NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents

ESTIMATION OF PARAMETERS OF SMALL DOMAINS IN FINITE POPULATIONS

PEISHUO CHEN

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science at
Concordia University
Montreal, Québec, Canada

September 1989

ISBN 0-315-51312-8

ABSTRACT

Estimation of Parameters of Small Domains In Finite Populations

Peishuo Chen

The estimation for small domains is to produce accurate estimator of small domain's characterics base on a sample taken from the large domain of study. The probiem arises: Limited resources dictated that the smaple should be relativly small, but representativenes requires that it should be widely spread. These two conditions combin to give small or zero sample size in a small domain. If we only use the smaple in a small domain to produce statistics, much information will be lost from the sample beyond the small domain. Then the variance of estimator for small domain is large because of the small sample size. But when estimator is obtained by a sample from whole area, this estimator for small domain will usually be design-biased. However, traditional probability sampling theory emphasizes that the estimator should be essentially design unbiased. We'll now give an approximately design-unbiased method. Let us consider the estimator under general regression model with random coefficients.

ACKNOWLEDEGEMENT

I would like to express my deepest gratitude to Dr. Y.P.Chaubey for his guidance, encouragement and support in the preparation of this thesis.

DEDICATION

I would like to dedicate this thesis to my wife Mrs.Zhiyin Deng and my parents Mr. and Mrs.Junliang Chen.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Estimation In Small Domains

A sample survey is usually designed within particular large geographical areas These geographical areas are usually called domain of study. Sometimes we are often faced with the problem of producing statistics for local small areas in character that are subdomains of the original domain of study. The subdomains are used to be called as small domains What we do is to produce accurate estimator of small domain's characteristics based on a sample taken from the large domain of study. The problem arises : Limited resources dictate that the sample should be relatively small, but representativeness requires that it should widely spread. These two conditions are combined to give small or even zero sample sizes in a small domain. If we only use the sample in small domain to produce the statistics, much information will be lost from the sample beyond the small domain, The variance of the estimation for such small domain in characteristics is large, because of the small sample size, Also no estimation will be obtained when sample size in the small domain is zero. But when estimation is obtained by a sample from a large area, this estimation for small domain will usually be design-biased. However, traditional probability sampling theory emphasizes that the estimator should be essentially design unbiased. We'll give anapproximately design-unbiased method. Usually, we assume that the characteristics y of population has model

