, $\hat{u}_i = \overline{Y}_i - \overline{X}_i'\hat{\beta}$ and $\hat{\beta} = (\overline{X}'\overline{X})^{-1}\,\overline{X}'\overline{Y}$.

The $MSE(\hat{\mu}_i)$ is then estimated by replacing θ_i with $\hat{\theta}_i$ in equations (2.23) and (2.24). For a detailed explanation about the small area estimators for a model based approach and their MSE estimator, refer to Prasad and Rao (1990) and Ghosh and Rao (1994).

2.3 Summary and Discussion

This chapter reviewed several methods of small area estimations. All have certain limitations and we discuss and identify them here. Gini's coefficient, Poverty Gap, Sen's Index and the FGT measure have all specifically been relevant only to income data, requiring positive values. Further, the major limitation of these poverty and need measures is that they focus on one variable, specifically income of a household or number of person below the poverty line. Income is not the only variable that determines the magnitude or extent of an individual's level of need. There are also such other correlated variables as level of education, unemployment rate, number of students attending school, number of hospitals and physicians in an area, which may have greater impact on the need level for that area than its income alone would explain. Although we did not review here, there are methods that have attempted to eliminate confounding effects in an effort to compare the need level of one small area to another small area. Carriere and Roos (1994, 1997) discussed techniques for comparing the small area variations in health care utilization by adjusting the rate for age\sex distributions across small areas. High rate of utilization would indicate an area with high discretionary practices among physicians, serious problems for the given condition, or requiring more need for health care resources. However, the traditional age\gender standardisation is incapable of adjusting for variables which are on a continuous scale. Although such a method is intuitive and easy to comprehend, a method is desired which will offset the limitations experienced by the current techniques while highlighting their strength. The small area estimation strategies we reviewed are found on certain distributional assumptions. In the next chapters we investigate the distributional properties of these techniques.

Chapter 3

Variance Estimation and Inference

We discuss inferential procedures following variance estimations of need indices using various strategies. Variance estimation strategies, using standard, jackknife and bootstrap methods, will be considered. Inference on the small area rates will be discussed on asymptotic, bootstrap-t and percentile methods.

3.1 Variance Estimation

In this section we derive the variance expressions for Gini's coefficient and Gini's distance. We then discuss variance estimation techniques using jackknife and bootstrap methods which are often applied when the form of the parameter estimator is complicated. First, we consider the standard non-parametric variance estimation of Gini's coefficient and Gini's distance. Secondly, we discuss the jackknife variance estimator of these estimators. Finally, we discuss the variance approximation using the bootstrap technique.

3.1.1 Standard Method

In this section we derive the approximate variance estimator for Gini's distance and Gini's coefficient. For n observed values X_i, $i=1,2, \ldots, n$, the Gini's distance given in equation (2.2) can also be expressed as:

$$
\begin{aligned}
G_{dist} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| X_i - X_j \right| \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_{ij}
\end{aligned}
$$

where $Y_{ij} = \left| X_i - X_j \right|$. Thus, the variance expression for Gini's distance is derived as

$$Var(G_{dist}) = \frac{1}{n^2(n-1)^2} Var\left(\sum_{i=1}^{n}\sum_{j=1}^{n} Y_{ij}\right)$$
$$= \frac{1}{n^2(n-1)^2} \sum_{i=1}^{n} Var(\sum_{j=1}^{n} Y_{ij}) + \frac{2}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j'>j\neq i}^{n} Cov(Y_{ij}, Y_{ij'})$$
$$= \frac{1}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} Var(Y_{ij}) + \frac{4}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j'>j\neq i}^{n} Cov(Y_{ij}, Y_{ij'}) \tag{3.1}$$

where $Var(Y_{ij})$ is the population variance of a variable Y_{ij}, $Cov(Y_{ij}, Y_{ij'})$ is the population covariance between Y_{ij} and $Y_{ij'}$. The variance expression in (3.1) can be rewritten as

$$Var(G_{dist}) = \frac{1}{n} C_1 \tag{3.2}$$

where $C_1 = \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} Var(Y_{ij}) + \frac{4}{(n-1)^2}\sum_{i=1}^{n}\sum_{j'>j\neq i}^{n} Cov(Y_{ij}, Y_{ij'})$. Then, $Var(G_{dist})$ is estimated by replacing $\hat{C}_1$ for C_1, where $\hat{C}_1 = \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} s_{Y_{ij}}^2 + \frac{4}{(n-1)^2}\sum_{i=1}^{n}\sum_{j'>j\neq i}^{n} s_{Y_{ij}Y_{ij'}}$,

$s_{Y_{ij}}^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_i\right)^2$, $\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$ and $s_{Y_{ij}Y_{ij'}} = \frac{1}{n}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_i\right)\left(Y_{ij'} - \bar{Y}_i\right)$.

Similarly, Gini's coefficient given in equation (2.1) can be written as

$$G = \frac{1}{2}\hat{R}$$

where $\hat{R} = \frac{\bar{Y}}{\bar{X}}$, $\bar{Y} = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n} Y_{ij}$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. If n is large, $\bar{X}$ should not differ greately from M and we can write

$$\hat{R} - R = \frac{\bar{Y} - R\bar{X}}{\bar{X}} \cong \frac{\bar{Y} - R\bar{X}}{M}.$$

Under this assumption, we have $\hat{R}-R$ converging in distribution to $\dfrac{\overline{Y}-R\,\overline{X}}{M}$.

Therefore,

$$E(\hat{R}-R) \cong \frac{E(\overline{Y} - R\overline{X})}{M} = 0$$

and

$$\begin{aligned} MSE(\hat{R}) = Var(\hat{R}) &\cong \frac{1}{M^2} Var\left(\overline{Y} - R\overline{X}\right) \\ &\cong \frac{1}{M^2}\left(Var(\overline{Y}) - 2R\,Cov(\overline{X},\overline{Y}) + R^2\,Var(\overline{X})\right) \end{aligned}$$

Therefore, the ratio estimate $\hat{R}$, for a large sample size *n*, can be assumed unbiased and asymptotically normal (Cochran, 1977). In general the following formula can be derived from $Var(\hat{R})$ to express the sampling variance of Gini's coefficient which is valid for large sample sizes *n*.

$$Var(G) \cong \frac{1}{4nM^2}\left(C_1 - 2RC_2 + R^2 C_3\right) \tag{3.3}$$

where $R = E(\hat{R})$, C_1 is as defined earlier, $C_2 = \dfrac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n} Cov\left(X_i, |X_i - X_j|\right)$ and $C_3 = Var(X)$. The expression $\hat{R} = \overline{Y}/\overline{X}$ is a ratio of sample means for two random variables. Assuming an ignorable sampling fraction, we estimate (3.3) by

$$\text{var}(G) = \frac{1}{4n\,\overline{X}^2}\left(\hat{C}_1 - 2\hat{R}\hat{C}_2 + \hat{R}^2\hat{C}_3\right) \tag{3.4}$$

where $\hat{C}_1 = \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} s_{Y_{ij}}^2 + \frac{4}{(n-1)^2}\sum_{i=1}^{n}\sum_{j'>j\neq i}^{n} s_{Y_{ij}Y_{ij'}}$, $\hat{C}_3 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$,

$$\hat{C}_2 = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n} Cov\left(X_i, |X_i - X_j|\right), Cov\left(X_i, |X_i - X_j|\right) = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n}(X_i - \bar{X})(Y_{ij} - \bar{Y})$$

and $\bar{Y} = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n} Y_{ij}$.

3.1.2 Jackknife Method

The jackknife method was initially introduced for the purpose of estimating and reducing the bias of an estimator. However, it has become a major tool after Tukey (1958) observed that the jackknife method could also be used to estimate the variance for any estimator of a characteristic. The jackknife method reduces the bias in the estimator and gives its variance estimate by deleting one datum each time from the original data set and recalculating the estimates based on the rest of the data (Shao and Tu, 1995).

We are interested in estimating a parameter ϑ and variance of $\hat{\vartheta}$ based on a given data set consisting of an identically and independently distributed sample $X_1, X_2, \ldots, X_n$ of size n taken from a distribution function F having observed values ($X_1=x_1, \ldots, X_n=x_n$). Let

$$\hat{\vartheta} = \hat{\vartheta}\left(x_1, x_2, \ldots, x_n\right)$$

$$\hat{\vartheta}^{(-i)} = \hat{\vartheta}\left(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right)$$

be estimators of ϑ from the complete sample set of n elements and from a data set obtained after deleting the ith observation, respectively. The ordinary jackknife estimator and its variance estimator for ϑ are given by

$$\hat{\vartheta}_J = \frac{1}{n}\sum_{i=1}^{n} \hat{\vartheta}_{(i)}$$

and

$$\begin{aligned} Var_J\left(\hat{\vartheta}\right) &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\vartheta}^{(-i)} - \hat{\vartheta}^{(\cdot)}\right)^2 \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\vartheta}_{(i)} - \hat{\vartheta}_J\right)^2 \end{aligned} \tag{3.5}$$

where $\hat{\vartheta}^{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\vartheta}^{(-i)}$ and $\hat{\vartheta}_{(i)} = n\hat{\vartheta} - (n-1)\hat{\vartheta}^{(-i)}$, which Tukey (1958) defined as the jackknife pseudo-values.

The jackknife method requires repeatedly computing the statistic n times, for which it is less dependent on model assumptions (Shao and Tu, 1995). The jackknife method in some cases can explicitly be a function of the data points $x_1, x_2, \ldots, x_n$. However, the jackknife method of variance estimation has been especially useful when we have a complicated form of an estimator.

3.1.3. Bootstrap Method

An alternative method of variance estimation is the bootstrap technique. The theoretical framework of the bootstrap approach to an estimator $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ is defined by replacing F with $\hat{F}$, an empirical probability distribution (Efron, 1982). Suppose that the data $\mathbf{X} = (X_1, X_2, ..., X_n)$ are identically and independently distributed from F, we obtain the bootstrap estimator of the variance of $\hat{\theta}$ by

$$\begin{aligned} \text{var}_b &= \int \left[\hat{\theta} - E_{\hat{F}}(\hat{\theta})\right]^2 d\prod_{i=1}^{n} \hat{F}(x_i) \\ &= \text{var}\left[\hat{\theta}(X_1^*, ..., X_n^*) \middle| X_1, ..., X_n\right] \end{aligned}$$

where $\{X_1^*, ..., X_n^*\}$ is a random sample drawn with replacement from the observed values $X = (X_1, X_2, ..., X_n)$ which is called a bootstrap sample.