$$y_k=\beta_0+\beta_1 x_{1k}+\ldots+\beta_p x_p \xi_k .$$

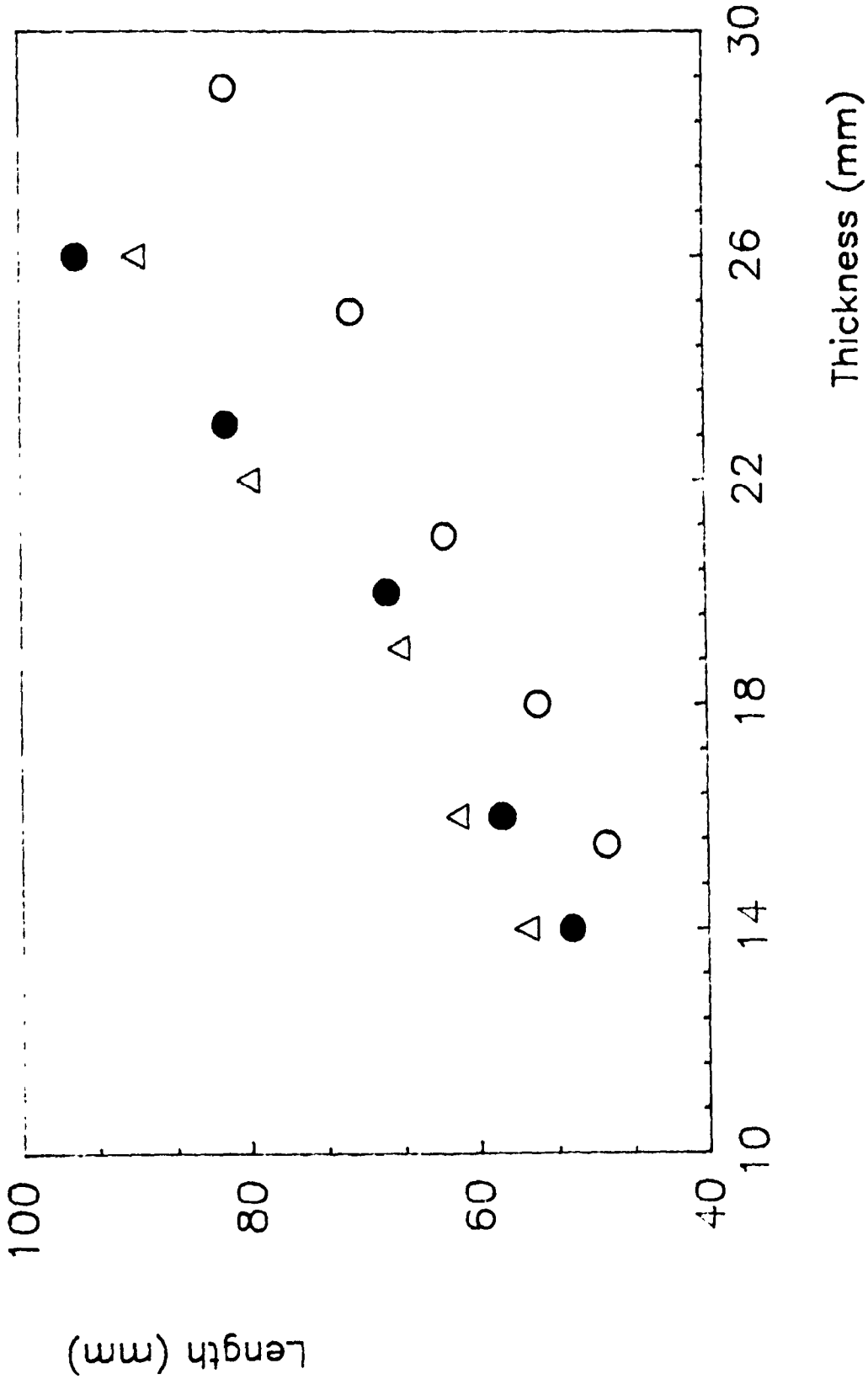


Fig.1 Length and thickness of three cucumbers

where $k \in$ domain of study, x_{pk}, $p=0,1,\cdots p$ are auxiliary variables In this thesis, we'll consider estimation for small domains, when (1) x_{pk} are random variables and (2) β_p are random variables. These cases are often encountered in practice.

For case (2), we shall introduce random regression coeffic.ents models by means of an example with data from a study of the relationship between the thickness and the length of cucumbers. In figure 1 the points represent observed values of length and thickness of three cucumbers for five consecutive days during the growing period. The different symbols represent different cucumbers.

It is seen that the points for each cucumber almost lie on a straght line Hence a straight line can be used to represent the relationship between thickness and length for a given cucumber. But each cucumber seems to have its own line. The cucumbers are chosen at random from a large number of cucumbers of a certain variety. Hence the regression lines must be considered as random. The individual cucumbers can be characterized by their straight-line relationships. To characterize the whole population of cucumbers , it is natural to look at the distribution of these lines.

1.2 Plan of The Thesis

In chapter 2 , we'll review some available techniques for estimation in small domains. Two concepts, design based approach and model based approach, will be presented first, then we'll outline two common estimators: ratio estimator and regression estimator. The synthetic estimator, an important method for estimating in small domains, also will

be introduced. In the last part of this chapter we consider the construction of small domain statistics as a prediction problem. Chapter 3, the details of generalized regression estimation under some superpopulation models will be discussed. Chapter 4, gives a numerical study of the methods discussed in this thesis.

1.3 Notation

Suppose that the finite population $U=\{ 1,2,\cdots k,\cdot\cdot N \}$ is divided into Q nonoverlapping domains, and the population also divided along the second dimension into H nonoverlapping categories (called groups), the population is cross classified into HQ cells.

$U_{\cdot q}$	Population of qth domain $q=1,2,\cdots Q$.
$U_{h\cdot}$	Population of hth group $h=1,2,\cdots H$.
U_{hq}	Population of qth group and hth domain.
$N_{\cdot q}$	Size of $U_{\cdot q}$.
$N_{h\cdot}$	Size of $U_{h\cdot}$.
N_{hq}	Size of U_{hq}.
s	A probability sample, of size n, given given by a sampling design $p(s)$.
$s_{\cdot q}$	A subset of sample s :$\{ k: k \in s$ and $k \in U_{\cdot q} \}$.
$s_{h\cdot}$	A subset of sample s :$\{ k: k \in s$ and $k \in U_{h\cdot} \}$.
s_{hq}	A subset of sample s :$\{ k: k \in s$ and $k \in U_{hq} \}$.
$\pi_k=p(k\epsilon s)$	Inclusion probability: the probability of kth population unit belongs to sample s.

$\pi_{kl}=p(k, l \in s)$ The probability of kth and qth population units belong to sample s.

$\bar{Y}= \frac{1}{N} \sum_{k\in U} y_k$ Population mean of the y-values.

$Y= N\bar{Y}$ Population total of the y-values.

$T_{\cdot q}=\sum_{k\in U_{\cdot q}} y_k$ qth domain total of the y-values.

$\sum_{U_{hq}} y_k, \sum_{s_{hq}} y_k$ Sums of the y-values over k in the indicated sets.

$\bar{y}=\frac{1}{n}\sum_{k\in s} y_k$ Sample mean of the y-values.

$\bar{X}=\frac{1}{N}\sum_{k\in U} x_k$ Population mean of the x-values.

$y=n\bar{y}$ Sample total of the y-values.

$\bar{x}=\frac{1}{n}\sum_{k\in s} x_k$ Sample mean of the x-values.

$x=n\bar{x}$ Sample total of the x-values.

ρ Correlation coefficient between x_i and y_i.

$S_y^2=\frac{1}{N-1}\cdot$ Population variance of y_k.

$\sum_{k\in U}(y_k-\bar{Y})^2$

$S_x^2=\frac{1}{N-1}\cdot$ population variance of x_k.

$\sum_{k\in U}(x_k-\bar{X})^2$

$S_{xy}=\rho S_x S_y$ Population covariance

$s_y^2 =\frac{1}{n-1}\cdot$ Sample variance of y_k.