The simplest and the most common method used to approximate var_b is the Monte Carlo approximation. The distribution $\hat{F}$ used to generate the bootstrap data set can be any estimator (parametric or non-parametric) of F based on $X_1, X_2, ..., X_n$ (Shao and Tu, 1995). The simplest and most commonly used non-parametric estimator of F is defined by

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x\}$$

where $I(A)$ is an indicator function of the set A. In order to carry out the estimation of the variance using the Monte Carlo method, we draw a bootstrap sample $\{X_{1b}^*, ..., X_{nb}^*\}$, $b=1, 2, ..., B$, independently from $\hat{F}$; compute $\hat{\theta}^b = \hat{\theta}(X_{1b}^*, ..., X_{nb}^*)$, conditional on $(X_1, X_2, ..., X_n)$ and obtain a Monte Carlo approximation of var_b by

$$\text{var}_b = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}^b - \hat{\theta}_b\right)^2 \tag{3.6}$$

where $\hat{\theta}_b = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^b$. In general, bootstrap and jackknife variance estimators are used when the estimator of a parameter of interest is complicated and deriving the exact form of the variance is difficult or cannot be written explicitly.

3.2. Method of Inference

We discuss methods of inferences for any real valued parameter $\vartheta = \vartheta(F)$based on the asymptotic method and bootstrap technique. The confidence intervals for the poverty and need indices can be constructed using asymptotic or bootstrap techniques (percentile and bootstrap - t methods).

3.2.1 Asymptotic Method

We discuss a testing procedure for comparing poverty and need indices for two or more small areas and the construction of confidence intervals from the asymptotic normal distribution theory. The null hypothesis

$$H_o \;:\; \vartheta_1 = \vartheta_2 = ... = \vartheta_m = \vartheta_o$$

states that the indices in m different areas are the same as the entire area, denoted by ϑ_o.

The test statistic proposed by Carriere and Roos (1994, 1997) can be used to test this hypothesis, and is given by

$$T^2 = \sum_{h=1}^{m}\left(\frac{\hat{\vartheta}_h - \vartheta_o}{\sqrt{\operatorname{var}(\hat{\vartheta}_h)}}\right)^2 \tag{3.7}$$

If the overall need index ϑ_o is unknown, an appropriate estimator is taken to make the inference and the statistics T^2 has an approximate chi-square distribution with *(m-1)* degrees of freedom, for large samples for each of the small areas.

Under similar assumptions, a 100(1 - α) % confidence interval for the poverty and need indices can be approximated by:

$$\hat{\vartheta} \pm Z_{\alpha/2}\sqrt{\operatorname{var}(\hat{\vartheta})}$$

For extremely skewed data, Carriere and Roos (1994, 1997) suggested a log transformation on the estimated rates before constructing confidence intervals for the rates.

3.2.2 Percentile Method

Let $\hat{\vartheta}^* = \vartheta(\hat{F})$ be a bootstrap estimator of a parameter ϑ and $\hat{\vartheta}$ an estimator from the original data. Define

$$R(t) = \text{Prob.}\{\hat{\vartheta}^* \le t\}$$

to be a probability distribution function for the bootstrap distribution of $\hat{\vartheta}^*$. If the Monte Carlo method is used to determine the bootstrap distribution, then *R(t)* is approximated by $\hat{R}(t) = \#\ \{\hat{\vartheta}^b \le t\}/B$ (Efron, 1982). For a given α,

$$\hat{\vartheta}_L = \hat{R}^{-1}(\alpha) \text{ and } \hat{\vartheta}_U = \hat{R}^{-1}(1-\alpha).$$

Then the 100 (1 - 2α) % central percentile confidence interval for ϑ is given by

$$\left[\hat{\vartheta}_L, \hat{\vartheta}_U\right].$$

3.2.3 The bootstrap - *t* method

Let $\{X_1^*, \ldots, X_n^*\}$ denotes an identically and independently distributed sample from $\hat{F}$, an estimator of F (parametric or nonparametric). The bootstrap - *t* method of constructing confidence interval depends on a statistic

$$\psi_n = \frac{\hat{\vartheta}_n - \vartheta}{\hat{\sigma}_n}$$

where $\hat{\vartheta}_n$ is an estimator of ϑ and $\hat{\sigma}_n^2$ is a variance estimator for $\hat{\vartheta}_n$. Assume that G_n is the distribution of ψ_n, which is unknown in most cases. It is most often estimated by the bootstrap estimator defined by

$$G_b(x) = \text{Prob}_* \{\psi_n^* \leq x\}$$

where $\psi_n^* = (\hat{\vartheta}_n^* - \hat{\vartheta}) / \hat{\sigma}_n^*$, and $\hat{\vartheta}_n^*$ and $\hat{\sigma}_n^*$ are the bootstrap analogues of $\hat{\vartheta}_n$ and $\hat{\sigma}_n$, respectively. Then, the 100 (1 - 2α) % confidence bound for *6* using this method is given by

$$\left[\hat{\vartheta}_L, \hat{\vartheta}_U\right]$$

where $\hat{\vartheta}_L = \hat{\vartheta}_n - \hat{\sigma}_n G_b^{-1}(1-\alpha)$ and $\hat{\vartheta}_{up} = \hat{\vartheta}_n - \hat{\sigma}_n G_b^{-1}(\alpha)$ which will be called the bootstrap - *t* lower and upper confidence bound (Shao and Tu, 1995).

Chapter 4

Simulation and Data Analysis

This chapter presents the analysis of simulated income data using poverty and needs measures. The summary measures of poverty and need indices discussed in chapter 2, the variance estimation of these estimated indices and the inference procedures described in chapter 3 are to be verified. We also discuss data reduction methods for reducing large numbers of variables to smaller numbers without loosing much information. The data from the Canada Census in 1991 will illustrate small area analysis of health care utilizations for 12 Regional Health Authorities (RHA) in the province of Manitoba.

4.1 Simulation Study

The purpose of the simulation is to validate and recommend an appropriate inference procedure for small area measures of needs. One thousand simulated samples were used in a study of income data of sizes 20, 40 and 100 independently generated from a log-normal distribution based on an average income of $23,800.00 and standard deviation of $3735.60. The average income and standard deviation used for the simulation study is obtained from the Canada Census data for the province of Alberta. The results presented in the simulation study include summary measures of estimators for 6 poverty and need indices. It also presents a gain in per cent of relative bias for empirical, bootstrap and jackknife methods of variance estimation compared to the theoretical variance of Gini's distance and Gini's coefficient. Finally, we present the confidence intervals for both need indices.

Table 1 reports the summary measures (Minimum, Median, Mean and Maximum of poverty and need indices) based on standard, bootstrap and jackknife methods. All poverty indices, other than Gini's coefficient and Gini's distance, depend mainly on the poverty line, which is estimated as one-half of the estimated median family income. These measures are determined from the one thousand simulated income data using the methods discussed in Chapter 2. The estimators obtained from the bootstrap and jackknife methods are also included. It can be seen from Table 1 that the range of all indices becomes narrower as the sample size increases from 20 to 100. In particular, the "mean" column denotes the estimate of the respective need index. All seem very similar with no substantial discrepancies.

Table 1: The summary measures of poverty and need indices.

Indices	No. of Obser.	Minimum			Median			Mean			Maximum		
		Standard method	Boot Method	Jack Method	Standard method	Boot Method	Jack Method	Standard method	Boot Method	Jack Method	Standard method	Boot Method	Jack Method
Poverty line	20	6819.34	7770.18	5934.08	12949.55	13653.56	11932.76	13148.72	13868.76	12125.37	22926.00	22527.62	22911.35
	40	8206.68	8214.97	6655.05	12350.70	12474.36	11789.02	12460.79	12596.23	11959.53	19117.18	19234.39	18470.33
	100	9199.50	9430.53	8987.79	11978.07	12085.82	11780.33	12140.95	12149.03	11890.27	15768.71	15398.38	15296.27
Gini's distance	20	7699.66	7564.56	7699.66	22147.04	21887.85	22147.04	22993.80	22933.75	22993.80	58503.14	60869.60	58503.14
	40	12292.07	11751.35	12292.07	22492.98	21856.25	22492.98	22957.13	22383.76	22957.13	40875.15	41275.49	40875.15
	100	15273.57	15159.17	15273.57	22900.26	22612.52	22900.26	23508.67	22834.23	23058.67	36772.40	37030.47	36772.40
Gini-Coeffi.	20	0.18520	0.17848	0.19495	0.35441	0.33922	0.37509	0.35678	0.34090	0.38050	0.59456	0.54007	0.70048
	40	0.23377	0.22508	0.24041	0.36704	0.35547	0.37849	0.36836	0.35663	0.38041	0.50503	0.48461	0.55141
	100	0.30053	0.29698	0.30372	0.37552	0.37062	0.38034	0.37653	0.37149	0.38142	0.47088	0.46534	0.48126
Sen's Index	20	0.00000	0.00000	-0.29045	0.06629	0.09138	0.06823	0.07122	0.09324	0.06462	0.24429	0.17887	0.37468
	40	0.00000	0.01391	-0.65713	0.06458	0.07340	0.06562	0.06611	0.07597	0.06114	0.16264	0.15526	0.43386
	100	0.01668	0.02311	-0.30275	0.06297	0.06288	0.06188	0.06442	0.06382	0.05905	0.14027	0.13210	0.46416
Povert-Gap	20	0.00000	0.00000	-0.19973	0.05113	0.05248	0.04569	0.05537	0.05561	0.04651	0.18452	0.14650	0.17766
	40	0.00000	0.00296	-0.16078	0.04827	0.04832	0.04492	0.04986	0.04957	0.04532	0.12496	0.11962	0.12790
	100	0.01395	0.01767	-0.05237	0.04655	0.04694	0.04451	0.04728	0.04764	0.04535	0.10541	0.09953	0.10508
FGT Measure	20	0.00000	0.00000	-0.09368	0.02048	0.02091	0.01625	0.02329	0.02376	0.01821	0.10134	0.08393	0.08486
	40	0.00000	0.00042	-0.03849	0.01925	0.01969	0.01694	0.02063	0.02079	0.01810	0.06672	0.06369	0.06498
	100	0.00422	0.00489	0.00121	0.01860	0.01885	0.01779	0.01935	0.01969	0.01839	0.05486	0.05195	0.05468

Table 2: Per cent bias of the variance from standard method relative to variance estimates obtained from empirical, jackknife and bootstrap methods.

Indices	Empirical Method			Bootstrap Method			Jackknife Method		
	Sample size			Sample size			Sample size		
	20	40	100	20	40	100	20	40	100
Gini's distance	16.24	3.06	-1.77	8.73	-1.66	-3.88	16.24	3.06	-1.77
Gini's coefficient	22.44	20.20	18.71	34.53	29.36	22.82	6.81	11.80	15.10

Table 2 report the per cent relative bias of the empirical, jackknife and bootstrap estimates of variance over the theoretical (standard) variance estimate of the two need indices, namely Gini's distance and Gini's coefficient. The per cent in relative bias (RB) is obtained from the expression given by

$$RB = \frac{\left[Var\ (\hat{\vartheta}) - E\left(Var\ (\hat{\vartheta}) \right) \right]}{Var\ (\hat{\vartheta})} x\ 100\ \%\ .$$

Here, $E\left(Var\,(\hat{\vartheta})\right)$ represents an average of the theoretical (standard) variance estimates of the two indices derived in section 3.2.1 and $Var(\hat{\vartheta})$ is an estimate for variance obtained from the empirical, jackknife and bootstrap methods, where all are determined from the one thousand simulated data.