$\sum_{k\epsilon s}(y_k-\bar{y})^2$

$s_x^2 = \frac{1}{n-1}\cdot$ Sample variance of x_k.

$\sum_{k\epsilon s}(x_k-\bar{x})^2$

$s_{xy}=\rho s_x s_y$ Sample covariance .

$C_{yy}=S_y^2/\bar{Y}^2$ Square of the coefficient of variation (cv) of y_k.

$C_{xx}=S_x^2/\bar{X}^2$ Square of the coefficient of variation (cv) of x_k.

$C_{xy}=S_{xy}/(\bar{X}\bar{Y})$ Relative covariance.

$f=n/N$ Sampling fraction.

E_M Model Expectation.

E_D Design Expectation.

V_M Variance under the model.

V_D Variance under the design.

V Variance of statistics.

1.4 Elementary Theorems (in SRSWOR)

SRSWOR stand for Simple Random Sample Without Replacement.

(1) $V_D(\bar{y}) = \frac{(1-f)}{n} S_y^2$

(2) $E_D(s_y^2) = S_y^2$

(3) $E_D(s_{xy}) = S_{xy}$

(4) $Cov(\bar{x},\bar{y}) = E(\bar{y}-\bar{Y})(\bar{x}-\bar{X})$

$$= \frac{(1-f)}{n} \frac{1}{n-1} \sum_{k\in s}(y_k-\bar{Y})(x_k-\bar{X})$$

$$= \frac{(1-f)}{n} \rho\, s_x\, s_y.$$

(5) Var $f(V,W) = E\ \mathrm{Var}[f(V,W)|W] + \mathrm{Var}\ E[f(V,W)|W]$

where V,W are random variables.

1.5 Introduction of Method of Estimation

One feature of theoretical statistics is the creation of a large body of theory that discusses how to make good estimators from data. In the development of theory, specifically for sample surveys, relatively little use has been made of this knowledge. Most of estimation methods in theoretical statistics assume that we know the functional form of the frequency distribution followed by the data in the sample, and the method of estimation is carefully geared to this type of distribution. The preference in sample survey theory has been to make, at most, limited assumptions about this frequency distribution (that it is very skew or rather symmetrical) and to leave its specific functional form out of the discussion.

Consequently, estimation techniques for sample survey work are at present restricted in scope. Some techniques will be considered in the following sections.

1.5.1 The Ratio Estimate

In the ratio method an auxiliary variate x_i, correlated with y_i, is obtained for each unit in the sample. The population total X of the x_i must be known.

The ratio estimate of Y, the population total of the y_i, is

$$Y_R = \frac{y}{x} X$$

$$= \frac{\bar{y}}{\bar{x}} X$$

where y,x are the sample totals of the y_i and x_i respectively.

If the quantity to be estimated is $\bar{Y}$, the population mean value of y_i, the ratio estimator is

$$\hat{\bar{Y}}_R = \frac{x}{y} \bar{X}$$

We now mention some results without proof.

Proposition 1.5.1

The ratio estimators of the population total ,Y, and the population mean, $\bar{Y}$, are, respectively,

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}} X \quad , \qquad \hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X}$$

In a simple random sample of size n (n large)

$$\mathbf{V}(\hat{Y}_R) \cong \frac{N^2(1-f)}{n} \left[\frac{\sum_{k \in U} (y_k - R x_k)^2}{N-1} \right] \qquad (1.1)$$

$$V(\hat{Y}_R) \cong \frac{1-f}{n}\left[\frac{\sum_{k\in U}(y_k - Rx_k)}{N-1}\right] \qquad (1.2)$$

where $R=X/Y$

Proposition 1.5.2

The estimated variance, $v(\hat{Y}_R)$, is given by

$$v(\hat{Y}_R) = \frac{N^2(1-f)}{n(n-1)} \sum_{k\in U}(y_k - \hat{R}x_k)^2$$

$$= \frac{N^2(1-f)}{n} (s_y^2 + \hat{R}\, s_x^2 - 2\hat{R}\, s_{xy})$$

where $\hat{R}=\bar{y}/\bar{x}$.

Remark 1.5.1

In general, the ratio estimate has a bias of order $1/n$. In practice, this quantity is usually unimportant in samples of moderate size.

Remark 1.5.2

The type of estimation of $Y = N\bar{y}$ where $\bar{y}$ is the mean per unit for the sample (in the simple random sampling) or a weighted mean per unit (in the stratified random sampling). Estimators of this kind are called estimators based on the *mean per unit* or estimators obtained by *simple expansion*.

Proposition 1.5.3

In the large samples, with simple random sampling, the ratio estimator $\hat{Y}_R$ has a smaller variance than the estimator $\hat{Y} = N\bar{y}$ obtained by simple expansion, if

$$\rho > \frac{1}{2} \frac{S_x}{\bar{X}} \Big/ \frac{S_y}{\bar{Y}}$$

$$= \frac{\text{coefficient of variation of } x_i}{2(\text{coefficient of variation of } y_i)}$$

$$= \frac{1}{2} \frac{cv(x)}{cv(y)}$$

We Will Now Specify Conditions Under Which The Ratio Estimator Is A Best Linear Unbiased Estimator

A well known result in regression theory indicates the type of population under which the ratio estimate may be called the best among a wide class of estimates. The result was first proved for infinite populations. Brewer(1963b) and Royall(1970a) extend the result to finite populations which holds if the following two conditions are satisfied.

1) The relation between y_i and x_i is a straight line though the origin.

2) The variance of y_i about this line is proportional to x_i.

A "best linear unbiased estimator" is defined as follow. Consider all estimators $\hat{Y}$ of Y that are linear function of the sample values y_i, They are of the form

$$l_1 y_1 + l_2 y_2 \cdots + l_n y_n$$

where the l's do not depend on the y_i, and may be functions of x's. The choice of l's is restricted to those that give unbiased estimation of Y. The estimator with the smallest variance is called the **best linear unbiased estimator** (BLUE).

Formally Brewer and Royall assume that the N population values (y_i, x_i) are a random sample from a superpopulation in which

$$y_i = \beta x_i + \xi_i \qquad (1.3)$$

where the ξ_i are independent of the x_i and $x_i > 0$. In arrays in which x_i is fixed, ξ_i has mean 0 and variance λx_i. The x_i $(i=1,2,\cdots N)$ are known.

The finite population total Y has been regarded as a fixed quantity. Under model (1.3), on the other hand, $Y=\beta X + \sum_{i=1}^{N} \xi_i$ is a random variable. In defining an unbiased estimator under this model, Brewer and Royall use a concept of unbiasedness which differs from that in randomization theory. They regard an estimator $\hat{Y}$ as unbiased if $\mathbb{E}(\hat{Y})=\mathbb{E}(Y)$ in repeated selections of the finite population **and** sample under the model. Such an estimator might be called *model-unbiased*. Thus under model (1.3) the ratio estimator $\hat{Y} = X\bar{y}/\bar{x}$ is the best linear unbiased estimator for any sample, random or not, selected solely according to the values of the x_i.

1.5.2 The Linear Regression Estimate

Like the ratio estimate, the linear regression estimate is designed to increase precision by the use of an auxiliary variate x_1 that is correlated with y_1. When the relation between y_1 and x_1 is examined, it may be found that, although the relation is approximate linear, the line does not go through the origin.

We suppose that y_i and x_i are each obtained for every unit in the sample and that the population mean $\bar{X}$ of the x_i is known. The linear regression estimate of $\bar{Y}$, the population mean of the y_i, is

$$\bar{y}_{lr}=\bar{y} + b(\bar{X}-x)$$

where the subscript *lr* denotes *linear regression* and b is an estimate of the change in y when x is increased by unity.

(A) Regression Estimates With Preassigned b

Let us now consider the following well known results

1) In simple random sampling, in which b_o is a preassigned constant, the linear regression estimate

$$\bar{y}_{lr}=\bar{y} + b_o\ (\bar{X} - \bar{x})$$

is unbiased, with variance

$$V(\bar{y}_{lr}) = \frac{1-f}{n} \frac{\sum_{k\in U}[(y_k-\bar{Y})-b_o(x_k-\bar{X})]^2}{N-1}$$

$$= \frac{1-f}{n}(S_y^2 - 2b_o S_{xy} + b_o^2 S_x^2)$$

2) The value of b_o that minimizes $V(y_{lr})$, is given by

$$b_o = B = \frac{S_{yx}}{S_x^2} = \frac{\sum_{k\in U}(y_k-\bar{Y})(x_k-\bar{X})}{\sum_{k\in U}(x-\bar{X})^2} \qquad (1.4)$$

and may be called the linear regression coefficient of y on x in the finite poppulation. Note that B does not depend on the properties of any sample that is drawn, and therefore could theoretically be preassigned. The resulting minimum variance is given by

$$V_{min}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2(1-\rho^2). \qquad (1.5)$$

where ρ is the population correlation coefficient between y and x.

(B) Regression Estimates When b Is Computed From The Sample

The equation (1.4) suggests that if b must be computed from the sample an effective estimate is likely to be the familiar least squares estimate of B, that is,

$$b = \frac{\sum_{k\in s}(y-\bar{y})(x-\bar{x})}{\sum_{k\in s}(x-\bar{x})^2} \qquad (1.6)$$

The theory of linear regression plays a prominent part in statistical methodology. The standard results of this theory are not entirely suitable for sample surveys because they require the assumptions that the population regression of y on x is linear, that the residual variance of y about the regression line is constant and that the

population is infinite.

We present an approach that makes no assumption of any specific relation between y and x .As in the analogous theory for the ratio estimate, only large-sample results are obtained.

With b as in (1.6),the linear regression estimator of Y in a simple random sample is

$$y_{lr} = \bar{y} + b(\bar{X}-x)$$
$$= \bar{y} - b(x-\bar{X}). \qquad (1.7)$$

Introduce the variate e_k defined by the relation

$$e_k = y_k - Y - B(x_k - X). \qquad (1.8)$$

Two properties of the e_k are

1) $\sum_{k\in U} e_k = 0$,

2) $\sum_{k\in U} e_k(x - \bar{X}) = 0$.

We now introduce some interesting results

Proposition 1.5.4

If b is the least squares estimate of B and

$$\bar{y}_{lr} = \bar{y} + b(\bar{X}-\bar{x})$$

then in simple random samples of size n,with n large,the variance of y_{lr} is given by

$$V(y_{lr}) \cong \frac{1-f}{n} S_y^2(1-\rho^2)$$

where $\rho = S_{yx}/S_y S_x$ is the population correlation coifficient between y and x.