It is seen from Table 2 that the bias of Gini's distance decreases as the sample size of the data increases. The variance estimate obtained from the empirical, bootstrap and jackknife overestimates the theoretical variance in small samples. As the sample size increases, the relative bias in estimating these indices becomes smaller approaching essentially zero. On the other hand the variance for Gini's coefficient performed unsatisfactorily, suggesting that the first order approximation was unsatisfactory, no matter which methods were used.

Table 3: 95% Confidence intervals for the poverty and need indices using the asymptotic, percentile and bootstrap-t methods.

Need indices	Method	Sample Size					
		20		40		100	
		L	U	L	U	L	U
Gini's Distance	Asymptotic	10362.74	35624.86	13736.20	32178.06	17008.21	29009.13
	Percentile	11645.45	37910.09	14749.93	32762.63	17594.89	29245.25
	Bootstrap-t	10680.10	35624.86	13070.98	31231.97	16652.47	28528.18
Gini's Coeff	Asymptotic	0.21170	0.50185	0.26940	0.46731	0.31572	0.43854
	Percentile	0.24357	0.47642	0.28810	0.45118	0.32215	0.43056
	Bootstrap-t	0.20890	0.49671	0.26982	0.46385	0.31584	0.43850

Tables 3 gives confidence limits for the need indices. The lower and upper confidence bound of poverty indices (based on the nominal error rate of 2.5% in each tail) is computed. Three different methods of constructing confidence intervals, asymptotic, percentile and bootstrap-t method, are applied.

As can be seen in Table 3, the confidence bounds obtained from the asymptotic method are very similar to those determined from the other two methods. The percentile method tends gave much narrower confidence intervals than the bootstrap-*t* method and the asymptotic results. As the sample size increases, the confidence intervals from the three methods were very similar with no discernible discrepancies of the coverage probabilities. Therefore, we suggest that the asymptotic approach discussed in section 3.2.1 may be an appropriate method to construct confidence interval of the indices.

A method commonly used to determine the accuracy of an approximation to a distribution is to compare the coverage probabilities. However, as we do not know the population quantities of a Gini's distance and Gini's coefficient, we did not consider such comparison.

4.2 Data Reduction Strategies

The main objective in this section is to reduce the dimension of the data to a univariate data without losing much information to consider the simple methods discussed in chapter 2. The success of data reduction will be determined by how much variance can be explained or captured by this univariate data. Then we propose that this new variable, which is a function of all relevant multivariables, can be used to give a need index for the areas. We plan to explore how the various adjusted index measures perform in capturing and comparing the needs across small areas.

4.2.1 Principal Component Method

For a p-variate data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_p)$ with correlation matrix $\boldsymbol{\rho}$, we consider principal component (PC) analysis to reduce the dimension by explaining the total variation of X's through a few underlying uncorrelated variables which are linear combinations of the original variables (Johnson and Wichern, 1992). Principal components depends mainly on the correlation matrix $\boldsymbol{\rho}$ (or covariance matrix Σ). Let the original variables X have a correlation matrix $\boldsymbol{\rho}$ with an eigen-value vector λ with $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_p \geq 0$ and its corresponding eigenvectors $\gamma = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p)$ and consider the following linear combinations:

$$Y = \gamma X$$

such that, $\quad Var(Y_i) = \mathbf{e}_i' \boldsymbol{\rho}\, \mathbf{e}_i = \lambda_i \quad Cov(Y_i, Y_k) = \mathbf{e}_i' \boldsymbol{\rho}\, \mathbf{e}_k = 0$, for $i \neq k$.

Therefore, the principal components are those uncorrelated linear combinations Y_1, Y_2, ..., Y_p. The first principal component has the largest variance. The variance explained by the kth principal component is given by

$$\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} \quad , \quad k = 1, 2, \ldots, p$$

The PC method is successful if it can reduce the p multi-variables down to a very few, while these variables can explain a substantial amount of variance. What we anticipate is that the PC method may be able to reduce the dimension of socio-economic and health relevant variables to one variable that explains a very high proportion of the variance. Our approach will work if the first PC explains most of the variance so that no real loss of information occurs. The first PC can be used to compute a needs measure using the indices discussed in chapter 2.

4.2.2 Factor Analysis Method

Similarly, if one can postulate a common factor model for **X**, then factor analysis can be used to postulate **X** as being linearly dependent upon a few unobservable random variables F_1, F_2, ..., F_m referred to as a common factors, with p additional sources of variation ε_1, ε_2, ..., ε_p . The factor model is given by

$$\mathbf{X} = \mu + \mathbf{L}\,\mathbf{F} + \varepsilon$$

where **L** is the matrix of factor loading, μ is the general mean vector and the covariance matrix of **X** is given by Σ. The first factor is one that explains the variance in **X** most. Similarly as in PC approach the success using factor analysis depends on how good the first few factors are in capturing and explaining the variance in **X**. Furthermore, the success of the factor analysis method depends on whether such a model is tenable. In the absence of such assumptions, the PC method of data reduction may be more robust. Similarly, the first factor can be used to compute a need measure.

4.2.3 Multiple Linear Model Method

This method is applicable when a clear dose-response relationship can be found (i.e. one response variable is identified as a dependent variable against several explanatory variables) as in

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,. \qquad (4.1)$$

We may predict the value for $\mathbf{y}$, $\hat{\mathbf{y}}$, for the responses explained or predicted by a set of explanatory variables, according to the model (4.1).

These predicted values can be used to give weighted average values for each area. In health services research, these values have been commonly used to give the measures of needs (Roos et. al. 1996a, 1996b).

4.3 Numerical Example: The Case of Manitoba

We compare various approaches for small area analyses. The purpose of this section includes studying the effect of employing various data reduction technique of a large number of variables into univariate data and applying needs index measures to the reduced one-dimensional data. These will be illustrated using the Manitoba data to estimate the average health status across 12 Regional Health Authorities (RHA). In the estimation of health statuses across small areas we also consider three different model based approaches. The auxiliary data in this example have six socio-economic variables, namely the per cent of unemployment among ages 15 to 24 and 45 to 54, the per cent female householders, the per cent of high school graduates among 25 to 34 year olds, the per cent women in the labour force, and the dwelling value all available for each of 241 municipalities ($n = 241$). Because the data were available in a standardized rate, the only appropriate measure of inequality is the Gini's distance. The sample size for RHA ranges from 1 to 43 municipalities.

One RHA was removed from the analysis because it contains only one municipality, and therefore we deal with the remaining 11 RHAs. We apply the data reduction techniques to reduce the dimension of the multivariables into one. The first principal component and the first factor explained 38.20% of the variance in the socio-economic variables. We use these first PC and first factor to compute the Gini's distance. Later in this section, comparisons will be made between the small area rates using the data reduction methods and those based on much fuller model based approaches. We have an overall Gini's distance of 1.5370 and 1.0151 obtained from the first principal component and first factor score, respectively, with the corresponding standard error estimates of 0.4970 and 0.9887.

Table 4: Gini's distance for the province of Manitoba using the first Principal Component.

RHA Code	Gini distance	95% C.I. Lower L	95% C.I. Upper L	Standard Error.
A	1.0095	0.6750	1.5096	0.2073
BN	1.5918	1.0656	2.3777	0.3259
BS	1.3255	0.5619	3.1269	0.5804
C	1.5008	1.0755	2.0942	0.2551
D	1.9076	1.6133	2.2557	0.1631
E	1.1080	0.8304	1.4785	0.1631
FB	2.4275	1.7878	3.2960	0.3788
G	3.2748	1.6002	6.7020	1.1966
GM	1.4931	0.9671	2.3051	0.3308
GS	0.7191	0.5492	0.9415	0.0989
K	0.8037	0.4727	1.3666	0.2177
Over all	1.5370			

Tables 4 and 5 summarizes the estimation results across the 11 RHAs. The confidence intervals for Gini's distance are obtained using the formula $\hat{\vartheta}\exp\{\pm Z_{\alpha/2}\,\mathrm{var}_l(\hat{\vartheta})\}$ (Carriere and Roos, 1994, 1997) where $\mathrm{var}_l(\hat{\vartheta})$ represents the standard error estimate of $ln(\hat{\vartheta})$ and approximated by $\mathrm{var}(\hat{\vartheta})/\hat{\vartheta}$ using a first order Taylor approximation. As the goal in small area analyses is to investigate the needs level and their variation across small areas, the Gini's distance score for each RHA against the overall measure will be examined.

Table 5: Gini's distance for each RHA using the first factor score.

RHA Code	Gini distance	95% C.I. Lower L	95% C.I. Upper L	Standard Error.
A	0.6667	0.4458	0.9971	0.1369
BN	1.0513	0.7038	1.5704	0.2153
BS	0.8755	0.3711	2.0652	0.3833
C	0.9912	0.7104	1.3832	0.1685
D	1.2599	1.0655	1.4898	0.1077
E	0.7318	0.5484	0.9765	0.1077
FB	1.6033	1.1808	2.1769	0.2502
G	2.1629	1.0569	4.4265	0.7903
GM	0.9861	0.6388	1.5224	0.2185
GS	0.4749	0.3627	0.6218	0.0653
K	0.5309	0.3122	0.9026	0.1438
Over all	1.0151			

Gini's distance across the small areas appear quite diverse (T^2=66.77 and 106.16 for the first PC and first factor score analysis, respectively). It is seen in both Tables 4 and 5 that both approaches led to the same conclusion. That is, the Gini's distance for three RHAs (D, FB and G) are greater than the overall indicating a greater disparity across the areas in their socio-economic conditions, while those of four RHAs (A, E, GS, and K) are smaller than the overall. The remaining four RHAs (BN, BS, C and GM) are similar to the overall.

Patterns of health care utilization may need to be assessed with detailed knowledge of socio-economic characteristics. Socio-economic status is an important determinant of health status and there is a broad range of conditions that can adversely affect the health of various segments of the population. We consider the more detailed model based approach for estimation of small area means summarized in chapter 2 in this data set. We will compare the results with those based on simple indices. Recall that, in using the simple indices, there was no dependent variable to be explained by the set of socio-economic variables, as in the model-based approaches. The dependent variable here is a composite measure of general health status for the municipalities (Mustard and Frohlich, 1995). The larger the score on health status, the poorer the municipality. Three different techniques, Multiple Linear (MLM), Variance Components (VCM) and models based on Ecological Level (EL) data, are considered to fit the model of the dependent data on the six socio-economic data, as described earlier. Table 6 shows the mean estimate of health status obtained from these models and their corresponding standard errors for each RHA.

Table 6: Model based small area estimation for health status and the standard error of these estimators for RHA's for the province of Manitoba.

RHA	Small Area Mean Estimators			Estimated Standard Error		
	MLM	VCM	EL model	MLM	VCM	EL model
A	0.3841	0.4132	0.4518	0.1108	0.1264	0.8663
BN	1.1229	0.9284	0.9303	0.1317	0.1805	1.6582
BS	0.3978	0.0858	0.0209	0.0868	0.2058	0.4275
C	0.6418	0.5352	0.5557	0.1180	0.1456	1.0404
D	0.9703	1.2032	1.5334	0.1153	0.2353	1.1766
E	0.7485	0.5390	0.5397	0.1170	0.1456	0.9663
FB	1.7954	2.1553	2.4119	0.1801	0.1994	1.6061
G	-0.1193	-0.3255	-0.3607	0.1071	0.2886	0.2686
GM	0.5168	0.4166	0.4350	0.1224	0.1199	1.0068
GS	0.2543	0.4433	0.5181	0.1059	0.1301	0.5580
K	-0.0442	-0.3860	-0.4623	0.1314	0.1996	0.2521
T^2	160.48	117.67	34.06			

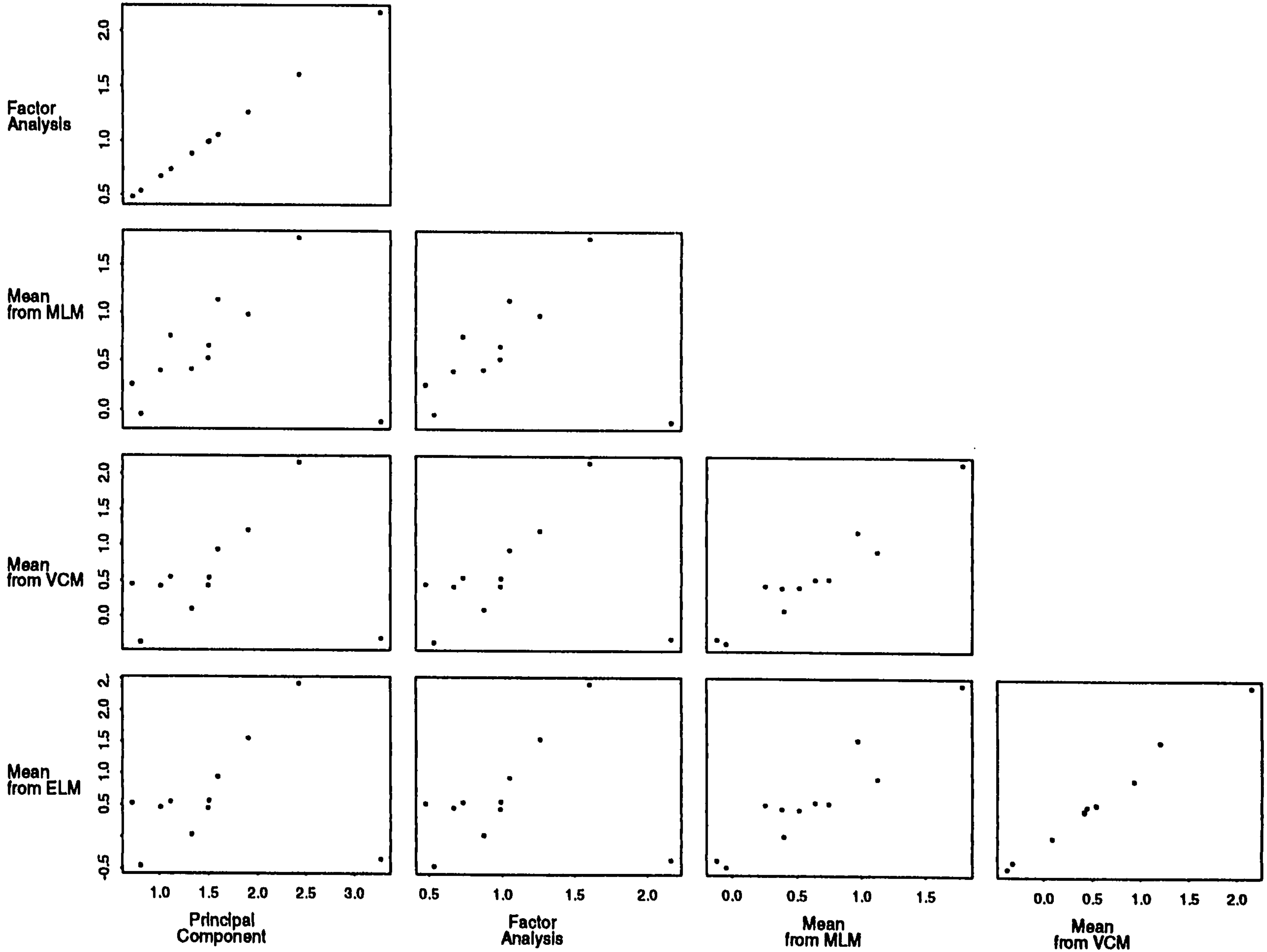
It can be seen from Table 6 that there is a significant variation in small area need inequality across the RHAs. Specifically, the RHA FB was found to be the one that is in need to improve its need level (health status) which has similar to the earlier findings with or without considering the composite score on health status. The results reported in Table 6 also shows that the standard errors of small area estimates obtained from the model based on ecological level data are significantly larger than those from the multiple linear and variance component models due to using summarized information. The model based estimates of needs from the multiple linear and variance component models give smaller standard errors because they use individual level data from each of the 241 municipalities. However, the overall ranking of RHAs in the order of their needs level was somewhat consistent. Table 7 reports that there is a strong linear relationship of the small area estimates of needs obtained from each model. The scatter plots given in Figures 2 also support this finding. Figure 2 shows scatter plots of the inequality estimates obtained from data reduced into one variable using principal component and factor analysis, and the model based small area needs estimates. The Gini's distance score of RHA G did not appear to correlate very well with the model based small area estimates due to small sample size: there are only three municipalities in this RHA. Other than this RHA, their rank order appears to be consistent with RHA coded FB the poorest and GS and K the healthiest.

Table 7: Pearson's coefficient of correlation of the health status estimates for 11 RHAs.

	Ecological level	Variance Com.	Multiple Linear	Principal Com.
Variance com.	0.9959			
Multiple Linear	0.9435	0.9645		
Principal Com.	0.2254 (0.8590)	0.2106 (0.8656)	0.2008 (0.9042)	
Factor Analysis	0.2254 (0.8590)	0.2106 (0.8656)	0.2007 (0.9042)	0.9999

* The numbers in parentheses indicate the correlation upon removing RHA G.

Figure 2: Scatter plots of inequality measures and model based small area estimates



4.4 Conclusion

Proxies of poverty or needs, such as health care and education resources are believed to supplement the profile for the standard of living of a household in a given area. In general, poverty and needs are defined in terms of adequacy of income, or more generally, of disposable resources, to support a minimum standard of decent living (Foday, 1995). Such techniques for defining poverty and needs are focused mainly on income, consumption and expenditure. In this thesis we stressed that income is not the only factor to consider in the study of the state of poverty. The impact of other factors such as level of education, state of health, the demographic structure, are all crucial determinants of poverty and needs. This thesis discussed methodological approaches for incorporating these factors into measures of needs analysis. One of the methods advocates multivariate data reduction strategy to reduce the dimension of variables, if possible to one dimension, without losing much information contained in the data. Therefore, depending on data availability, one may take this one-dimensional variable to determine the extent and severity of poverty and need for a given area.

We presented methods of variance estimation and inference procedures for Gini's distance and Gini's coefficient. The per cent in relative bias is used to compare the standard variance estimate with those estimates obtained using empirical, jackknife and bootstrap methods based on one thousand simulated income data. The results reported in Table 3 shows that the per cent in relative bias of Gini's distance is close to zero for large sample size. Whereas, the variance estimate of Gini's coefficient from the standard method is not satisfactory, which indicates that Taylor's first order approximation does not provide precise estimate of variance.

Different methods of small area estimation and their variance estimation strategies along with inference methods have been reviewed in this thesis. Such indirect estimation includes design based, synthetic, composite and model based estimators. Based on these small area estimation strategies, the health status data collected in the province of Manitoba have been utilized to investigate the level of needs for the RHAs. There are broad range of conditions that can adversely affect needs (health status) of various segments of the population in a given area. In this thesis we compared the interpretations of needs across Manitoba RHAs based on various approaches. Unless the sample size is too small, the overall explanation seem consistent whether one uses model based approaches or simple index measures.

REFERENCES

Battese, G.E. and Fuller, W.A. (1981). *Prediction of County Crop Areas Using Satellite Data.* Proceeding of the Section on Survey Research Methods, American Statistical Association, 500-505.

Carriere, K.C. and Roos, L.L. (1994). *Comparing Standardised Rates of Events.* American Journal of Epidemiology, 140, 472-82.

Carriere, K.C. and Roos, L.L. (1997). *A Method of Comparison for Standardized Rates of Low-Incidence Events.* Medical Care, 35, 57 – 69.

Cochran, W.G. (1977). *Sampling Techniques* (Third Edition). New York: John Wiley and Sons.

Dagum, C. (1983). *Income Inequality Measures.* Encyclopedia of Statistical Sciences, 4, 34-40.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia.

Foday, J. (1995). *Poverty in Africa: Towards a Methodology and Frame Works for Analysis.* Paper Presented at Joint IAOS/AFSA Conference, Addis Ababa.

Foster, J., Greer, J. and Thurbecke, E. (1984). *Notes and Comments on a class of Decomposable Poverty Measures.* Econometrica, 52, 761-66.

Ghosh, M. and Rao, J.N.K. (1994). *Small Area Estimation: An Appraisal.* Statistical Sciences, 9, 55-93.

Henderson, C.R. (1975). *Best Linear Unbiased Estimation and Prediction under a Selection Model.* Biometrics, 31, 423-47.

Johnson, A.E. (1993). *Small Area Estimation.* Proceedings of the Section on Survey Research Methods, American Statistical Association, 352 - 357.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis.* New Jersey: Prantice-Hall.

Mustard, C.A. and Frohlich, N. (1995). *Socioeconomic Status and the Health of the Population.* Med Care 33(12): DS43 – DS54.

Prasad, N.G.N. and Rao, J.N.K. (1990). *On the Estimation of Mean Square Error of Small Area Predictors.* Journal of American Statistical Association, 85, 163-171.