Proposition 1.5.5

A sample estimate of $V(\bar{y}_{lr})$, valid in large samples, is given by

$$v(\bar{y}_{lr}) = \frac{1-f}{n(n-2)}\sum_{k\in s}[(y_k-\bar{y})-b(x_k-\bar{x})]^2$$

$$= \frac{1-f}{n(n-2)}\left\{\sum_{k\in s}(y_k-\bar{y})^2 - \frac{[\sum_{k\in s}(y_k-\bar{y})(x_k-\bar{x})]^2}{\sum_{k\in s}(x_k-\bar{x})^2}\right\}$$

the latter being the usual short-cut computing formula.

We now consider large-sample comparison with the ratio estimate and the mean per unit For these comparisons the sample size must be large enough so that the approximate formulas for the variances of the ratio and regression estimates are valid. The three comparable variances for the estimated population mean $\bar{Y}$ are as follows.

Proposition 1.5.6

$$V(\bar{y}_{lr}) = \frac{N-n}{Nn} S_y^2(1-\rho^2) \qquad \text{(regression)}$$

$$V(\bar{y}_R) = \frac{N-n}{Nn} (S_y^2+R^2S_x^2-2R\rho S_yS_x) \qquad \text{(ratio)}$$

$$V(\bar{y}) = \frac{N-n}{Nn} S_y^2 \qquad \text{(mean per unit)}$$

It is apparent that the variance of the regression estimate is smaller than that of the mean per unit unless $\rho=0$, in which case the two variances are equal.

The variance of the regression estimate is less than that of the ratio estimate if

$$-\rho^2S_y^2<R^2S_x^2-2R\rho S_yS_x .$$

This is equivalent to the inequalities

$$(\rho S_y-RS_x)^2>0 \qquad \text{or} \qquad (B-R)^2>0 .$$

Thus the regression estimate is more precise than the ratio estimate unless B=R. This occurs when the relation between y_k and x_k is a straight line through the origin.

1.5.3 Bias of The Linear Regression Estimate

We can show that the estimator $\bar{y}_{lr}$ has a bias of order $1/n$ in simple random sampling. We can write

$$\mathbb{E}(\bar{y}_{lr}) = \bar{Y} - \mathbb{E}b(\bar{x}-\bar{X}) ,$$

therefore for the bias is $-\mathbb{E}b(\bar{x}-\bar{X}) = -\mathrm{cov}(b, \bar{x})$. The leading term in the bias turns out to be

$$\frac{-(1-f)}{n} \frac{\mathbb{E}e_i(x_i-X)^2}{S_x^2}$$

This term represents a contribution from the *quadratic* component of the regression of y on $\bar{x}$. Thus, if a sample plot of y_k against x_k appears approximately linear, there should be little risk of major bias in $\bar{y}_{lr}$

1.5.4 The Linear Regression Estimator Under A Linear Regression Model

Suppose that the finite population values $y_k (k=1,2,\ldots N)$ are randomly drawn from an infinite superpopulation in which

$$y = \alpha + \beta x + \xi$$

where the ξ are independent, with means 0 and variance σ_ξ^2 for fixed x By

direct substitution from the model we find that

$$b = \frac{\sum_{k\in s} y_k (x_k - \bar{x})}{\sum_{k\in s} (x_k - \bar{x})^2}$$

$$= \beta + \frac{\sum_{k\in s} \xi_k (x_k - \bar{x})}{\sum_{k\in s} (x_k - \bar{x})^2} .$$

Hence
$$\bar{y}_{lr} - \bar{Y} = (\bar{\xi}_n - \bar{\xi}_N) + (\bar{X} - \bar{x}) \frac{\sum_{k\in s} \xi_k (x_k - \bar{x})}{\sum_{k\in s} (x_k - \bar{x})^2} \qquad (1.9)$$

where $\bar{\xi}_n$ and $\bar{\xi}_N$ are means over the sample and the finite population. It follows from(1 9) that under this model, $\mathbb{E}(\bar{y}_{lr} - \bar{Y}) = 0$, so, that $\bar{y}_{lr}$ is model-unbiased for any size of sample.

As regards the variance, it follows from(1.9), for a given set of x's,

$$\mathbb{V}(\bar{y}_{lr}) = \mathbb{E}(\bar{y}_{lr} - \bar{Y})^2$$

$$= \sigma_\xi^2 \left[\left(\frac{1}{n} - \frac{1}{N} \right) + \frac{(\bar{X} - \bar{x})^2}{\sum_{k\in s} (x_k - \bar{x})^2} \right]$$

This result holds for any n>1 and sample selected soley from the values of x. This approach and its generalization to the case of unequal residual variances were given by Royall (1970). Under this model a purposive sample plan that succeeded in making $\bar{x} = \bar{X}$ would minimize $\mathbb{V}(\bar{y}_{lr})$ for a given n.

Also, for any sample selected solely according to the values of the x, the usual least squares estimator

$$s_\xi^2 = \sum_{k\in s} [(y_k - \bar{y}) - b(x_k - \bar{x})]^2 / (n-2)$$

is a model-unbiased estimator of σ_ξ^2 for n>2.

Thus, in problems in which this model applies, simple exact results

about the mean and variance of $\bar{y}_{lr}$ can be established, values of the x_k, the random element being supplied gratis by the distribution of the ξ's assumed in the model.

CHAPTER 2

SOME AVAILABLE TECHNIQUES FOR SMALL DOMAIN ESTIMATIOM

2.1 Design Based And Model Based Appraoches

The essence of survey sampling consists of selection of a part of a finite collection of units, followed by giving of statements about the entire collection on the basis of the selected part. Two ways of having a finite collection of units are:

(1) The Fixed Population Approach:

With each population unit is associated a fixed but unknown real number, that is, the value of the variable y under study.

(2) The Superpopulation Approach:

With each population unit is associated a random variable for which a stochastic structure is specified; the actual value associated with a population unit is treated as the outcome of this random variable.

The fixed population approach is the traditional one in survey sampling. However, there are also early examples of the superpopulation approach: [see Cochran (1939, 1949), Deming and Stephen (1941)]. Many recent important contributions to finite population inference theory take the superpopulation approach. This is a very promising approach in survey sampling and may be meaningful in terms of changing population values with respect to the time.

As an example, suppose that the units are farms and that the characteristics under study is the yield of wheat in a given year. One approach to the inference problem, often used in standard texts, is the

following: It is assumed that to each farm in the population corresponds a fixed but unknown real number representing the yield of that particular farm in the particular year. When a farm has been selected for the sample, it is further more assumed that the yield is a fixed real number measured without error.

A different approach to the problem is to treat the yields of farms in the population as numbers generated under a stochastic model. Such models often incorporate auxiliary knowledge.

A crude but frequently effective model may simply postulate a linear stochastic relationship, for example, the yield of wheat apart from an error term of zero, expected value, proportional to size of the farm in acres, x_k, which is assumed known from a previous year. The model is,

$$y_k = \beta x_k + \xi_k \ ,$$

$$E_M (\xi_k | x_k) = 0 \ , \qquad k=1,2,\cdots N \ .$$

Moreover, the model considers that the unknown proportionality factor β is common to all farms. It would be determined in the particular year, by the average propensity of farms to devote acreage to wheat, by average yield per acre that year.

An estimate $\hat{\beta}$ of the unknown proportionality factor can be obtained from a sample of farms. For if any one farm is not in the sample, the value $\hat{\beta} x_k$ should provide an effective prediction of wheat yield, thereby permitting a prediction of total yield in the population.

In the Fixed Population Approach the variation of the estimator is entirely due to the sampling design chosen to select the units in the sample. However, in the super population approach it will also depend on the stochastic model generating the population, thus, the basis of

comparison of two estimations under the two approaches is $E[(\hat{\theta} - \theta)^2]$, the average MSE under the model, where $\hat{\theta}$ is an estimators of θ .We often strive for unbiased estimators in which case the criterion of comparison becomes the average variance.

2.2 Synthetic Approach

In many surveys, the target population is national in character, the primary aim being to produce accurate estimates of national characteristics. Limited resources dictate that the sample should be relatively small, but representativeness requires that it should be spread widely over the target population .These two conditions combine to give small or even zero sample size in certain subpopulation (called small domain), say province; yet increasingly, administrative decisions are based on data for small domain and several methods have been proposed. Among there is the method of synthetic estimation. This method was proposed in the U.S. Nation Center for Health Statistics (1968). Gonzalez (1973) described the method of synthetic estimation as follows:

An unbiased estimate is obtained from a sample survey for a large area. Then this estimate is used to derive estimates for subareas on the assumption that the small areas have the same characteristics as the large area, we identify these estimates as synthetic estimates.

Suppose that the finite population $U=\{1,2,...N\}$ is cross classified into HQ cells, Q domains and H groups. The domains may be numerous, say, small geographical areas of a sampled country. THe groups , may be not

more than ten, could be based on age, sex and so on.

$$N = \sum_{h=1}^{H} N_{h\cdot} = \sum_{q=1}^{Q} N_{\cdot q} = \sum_{h=1}^{H}\sum_{q=1}^{Q} N_{hq}$$

The N_{hq} are assumed known from a previous census or any other reliable source. This auxiliary information is the necessary prerequisite for significant improvement over the most basic estimators and is the essential ingredient in the synthetic techniques.

A probability sample s of size n, a subset of U, is drawn by a given sampling design $p(s)$ that determines the inclusion probabilities $\pi_k = p(k \epsilon s)$ and $\pi_{kl} = p(k, l \epsilon s)$. The parts of s that fall within $U_{\cdot q}$, $U_{h\cdot}$, and U_{hq} are denoted by $s_{\cdot q}$, $s_{h\cdot}$ and s_{hq}, respectively. Their respective sizes are denoted by $n_{\cdot q}$, $n_{h\cdot}$ and n_{hq}.

Associated with the kth population unit is the value y_k of the variable of character y. Having drawn s, we observe y_k for $k \in s$. The problem is to estimate the domain total $T_{\cdot q} = \sum_{k \in U_{\cdot q}} y$ for $q = 1, 2, \ldots Q$.

Notation such as $\sum_{U_{hq}} y_k$, $\sum_{s_{hq}} y_k$ and so on, denote sums over k in the indicated sets. I will write $\bar{y}_{hq}$ for $\sum_{U_{hq}} y_k / N_{hq}$, $\bar{y}_{h\cdot}$ for $\sum_{U_{h\cdot}} y / N_{h\cdot}$ and so on. For means at the sample level, an s is added: $\bar{y}_{s_{hq}}$ stands for $\sum_{s_{hq}} y_k / n_{hq}$, $\bar{y}_{s_{h\cdot}}$ for $\sum_{s_{h\cdot}} y_k / n_{h\cdot}$, and so on.