Rao, J.N.K. (1986). *Synthetic Estimators, SPREE and Best Model-Based Predictors on Small Area Means.* Technical Report 97, Carlton University, Laboratory for Research in Statistics and Probability, Ottawa.

Roos, N.P., Black C.D., Wade J., Decker K., et al. (1996a). *How Many General Surgeons do you need in rural areas? Three Approaches to Physicians Resource Planning in Southern Manitoba.* Can Med Assoc., 155(4), 395-401.

Roos N.P., Fransoo R., Bogdanovic B., et al. (1996b). *Needs-Based Planning for Manitoba's Generalist Physicians.* Manitoba Centre for Health Policy and Evaluation, Winnipeg.

Sen, A. (1976). *Poverty: An Ordinal Approach to Measurement.* Econometrica, 44, 219-31.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* New York: Springer-Verlag.

Shao, J. and Rao, J.N.K. (1993). *Standard Errors for Low Income Proportion Estimated from Stratified Multi-Stage Samples.* Sankhya B, 55, 393-414.

Todaro, M.P. (1990). *Economic Development in the Third World* (Fourth Edition). New York: Longman.

Tukey, J. (1958). *Bias and Confidence in not quite Large Sample. Abstract.* Annals of Mathematical Statistics, 29, 614.

Wondmagegnehu, E.T. (1995). *Statistical Measures of Poverty with Reference to Ethiopia.* M.Sc. Thesis, School of Graduate Studies, Addis Ababa University, Addis Ababa.

Wondmagegnehu, E.T. and Carriere, K.C. (1997). *Statistical measures of poverty and needs.* Proceedings of the Section of Epidemiology, American Statistical Association, 120-125.

APPENDIX

A.1 Simulation Program (in C language)

```
/*********************************************************************
* The following program is used to calculate the value for poverty and*
* need indices with their corresponding standard errors using a *
* standard, Jackknife and bootstrap method.It also uses to find The *
* relative bias of MSE's for two  need indices, namely gini_distance *
* and gini-coefficient.
*********************************************************************/

#include <stdio.h>
#include <math.h>
#include <string.h>
#include <time.h>
#include <stdlib.h>

#define  sim_size 1000
#define  size  20
#define  boot_size  200
#define RANDMAX  32767

double gini_coef(double [], int);
double sens_index(double [], int);
double poverty_gap(double [], int);
double fgt_measure(double [], int);
double income_gap(double [], int);
double gini_less_pov(double [], int);
int count_less(double [], int);
double pov_line(double [], int);
void arrange(double [], int);
double average(double [], int);
double pseudo(double, double, int);
void create_matrix(double [], double [][size-1]);
void bootstrap_sample(double [], double [], int, int);
double variance(double [], int);
double var_gini_distance(double [], int);
double var_gini_coef(double [], int);
double gini_distance(double [], int);
void generate_data(double [], int, int);


main ()
{ /*Begining of the main program*/


int num1, num2, num3, num4, num5, num, ind;
unsigned int seed;
double *pov_lin, *gini_c, *sen_in, *pov_gp, *fgt_m;
double jack_var_pl, jack_var_gini, jack_var_sen, jack_var_pg,
jack_var_fgt;
double *arrayofy, *row_matrix, matrix[size][size-1] ;
```

```
double *pl_pseudo, *gini_pseudo, *sen_pseudo, *pg_pseudo, *fgt_pseudo;
double *pl_data, *gini_data, *sen_data, *pg_data, *fgt_data;
double *arrayindex, *pl_boot, *gini_boot, *sen_boot, *pg_boot,
*fgt_boot;
double *pl_bootstrap, *gini_bootstrap;
double *pg_bootstrap, *sen_bootstrap, *fgt_bootstrap;
double var_pl_boot, var_gini_boot, var_pg_boot, var_sen_boot,
var_fgt_boot;
double *pl_jackknife, *gini_jackknife;
double *sen_jackknife, *pg_jackknife, *fgt_jackknife;
double pl_sim_data, gini_sim_data, sen_sim_data, pg_sim_data,
fgt_sim_data;
double var_pl_data, var_gini_data, var_sen_data, var_pg_data,
var_fgt_data;
double pl_sim_jack, gini_sim_jack, sen_sim_jack, pg_sim_jack,
fgt_sim_jack;
double pl_sim_boot, gini_sim_boot, pg_sim_boot, sen_sim_boot,
fgt_sim_boot;
double *gini_dist_var, *pov_gap_var, *gini_coef_var;
double *gini_dist_data, *gini_dist_jackknife, *gini_di,
*gini_dist_pseudo;
double *gini_dist_boot, *gini_dist_bootstrap, ginidist_sim_data;
double var_ginidist_data, ginidist_sim_jack, jack_var_ginidist;
double ginidist_sim_boot, var_ginidist_boot;
FILE  *out, *data01, *data02, *data03, *data04;


arrayofy = malloc(size*sizeof(double));
pl_data = malloc(sim_size*sizeof(double));
gini_data = malloc(sim_size*sizeof(double));
sen_data = malloc(sim_size*sizeof(double));
pg_data = malloc(sim_size*sizeof(double));
fgt_data = malloc(sim_size*sizeof(double));
pl_jackknife = malloc(sim_size*sizeof(double));
gini_jackknife = malloc(sim_size*sizeof(double));
sen_jackknife = malloc(sim_size*sizeof(double));
fgt_jackknife = malloc(sim_size*sizeof(double));
pg_jackknife = malloc(sim_size*sizeof(double));
pl_bootstrap = malloc(sim_size*sizeof(double));
gini_bootstrap = malloc(sim_size*sizeof(double));
sen_bootstrap = malloc(sim_size*sizeof(double));
pg_bootstrap = malloc(sim_size*sizeof(double));
fgt_bootstrap = malloc(sim_size*sizeof(double));
gini_dist_var = malloc(sim_size*sizeof(double));
pov_gap_var = malloc(sim_size*sizeof(double));
gini_coef_var = malloc(sim_size*sizeof(double));
gini_dist_data = malloc(sim_size*sizeof(double));
gini_dist_jackknife = malloc(sim_size*sizeof(double));
gini_dist_bootstrap = malloc(sim_size*sizeof(double));

data01 = fopen("out20_data", "w");
data02 = fopen("out20_jack", "w");
data03 = fopen("out20_boot", "w");
data04 = fopen("out20_var", "w");

fprintf(data01, "THE FOLLOWING OUTPUT IS INDIVIDUAL MEASURES OF INDICES
FROM THE ORIGINAL DATA \n");
```

```
fprintf(data01, "CONSISTING OF %4d INCOME DATA FOR EACH SIMULATION
\n\n", size);
fprintf(data01, "Simu. No. Pov. Line   Gini-Dist  Gini-coef  pov-gap
sen's-index fgt-meas \n");

fprintf(data02, "THE FOLLOWING OUTPUT IS INDIVIDUAL MEASURES OF INDICES
FROM THE JACKKNIFE METHOD\n");
fprintf(data02, "CONSISTING OF %4d INCOME DATA FOR EACH SIMULATION
\n\n", size);
fprintf(data02, "Simu. No. Pov. Line   Gini-Dist  Gini-coef  pov-gap
sen's-index fgt-meas \n");

fprintf(data03, "THE FOLLOWING OUTPUT IS INDIVIDUAL MEASURES OF INDICES
FROM THE BOOTSTRAP METHOD\n");
fprintf(data03, "CONSISTING OF %4d INCOME DATA FOR EACH SIMULATION
\n\n", size);
fprintf(data03, "Simu. No. Pov. Line   Gini-Dist  Gini-coef  pov-gap
sen's-index fgt-meas \n");

fprintf(data04, "THE FOLLOWING RESULTS REPRESENTS VARIANCE OF THE Two
INDICES \n");
fprintf(data04, "CONSISTING OF %4d INCOME DATA FOR EACH SIMULATION
\n\n", size);
fprintf(data04, "Simul. No.    Gini-Distance    Gini-Coefficient
\n");

for (num5 = 1; num5 <= sim_size; num5++)
 {

   generate_data(arrayofy, size, num5);

   gini_dist_var[num5-1] = var_gini_distance(arrayofy, size);
   gini_coef_var[num5-1] = var_gini_coef(arrayofy, size);

   pl_data[num5-1] = pov_line(arrayofy, size);
   gini_data[num5-1] = gini_coef(arrayofy, size);
   gini_dist_data[num5-1] = gini_distance(arrayofy, size);
   sen_data[num5-1] = sens_index(arrayofy, size);
   pg_data[num5-1] = poverty_gap(arrayofy, size);
   fgt_data[num5-1] = fgt_measure(arrayofy, size);

   row_matrix = malloc((size-1)*sizeof(double));
   pov_lin = malloc(size*sizeof(double));
   gini_c = malloc(size*sizeof(double));
   gini_di = malloc(size*sizeof(double));
   sen_in = malloc(size*sizeof(double));
   pov_gp = malloc(size*sizeof(double));
   fgt_m = malloc(size*sizeof(double));

   create_matrix(arrayofy, matrix);

   for (num1=0; num1< size; num1++)
     {
       for(num2=0; num2 < size-1; num2++)
         {
           row_matrix[num2] = matrix[num1][num2];
         }
```

```
        pov_lin[num1] = pov_line(row_matrix, size-1);
        gini_c[num1] = gini_coef(row_matrix, size-1);
        gini_di[num1] = gini_distance(row_matrix, size-1);
        sen_in[num1] = sens_index(row_matrix, size-1);
        pov_gp[num1] = poverty_gap(row_matrix, size-1);
        fgt_m[num1] = fgt_measure(row_matrix, size-1);
      }

     free(row_matrix);

     pl_pseudo = malloc(size*sizeof(double));
     gini_pseudo = malloc(size*sizeof(double));
     gini_dist_pseudo = malloc(size*sizeof(double));
     sen_pseudo = malloc(size*sizeof(double));
     pg_pseudo = malloc(size*sizeof(double));
     fgt_pseudo = malloc(size*sizeof(double));

    /*THE FOLLOWING LOOP USES TO FIND PSUEDO VALUES FOR POVERTY
INDICES*/

    for(num3=0; num3 < size; num3++)
       {
         pl_pseudo[num3] = pseudo(pl_data[num5-1], pov_lin[num3], size);
         gini_pseudo[num3] = pseudo(gini_data[num5-1], gini_c[num3],
size);
         gini_dist_pseudo[num3] = pseudo(gini_dist_data[num5-1],
gini_di[num3], size);
         sen_pseudo[num3] = pseudo(sen_data[num5-1], sen_in[num3],
size);
         pg_pseudo[num3] = pseudo(pg_data[num5-1], pov_gp[num3], size);
         fgt_pseudo[num3] = pseudo(fgt_data[num5-1], fgt_m[num3], size);

       }

    free(pov_lin);
    free(gini_c);
    free(gini_di);
    free(sen_in);
    free(pov_gp);
    free(fgt_m);

    pl_jackknife[num5-1] = average(pl_pseudo, size);
    gini_jackknife[num5-1] = average(gini_pseudo, size);
    gini_dist_jackknife[num5-1] = average(gini_dist_pseudo, size);
    sen_jackknife[num5-1] = average(sen_pseudo, size);
    pg_jackknife[num5-1] = average(pg_pseudo, size);
    fgt_jackknife[num5-1] = average(fgt_pseudo, size);


    free(pl_pseudo);
    free(gini_pseudo);
    free(gini_dist_pseudo);
    free(sen_pseudo);
    free(pg_pseudo);
    free(fgt_pseudo);

    /* THE FOLLOWING PART DOES COMPUTATION USING BOOTSTRAP METHOD */
```

```
    arrayindex = malloc(size*sizeof(double));
    pl_boot = malloc(boot_size*sizeof(double));
    gini_boot = malloc(boot_size*sizeof(double));
    gini_dist_boot = malloc(boot_size*sizeof(double));
    sen_boot = malloc(boot_size*sizeof(double));
    pg_boot = malloc(boot_size*sizeof(double));
    fgt_boot = malloc(boot_size*sizeof(double));

    for (num4=1; num4 <= boot_size; num4++)
      {
        seed = rand();
        bootstrap_sample(arrayofy, arrayindex, size, seed);
        pl_boot[num4-1] = pov_line(arrayindex, size);
        gini_boot[num4-1] = gini_coef(arrayindex, size);
        gini_dist_boot[num4-1] = gini_distance(arrayindex, size);
        sen_boot[num4-1] = sens_index(arrayindex, size);
        pg_boot[num4-1] = poverty_gap(arrayindex, size);
        fgt_boot[num4-1] = fgt_measure(arrayindex, size);
      }

    free(arrayindex);

    /*ESTIMATION OF POVERTY AND NEED INDICES USING BOOTSTRAP METHOD*/

    pl_bootstrap[num5-1] = average(pl_boot, boot_size);
    gini_bootstrap[num5-1] = average(gini_boot, boot_size);
    gini_dist_bootstrap[num5-1] = average(gini_dist_boot, boot_size);
    pg_bootstrap[num5-1] = average(pg_boot, boot_size);
    sen_bootstrap[num5-1] = average(sen_boot, boot_size);
    fgt_bootstrap[num5-1] = average(fgt_boot, boot_size);

    free(pl_boot);
    free(gini_boot);
    free(gini_dist_boot);
    free(sen_boot);
    free(pg_boot);
    free(fgt_boot);

  fprintf(data01, "%4d %13.2f %12.2f %9.5f", num5, pl_data[num5-1],
gini_dist_data[num5-1], gini_data[num5-1]);
  fprintf(data01, "%11.5f %9.5f %9.5f \n", pg_data[num5-1],
sen_data[num5-1], fgt_data[num5-1]);

  fprintf(data02, "%4d %13.2f %12.2f %9.5f", num5, pl_jackknife[num5-
1], gini_dist_jackknife[num5-1], gini_jackknife[num5-1]);
  fprintf(data02, "%11.5f %9.5f %9.5f \n", pg_jackknife[num5-1],
sen_jackknife[num5-1], fgt_jackknife[num5-1]);

  fprintf(data03, "%4d %13.2f %12.2f %9.5f", num5, pl_bootstrap[num5-
1], gini_dist_bootstrap[num5-1], gini_bootstrap[num5-1]);
  fprintf(data03, "%11.5f %9.5f %9.5f \n", pg_bootstrap[num5-1],
sen_bootstrap[num5-1], fgt_bootstrap[num5-1]);

  fprintf(data04, "%5d %20.2f %16.5f \n", num5, gini_dist_var[num5-1],
gini_coef_var[num5-1]);
```

```
}

pl_sim_data = average(pl_data, sim_size);
gini_sim_data = average(gini_data, sim_size);
ginidist_sim_data = average(gini_dist_data, sim_size);
sen_sim_data = average(sen_data, sim_size);
pg_sim_data = average(pg_data, sim_size);
fgt_sim_data = average(fgt_data, sim_size);

var_gini_data = variance(gini_data, sim_size);
var_ginidist_data = variance(gini_dist_data, sim_size);

pl_sim_jack = average(pl_jackknife, sim_size);
gini_sim_jack = average(gini_jackknife, sim_size);
ginidist_sim_jack = average(gini_dist_jackknife, sim_size);
sen_sim_jack = average(sen_jackknife, sim_size);
pg_sim_jack = average(pg_jackknife, sim_size);
fgt_sim_jack = average(fgt_jackknife, sim_size);

jack_var_gini = variance(gini_jackknife, sim_size);
jack_var_ginidist = variance(gini_dist_jackknife, sim_size);

pl_sim_boot = average(pl_bootstrap, sim_size);
gini_sim_boot = average(gini_bootstrap, sim_size);
ginidist_sim_boot = average(gini_dist_bootstrap, sim_size);
sen_sim_boot = average(sen_bootstrap, sim_size);
pg_sim_boot = average(pg_bootstrap, sim_size);
fgt_sim_boot = average(fgt_bootstrap, sim_size);

var_gini_boot = variance(gini_bootstrap, sim_size);
var_ginidist_boot = variance(gini_dist_bootstrap, sim_size);

out = fopen("out01_20", "w");

fprintf(out, "THE FOLLOWING RESULT GIVES VALUES OF NEED INDICES AND
\n");
fprintf(out, " THEIR VARIANCE FROM THE ORIGINAL DATA OF %d SIMULATION
SIZE \n\n", sim_size);
fprintf(out, " Poverty Line of the original data =  $%.2f \n",
pl_sim_data);
fprintf(out, " Gini_Coefficient of the original data = %7.5f \n",
gini_sim_data);
fprintf(out, " Gini_distance of the original data = %7.5f \n",
ginidist_sim_data);
fprintf(out, " Sen's Index of the original data = %7.5f \n",
sen_sim_data);
fprintf(out, " Poverty Gap of the original data = %7.5f \n",
pg_sim_data);
fprintf(out, " FGT Measure of the original data = %7.5f \n\n\n",
fgt_sim_data);

fprintf(out, "Variance of gini coef from the data = %15.11f \n",
var_gini_data);
fprintf(out, "Variance of gini dist from the data = %15.11f \n",
var_ginidist_data);
```

```
fprintf(out, "THE FOLLOWING RESULT GIVES VALUES FROM THE JACK METHOD
\n\n");
fprintf(out, " poverty Line from Jack = $%.2f \n", pl_sim_jack);
fprintf(out, " Gini-Coeff from jack = %7.5f \n", gini_sim_jack);
fprintf(out, " Gini-dist from jack = %7.5f \n", ginidist_sim_jack);
fprintf(out, " Poverty Gap from jackknife = %7.5f \n", pg_sim_jack);
fprintf(out, " Sens Index from jack = %7.5f \n", sen_sim_jack);
fprintf(out, " FGT Measure from jack = %7.5f \n\n\n", fgt_sim_jack);

fprintf(out, " Variance of Gini Coeff from jack = %15.11f \n",
jack_var_gini);
fprintf(out, " Variance of Gini dist from jack = %15.11f \n",
jack_var_ginidist);

fprintf(out, "THE FOLLOWING RESULT GIVES VALUES FROM THE BOOT METHOD
\n\n");
fprintf(out, " poverty Line from bootstrap = $%.2f \n", pl_sim_boot);
fprintf(out, " Gini-Coeff from bootstrap = %7.5f \n", gini_sim_boot);
fprintf(out, " Gini-dist from bootstrap = %7.5f \n",
ginidist_sim_boot);
fprintf(out, " Poverty Gap from bootstrap = %7.5f \n", pg_sim_boot);
fprintf(out, " Sens Index from bootstrap = %7.5f \n", sen_sim_boot);
fprintf(out, " FGT Measure from bootstrap = %7.5f \n\n\n",
fgt_sim_boot);

fprintf(out, " Variance of Gini Coeff from bootstrap = %15.11f \n",
var_gini_boot);
fprintf(out, " Variance of Gini dist from bootstrap = %15.11f \n",
var_ginidist_boot);

fprintf(out, " Theoretical variance for Gini-coeff. = %7.5f \n",
average(gini_coef_var, sim_size));
fprintf(out, " Theoretical variance for Gini-dist. = %12.2f \n",
average(gini_dist_var, sim_size));


}/* End of the main program*/

/* CREATE A MATRIX OF ORDER K BY K-1 FROM AN ARRAY OF SIZE K, WHERE iTH
ROW
   IS OBTAINED BY DELETING THE iTH ELEMENT OF THE ARRAY, WHICH IS USED
FOR
   JACKKNIFE METHOD */

void create_matrix(double a[], double result[size][size-1])

{
  int i, j;


  for (i=0; i<= size - 1; i++)
    {
      for(j=0; j<size-1; j++)
        {
          if (j < i)
            result[i][j] = a[j];
          else
```

```
            result[i][j] = a[j+1];
        }
    }
}


/*IT CALCULATES GINI-COEFFICIENT OF A GIVEN ARRAY*/

double gini_coef(double a[], int step)

{

int   i, j, k, count;
double   sum_var, gini, sumofy, sumofx;


sumofy = 0;
sumofx = 0;
for (i=0; i <= step-2; i++)
 {
   for ( j=i+1; j<=step-1; j++)
     {
       sumofy += fabs(a[i] -a[j]);
     }
   sumofx +=a[i];
  }
sumofx += a[step -1];
gini = sumofy/( (double) step * sumofx);

return gini;


}

/*This function calculates Gini-distance of a given array of numbers*/

double gini_distance(double b[], int step)

{

  int  i, j, count;
  double sum, gini_dis;

  sum = 0;
  count = 0;

  for (i = 0; i < step-1; i++)
    {
      for (j = i + 1; j <= step-1; j++)
        {
           sum += fabs(b[i] - b[j]);
           count += 1;
        }
    }
  gini_dis = sum/(double) count;

  return gini_dis;
```

```
}

/*THIS FUNCTION CALCULATES SEN'S INDEX OF AN ARRAY*/

double sens_index(double c[], int step)

{
  int count;
  double sen, h_ratio, in_gap, gini_less;

  count = count_less(c, step);

  h_ratio = (double) count/(double) step;
  in_gap = income_gap(c, step);
  gini_less = gini_less_pov(c, step);
  sen = h_ratio*(in_gap + (1.0 - in_gap)*gini_less);

  return sen;

}

/*THIS USES TO OBTAIN POVERTY GAP OF A GIVEN ARRAY*/

double poverty_gap(double e[], int step)

{
  int i, count;
  double poverty_line, sum_pg, p_gap;

  arrange(e, step);
  count = count_less(e, step);
  poverty_line = pov_line(e, step);
  sum_pg = 0;
  for (i=0; i <= count - 1; i++)
    sum_pg += (poverty_line - e[i]) / poverty_line;

  p_gap = sum_pg / (double) step;

  return p_gap;
}

/*COMPUTES FGT MEASURE OF AN INCOME DATA*/

double fgt_measure(double f[], int step)

{
  int i, count;
  double poverty_line, sum_fgt, fgt_index;

  arrange(f, step);
  poverty_line = pov_line(f, step);
  count = count_less(f, step);
  sum_fgt = 0;
  for (i=0; i <= count - 1; i++)
    sum_fgt += pow((poverty_line - f[i]) / poverty_line, 2.0);
```

```
  fgt_index = sum_fgt / (double) step;

  return fgt_index;

}

/*THIS FUNCTION IS USED TO CALCULATE VARINCE OF GINI-DISTANCE FOR A
GIVE ARRAY*/

double var_gini_distance(double a[], int step)

{

  int i, j, k;
  double sum_var, sum_av, sum_sq, mean[size], var[size], sum_cov;
  double var_dist, sum_sq_cov, const1, const2;

  sum_var = 0;
  for(i=0; i<step; i++)
    {
      sum_av = 0;
      sum_sq = 0;
      for(j=0; j<step; j++)
        {
          sum_av += fabs(a[i] - a[j]);
          sum_sq += pow(a[i] - a[j], 2.0);
        }
      mean[i] = sum_av / (double) step ;
      var[i] = pow(step - 1.0, -1.0)*(sum_sq - step*pow(mean[i], 2.0));
      sum_var += var[i];
    }

   sum_cov = 0;
   for(i=0; i<step; i++)
     {
       for(j=i+1; j<step; j++)
         {
           sum_sq_cov = 0;
           if (i != j)
             {
               for(k=0; k<step; k++)
                   sum_sq_cov += (fabs(a[i] - a[k]) -
mean[i])*(fabs(a[j] -
                                                    a[k]) - mean[j]);
               sum_cov += sum_sq_cov / (double) step;
             }
         }
     }

  const1 = pow(step * pow(step - 1, 2.0), -1.0);
  const2 = 4*pow(step * pow(step - 1, 2.0), -1.0);

  var_dist = const1*sum_var + const2*sum_cov;

  return var_dist;

}
```

```
/*THIS PART OF THE PROGRAM IS USED TO CALCULATE VARINCE OF A GINI-
COEFFICIENT*/

double var_gini_coef(double a[], int step)

{
  int i, j, count;
  double meanofx, meanofy, b[(size*(size - 1))/2];
  double var_x, var1, ratio, sumsqxy, covar, devofx, devofy;
  double coef1, coef2, var_coef;


  var_x = variance(a, step);
  var1 = var_gini_distance(a, step);

  count = 0;
  for(i = 0; i<step; i++)
    {
      for(j = i+1; j<step; j++)
        {
          b[count] = fabs(a[i] - a[j]);
          count += 1;
        }
    }

  meanofx = average(a, step);
  meanofy = average(b, count);

  sumsqxy = 0;
  for ( i=0; i<step; i++)
    {
      devofx = a[i] - meanofx;
      for ( j=i+1; j<step; j++)
        {
          devofy = fabs(a[i] - a[j]) - meanofy;
          sumsqxy += devofx*devofy;
        }
    }

  covar = sumsqxy/ (double) count;

  ratio = meanofy / meanofx;

  coef1 = pow(4.0*step*meanofx*meanofx, -1.0);
  coef2 = var1 - 2.0*ratio*covar*(2.0/(step*(step - 1))) + pow(ratio,
2.0)*var_x;

  var_coef = coef1*coef2;


  return var_coef;

}

/*THIS FUNCTION COMPUTES VARIANCE OF A GIVEN ARRAY OF NUMBERS*/
```

```
double variance(double a[], int step)

{
   int  i;
   double est, sum, var;

   est = average(a, step);
   sum = 0.0;
   for (i=0; i<step; i++)
     sum += pow(a[i] - est, 2.0);

   var = sum / (double) (step - 1);

   return var;

}

/*CALCULATES AVERAGE VALUE*/

double average(double c[], int step)

{
  int i;
  double sum_av, aver;

  sum_av = 0;
  for ( i=0; i<= step - 1; i++)
    sum_av += c[i];
  aver = sum_av / (double) step ;

  return aver;

}

/*COMPUTES PSEUDO VALUE*/

double pseudo(double z1, double z2, int z3)

{
  double value;

  value = z3*z1 - (z3 - 1)*z2;

  return value;

}


/*COMPUTES INCOME GAP OF A GIVEN INCOME DATA WHICH WE USE IT FOR
SEN'S INDEX*/

double income_gap(double c[], int step)

{

  int i, count;
  double poverty_line, sum, inc_gap;
```

```
  poverty_line = pov_line(c, step);
  count = count_less(c, step);
  arrange(c, step);
  sum = 0;
  for(i=0; i<= count-1; i++)
    {
       sum += (poverty_line - c[i])/poverty_line;
    }

  inc_gap = (1.0/(double) count )*sum;

  return inc_gap;
}


/*CALCULATES GINI-COEFFICIENT FOR THOSE INCOME DATA BELOW THE POVERTY
LINE,
AGAIN USED FOR SEN'S INDEX*/


double gini_less_pov(double c[], int step)


{

  int i, j, count;

  double sum_den, sum_num, gin_less;


  arrange(c, step);
  count = count_less(c, step);
  sum_den = 0;
  sum_num = 0;
  for( i=0; i<= count - 2; i++)
    {
      for(j = i+1; j<= count - 1; j++)
        {
          sum_num += fabs(c[i] - c[j]);
        }
      sum_den += c[i];
    }
  sum_den += c[count - 1];

  gin_less = sum_num / ((double) count*sum_den);

  return gin_less;

}


/*THIS FUNCTION COUNTS THE NUMBER OF INCOME DATA BELOW THE POVERTY
LINE*/
```

```
int count_less(double c1[], int step)

{
  int i, count;
  double poverty_line;

  poverty_line = pov_line(c1, step);
  count = 0;
  for(i=0; i <= step - 1; i++)
    if (c1[i] < poverty_line)
      count += 1;

  return count;

}


/*THIS IS USED TO OBTAIN POVERTY LINE*/

double pov_line(double c2[], int step)

{
  double median;
  int j;

  arrange(c2, step);
  if (step % 2 == 0 )
     median = (c2[step/2] + c2[step/2 + 1])/2.0;
   else
     median = c2[(step + 1)/2];

   return median/2.0;

 }


/*THIS FUNCTION ARRANGES THE GIVEN ARRAY OF DATA IN ASSENDING ORDER*/

void arrange(double d[], int step)

{

 int i, k;
 double hold;

  for (i=0; i<=step-1; i++)
   {
    for(k=0; k<=step-2; k++)
     {
      if( d[k] > d[k+1])
        {
          hold = d[k];
          d[k] = d[k+1];
          d[k+1] = hold;
        }
     }
   }
```

```
}

/*THE PURPOSE OF THIS FUNCTION IS TO SELECT BOOTSTRAP SAMPLE AND PASS
IT
FOR THE PURPOSE OF DETERMINING BOOTSTRAP ESTIMATES*/

void bootstrap_sample(double a[], double b[], int step, int n)

{
  int   i, ind;
  unsigned int m;


      m = n;
      srand(m);
      for (i=0; i < step; i++)
        {
          ind = rand() % step;
          b[i] = a[ind];
        }

}

/*THIS IS USED TO GENERATE INCOME DATA FOR EACH SIMULATION PROCESS*/

void generate_data(double a[], int con, int seed)

{
  unsigned int m;
  int k;
  double u1, u2, st_norm, norm, mean, std;

  mean = 10.07;
  std = 0.70;
  m = seed;
  srand(m);
  for (k = 1; k <= con; k++)
    {
      u1 = (double) (1.0 + rand() % RANDMAX)/(1.0 + RANDMAX);
      u2 = (double) (1.0 + rand() % RANDMAX)/(1.0 + RANDMAX);
      st_norm = sqrt(-2*log(u1))*cos(2*3.14*u2);
      norm = mean + std*st_norm;
      a[k-1] = exp(norm);
    }
}
```