Often the survey is not or can not be designed to achieve highly efficient estimates for every one of numerous domains. Although n may be several thousand, there is often a lack of observations in given domain (*i.e.* $n_{\cdot q}$ is small) In practice, the n_{hq} may be zero for some

cells. Assume, however, that the sample sizes for the groups, $n_{h\cdot}$ are moderate to large.

We refer to simple random sampling (srs) with $\pi=n/N$ for all k, where n is fixed but the $n_{h\cdot}$ and $n_{\cdot q}$ are random.

The synthetic estimator of $T_{\cdot q}$ is obviously

$$\hat{T}_{syq} = \sum_{h=1}^{H} N_{hq} \bar{y}_{s_{h\cdot}} \qquad (2.1)$$

The term synthetic was used because these estimates were not derived directly from survey results. However, this particular method of borrowing information from similar small domains is to increase the accuracy of the resulting estimates.

Synthetic estimator has the outstanding advantage in small variance. But this estimator has the design bias $\sum_{h=1}^{H} N_{hq}(\bar{y}_{h\cdot} - \bar{y}_{hq})$ under srs or strs (stratified sampling).

The advantage of small variance of synthetic estimator is based on domains resemble each other. If the assumption is false, we can adjust the estimator for small domains by weights. The weights are based on the prior information.

Also proper bias is important. We shall see how to eliminate the effect of the bias. There is another estimator

$$\hat{T}_{MOq} = \sum_{h=1}^{H} [N_{hq}\bar{y}_{h\cdot} + n_{hq}(\bar{y}_{hq} - \bar{y}_{h\cdot})]$$

The estimator consists of synthetic estimator plus a correction term, $\sum_{h=1}^{H} n_{hq}(\bar{y}_{hq} - \bar{y}_{h\cdot})$, which works in reducing bias.

Särndal (1981) proposed the generalized regression estimator which

reduces to the following formula when the simple prediction approach is taken under the model that the mean in the hth group is the same for each domain,

$$\hat{T}_{GOq} = k\epsilon U[N_{hq}y_{s_{h\bullet}} + \hat{N}_{hq}(y_{s_{hq}} - y_{s_{h\bullet}})]$$

where $\hat{N}_{hq}=n_{hq}N_{h\bullet}/n_{h\bullet}$.

Särndal points that $\hat{T}_{SYq}$ and $\hat{T}_{MOq}$ should be similar in bias and mean squared error(**MSE**)properties for small samples and shows that the synthetic estimator has mean squared error advantage in small samples under the assumed model, however,with moderate to large samples,the generalized regression estimator gives smaller **MSE** under the deviations from the model. The latter also has advantage that estimators of variance and confidence intervals can be easily obtained.

Other estimation approaches have been proposed, including estimators as a weighted combination of a design-unbiased estimator and a design-biased but low variance estimator.Such attempts have included the use of shrinkage(or James-Stein) estimators. Some references in this direction are Schaible(1979),Fay and Herriott(1979),Battese and Fuller(1981),and Drew, Singh, and Choudry(1982). As Little (1983*a*,*b*) pointed out,one can construct by empirical Bayes methods a combined estimator such that shrinkage toward the design-biased component tends to zero in large samples.So the entire weight tends to be put on the design-consistent compoment.The resulting combined estimator will therefore be design consistent,but the weights are difinitely more complicated.With empirical Bayes and similar techniques, the weights are complex functions of the estimated population or model

variance. Consequently it is difficult to evaluate their design bias and design variance by analytical methods. Simplifying the weights, Drew ,Singh,and Choudry(1982) suggested a sample depend estimator whose performance,studied by simulation,seems promising.

2.3 The Prediction Approach

In the prediction approach, assumptions about population structure are made explicit by writing them in the form of a statistical model.Estimators are then derived from the model by using some criterion such as least squares.By using this approach all assumptions are quite explicit and thus open to challenge and to test using the data.This approach contrasts with the rather intuitive and ad hoc way in which various synthetic estimators have been proposed.

D.Holt,T.M.F.Smith and T.J.Tomberlin(1979) considered the following estimator.

The model is

$$y_{hqk} = \beta_h + \xi_{hqk} \quad , \qquad h = 1,2,\cdots H \quad q = 1,2,\cdots Q \quad k = 1,2,\cdots N_{hq} \qquad (2.2)$$

with $\mathbb{E}(\xi_{hqk}) = 0, \mathbb{V}(\xi_{hqk}) = \sigma^2$, and $\mathrm{Cov}(\xi_{hqk}, \xi_{hkl})=0$, when k, l are different.The population is cross classified by H groups and Q domains as above.This model is a simple one-way analysis of variance model with the hth group mean β_h the same for all of h.In any particular case,the appropriateness of this model may be examined in the usual ways,including the calculation of measures of fit and the use of other

diagnostic procedures such as plots of residuals.

The assumption of homoscedasticity is unnecessarily restrictive and can be replaced by different variances within groups or even a more general variance structure. The assumption of different variance in each group will not alter the form of the prediction estimator but will change its variance and also the estimator of variance. Laake(1979) has developed the prediction estimator and its variance under a more general variance structure.