A.2 Analysis Program (S-plus)

```
# This program is an Splus program used to calculate model based small
# area rates of needs (health status) and their corresponding MSE
# estimates on the health status data for the RHAs in the province of
# Manitoba. Two model based approaches, VCM and model based on EL data,
# is applied in this program.

Tarea<-11
Tvariable<-6

X10<-matrix(0, 242, 7)

for(i in 1:242)
   {
     for(j in 4:9)
       {
         X10[i,j-3]<-dat.data[i,j]
       }
     for(j in 18:18)
       {
         X10[i,7]<-dat.data[i,j]
       }
   }
dimnames(X10)<-list(NULL, c("UNMR15N", "UNMR45n", "fparenn", "EDHp25N",
                               "LFWOMPN", "DWoosnn", "assoc."))

meanX10<-matrix(0, 11, 6)
for(i in 1:12)
  {
    for(j in 1:6)
      {
        if(i < 8)
           meanX10[i,j]<-mean(X10[X10[ ,7]==i,j])
        else if(i > 8)
           meanX10[i-1,j]<-mean(X10[X10[ ,7]==i,j])
      }
  }

nsize<-read("samplesize", T, 11)
meanofy<-read("mean_y", T, 11)

X11<-matrix(0, 242, 6)
for(i in 1:242)
  {
    for(j in 1:6)
      X11[i,j]<-X10[i,j]
  }

X12<-solve(t(X11)%*%X11)

newmat10<-matrix(0,6,6)
for(i in 1:11)
  {
    a<-matrix(meanX10[i,],6,1)
```

```
    b<-matrix(meanX10[i,],1,6)
    ab<-nsize[i]*(a%*%b)
    newmat10<-newmat10 + ab
  }

X13<-X12%*%newmat10

trace10<-0
for(i in 1:6)
   trace10<-trace10 + X13[i,i]

nstar10<-242 - trace10
ntotal10<-0
for(i in 1:11)
  ntotal10<-ntotal10 + nsize[i]

X14<-matrix(0, 242, 8)
for(i in 1:242)
  {
    for(j in 1:8)
      {
        if(j == 1)
          X14[i,j]<-dat.data[i,3]
        else
          X14[i,j]<-X10[i,j-1]
      }
  }

dimnames(X14)<-list(NULL, c("healf9", "UNMR15N", "UNMR45n", "fparenn",
"EDHP25N", "LFWOMPN", "DWoosdn", "assoc."))

X10dat.data<-data.frame(X14)

fit10.lm<-lm(healf9 ~ UNMR15N + UNMR45n + fparenn + EDHP25N + LFWOMPN +
                                        DWoosdn, data = X10dat.data)

sumres10<-0
for(i in 1:242)
   sumres10<-sumres10 + resid(fit10.lm)[i]^2

sigma01<- (1.0/(ntotal10 - Tarea - Tvariable + 1))*sumres10

sum10<-c(rep(0,11))
for(i in 1:11)
  {
    for(j in 1:6)
      sum10[i]<-sum10[i] + coefficients(fit10.lm)[j+1]*meanX10[i,j]
  }

Xdiff10<-matrix(0, 242, 8)
for(i in 1:242)
  {
     for(j in 1:1)
      {
       if( X14[i,8] < 8 )
         Xdiff10[i,j]<-X14[i,j] - meanofy[X14[i,8]]
       else if (X14[i, 8] == 8)
```

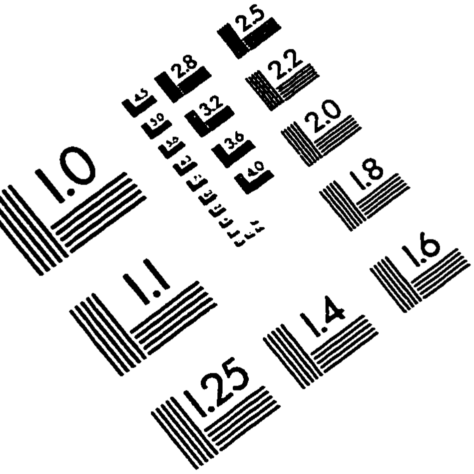
```
          Xdiff10[i,j]<-0
        else
          Xdiff10[i,j]<-X14[i,j] - meanofy[X14[i,8] - 1]
       }

      for(j in 2:7)
        {
         if(X14[i,8] < 8)
          Xdiff10[i,j]<-X14[i,j] - meanX10[X14[i,8],j-1]
         else if(X14[i,8] == 8)
          Xdiff10[i,j]<-0
         else
          Xdiff10[i,j]<-X14[i,j] - meanX10[X14[i,8] - 1,j-1]
        }
      for(j in 8:8)
        Xdiff10[i,j]<-X14[i,j]
  }

dimnames(Xdiff10)<-list(NULL, c("healf9dif", "UNMR15Ndif",
"UNMR45ndif","fparenndif", "EDHP25Ndif", "LFWOMPNdif", "Dwoosdndif",
"assoc."))

Xdiff10.data<-data.frame(Xdiff10)

fit11.lm<-lm(healf9dif ~ UNMR15Ndif + UNMR45ndif + fparenndif +
EDHP25Ndif + LFWOMPNdif + Dwoosdndif, data = Xdiff10.data)

sumres11<-0
for(i in 1:242)
  sumres11<-sumres11 + resid(fit11.lm)[i]^2

sigma02<-(nstar10 ^(-1))*(sumres11 - (ntotal10 - Tvariable)*sigma01)
if(sigma02 < 0)
  sigma02<-0

gamma10<-c(rep(0,11))
for(i in 1:11)
  {
    gamma10[i]<-sigma02/(sigma01/nsize[i] + sigma02)
  }

areamean10<-c(rep(0,11))
for(i in 1:11)
  {
    areamean10[i]<-sum10[i] + gamma10[i]*(meanofy[i] - sum10[i])
  }

count01<-0
count02<-1
Z10<-matrix(0, 241, 11)
for(j in 1:11)
  {
    count01<-count01 + nsize[j]
    for(i in count02:count01)
      {
        Z10[i,j]<-1
      }
```

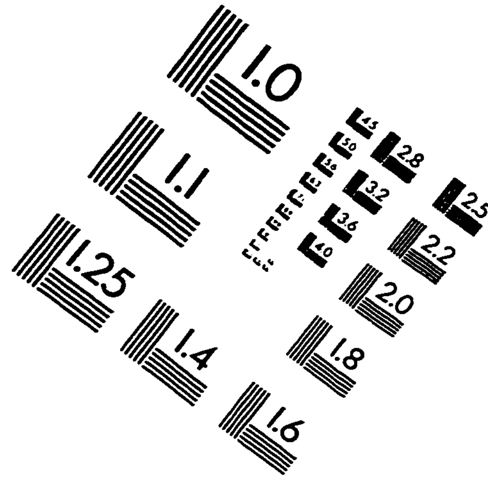
```
    count02<-count02 + nsize[j]
  }

G10<-sigma02*(Z10%*%t(Z10))
R10<-matrix(0, 241, 241)
for(i in 1:241)
  {
    R10[i,i]<-sigma01
  }

V10<-R10 + G10
Xtem10<-matrix(0, 241, 6)
for(j in 1:6)
  {
    for(i in 1:140)
      Xtem10[i,j]<-X10[i,j]
    for(i in 142:242)
      Xtem10[i-1,j]<-X10[i,j]
  }

XVX10<-solve(t(Xtem10)%*%V10%*%Xtem10)

newmat10<-matrix(0, 6, 6)
nsqtotal10<-0
for(i in 1:11)
  {
    a<-matrix(meanX10[i, ], 6, 1)
    b<-matrix(meanX10[i, ], 1, 6)
    ab<-(nsize[i]^3)*(a%*%b)
    newmat10<-newmat10 + ab
    nsqtotal10<-nsqtotal10 + (nsize[i]^2)
  }

newmat11<-solve(t(Xtem10)%*%Xtem10)%*%newmat10
trace11<-0
for(i in 1:6)
  trace11<-trace11 + newmat11[i,i]

ndbstar10<-nsqtotal10 - trace11

varsigma10<-2*(sigma01^2)*(1.0/(ntotal10 - Tarea - Tvariable + 1))
const01<-(1/(ntotal10 - Tvariable - Tarea +1))*(Tarea-1)^2*sigma01^2
const02<-2*nstar10*sigma01*sigma02 + ndbstar10*sigma02
varsigma11<-(2/nstar10^2)*(const01 + const02)
covar10<- -(Tarea-1)*(1.0/nstar10)*varsigma10

mse10<-c(rep(0,11))
for(i in 1:11)
  {
    g01<-(1 - gamma10[i])*sigma02
    g02<- (meanX10[i,] -
gamma10[i]*meanX10[i,])%*%XVX10%*%t(meanX10[i,] -
              gamma10[i]*meanX10[i,])
    const03<-(1.0/nsize[i]^2)*(sigma02 + sigma01/nsize[i])^(-3)
    const04<-(sigma01^2)*varsigma11 + (sigma02^2)*varsigma10 -
              2*sigma01*sigma02*covar10
    g03<-const03*const04
```

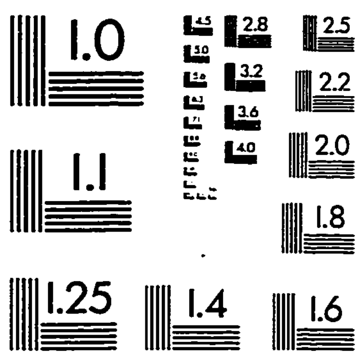
```
    mse10[i]<-g01 + g02 + 2*g03
  }

meanXY<-matrix(0, 11, 7)
for(i in 1:11)
  {
    for(j in 1:7)
      {
        if(j == 1)
          meanXY[i,j]<-meanofy[i]
        else
          meanXY[i,j]<-meanX10[i,j-1]
      }
  }

dimnames(meanXY)<-list(NULL, c("meanhealf9", "meanUNMR15N",
"meanUNMR45n", "meanfparenn", "meanEDHp25N", "meanLFWOMPN",
"meanDWoosnn"))

meanXY.data<-data.frame(meanXY)

fitmean.lm<-lm(meanhealf9 ~ meanUNMR15N + meanUNMR45n + meanfparenn +
           meanEDHp25N + meanLFWOMPN + meanDWoosnn, data = meanXY.data)

fayvar<-c(rep(0,11))
for(i in 1:12)
  {
    if(i < 8)
      fayvar[i]<-var(dat.data[dat.data[ ,18]==i,3])
    else if (i > 8)
      fayvar[i-1]<-var(dat.data[dat.data[ ,18]==i,3])
  }

faysum1<-0
faysum2<-0
XXinv<-t(meanX10)%*%meanX10
for(i in 1:11)
  {
    faysum1<-faysum1 + resid(fitmean.lm)[i]^2
    faysum2<-faysum2 + fayvar[i]*(1 -
meanX10[i,]%*%XXinv%*%t(meanX10[i,]))
  }

A<-(Tarea - 1)^(-1)*(faysum1 - faysum2)
if(A < 0)
  A<-0

Vfay<-matrix(0, 11, 11)
for(i in 1:11)
  Vfay[i,i]<-A + fayvar[i]


faybeta<-solve((t(meanX10)%*%solve(Vfay)%*%meanX10))%*%t(meanX10)%*%
                                           solve(Vfay)%*%t(meanofy)

faymean<-c(rep(0,11))
for(i in 1:11)
```

```
  {
    faymean[i]<-meanX10[i,]%*%faybeta +(A/(A + fayvar[i]))*(meanofy[i]
-
    meanX10[i,]%*%faybeta)
  }

faysum3<-0
faysum4<-0
for(i in 1:11)
  {
    faysum3<-faysum3 + 2*A*fayvar[i]/Tarea
    faysum4<-faysum4 + (fayvar[i]^2)/Tarea
  }


varA<-2*(Tarea^(-1))*(A^2 + faysum3 + faysum4)

faymse<-c(rep(0,11))
for(i in 1:11)
  {
    fayg1<-A*fayvar[i]*((A + fayvar[i])^(-1))
    c1<-(fayvar[i]^2)*((A + fayvar[i])^(-2))
    fayg2<-
c1*meanX10[i,]%*%solve(t(meanX10)%*%solve(Vfay)%*%meanX10)%*%t(meanX10[
i,])
    fayg3<-(fayvar[i]^2)*((A + fayvar[i])^(-3))*varA
    faymse[i]<-fayg1 + fayg2 + 2*fayg3
  }
```

IMAGE EVALUATION TEST TARGET (QA-3)

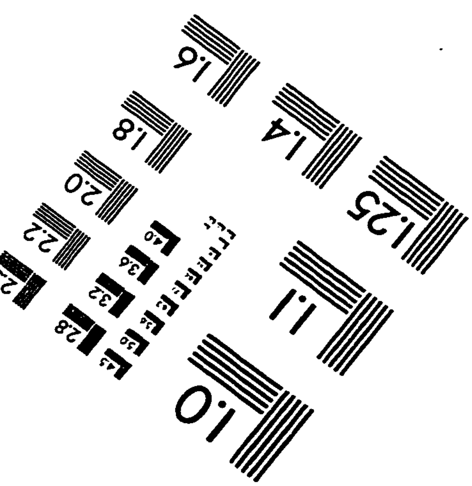






150mm

6″



APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989