Under the model (2.2), the **BLU** estimator of β_h is given by

$$\hat{\beta}_h = \bar{y}_{h\cdot\cdot}$$

$$= \Sigma_q \Sigma_{k\in s} y_{hqk} / \Sigma_q n_{hq} \quad ,$$

The **BLU** estimator of $\hat{T}_q$ is given by

$$\hat{T}_q = \Sigma_h \Sigma_{k\in s} y_{hqk} + \Sigma_h \Sigma_{k\notin s} \hat{\beta}_h$$

$$= \Sigma_h n_{hq} (\bar{y}_{hq\cdot} - \bar{y}_{h\cdot\cdot}) + \Sigma_h N_{hq} \bar{y}_{h\cdot\cdot}, \qquad (2.3)$$

When N_{hq} is much larger than n_{hq} ($N_{hq} \gg n_{hq}$), the second term in (2.3) will dominate and we obtain the synthetic estimator (2.2). The difference between the estimators is the result of the fact that the predictive estimators(2.3) recognizes that observed values are known exactly and so prediction is only required for the $(N_{hq}-n_{hq})$ unobserved values in each cell. Laake(1979) has shown that the variance of the prediction estimator(2.3) is smaller than that of the synthetic estimator(2.2) evaluated under the model (2.1). We note that in the limit when $n_{hq}=N_{hq}$ and the small domain total is known exactly, then (2.3) gives this value while (2.2) does not.

We can show under model (2.1) that the mean squared error of the

predictive estimator (2.3)is given by

$$\mathbf{MSE}(\hat{T}_q) = \sigma^2 \sum_h \frac{(N_{hq}-n_{hq})}{n_{\cdot h}} (N_{hq}-n_{hq}+n_{h\cdot})$$

$$\cong \sigma^2 \sum_h N_{hq}^2 / n_{h\cdot}, \qquad \text{for } N_{hq} \gg n_{hq}, \qquad (2.4)$$

where all expectations are taken over the distribution of the model error term β.We note again that if $n_{hq}=N_{hq}$,this becomes zero,as it should.

The variance σ^2 can be estimated from linear model theory by using the residual sum of squares.Under the model(2.1) we obtain

$$(n-H)\ \hat{\sigma}^2 = \sum_{k\in s} (y_k - y_{h\cdot})^2.$$

As we noted before,a different variance structure on the model would lead to a different expression for $\mathbf{MSE}(\hat{T}_q)$.Under suitable assumptions this modified $\mathbf{MSE}(\hat{T}_q)$ is still estimable using the theory for generalized least squares.

As we have indicated, any model is likely to be only an approximation of the actual situation.For this reason, we examine the behavior of the estimator(2.3),corresponding to the usual synthetic estimator,under alternative models of the population structure.For example, if the models of a constant mean for each group over all small domains were false, then a more appropriate model might be one that makes no such assumption; for example,

$$y_{hqk} = \mu_{hq} + \xi_{hqk} \qquad (2.5)$$

where $\mathrm{var}(\xi_{hqk}) = \sigma_{\cdot}^2$.The **BLU** estimator of μ_{hq} is the sample cell mean $\bar{y}_{hq\cdot}$ and we see that no information is borrowed from other cells

because the model does not allow for that. The estimator of T_q is $\hat{T}_q^{\bullet} = \sum_h N_{hq}\bar{y}_{hq\bullet}$, the poststratified estimator. If the estimator $\hat{T}_q$ (2.3), is used under model (2.4) instead of the **BLU** estimator $\hat{T}_q^{\bullet}$, then the bias of $\hat{T}_q$ is given by

$$\begin{aligned}
\text{Bias}(\hat{T}_q) &= \mathbb{E}(\hat{T}_q - T_q) \\
&= \mathbb{E}[\textstyle\sum_h \sum_{k\in s} y_{hqk} + \sum_h (N_{hq}-n_{hq})\bar{y}_{h\bullet\bullet} - \sum_h \sum_k y_{hqk}] \\
&= \textstyle\sum_h (N_{hq}-n_{hq}) \left[\dfrac{\sum_{q'} n_{hq'}\mu_{hq'}}{\sum_{q'} n_{hq'}} - \mu_{hq} \right].
\end{aligned}$$

$$\begin{aligned}
\mathbf{MSE}(\hat{T}_q) &= \mathbb{E}(\hat{T}_q - T_q)^2 \\
&= [\text{Bias}(\hat{T}_q)]^2 + \sigma_{\bullet}^2 \textstyle\sum_h \dfrac{(N_{hq}-n_{hq})}{n_{h\bullet}} (N_{hq}-n_{hq}+n_{h\bullet}).
\end{aligned}$$

Thus, the **MSE** is the sum of the square of the bias plus the **MSE** of the true model given by(2.4) but with σ^2 replaced by the true variance $\sigma_{\bullet}^2$. this result is true in general.

P. Laake published (1979) a closely related paper in predictive approach to small domain estimation. In this paper he derived optimal predictors under a superpopulation probability model.

2.4 The summary of synthetic approach and prediction approach

Generally speaking, synthetic estimators make use of auxiliary information. For any small domain characteristics of interest, the simplest synthetic estimator is obtained by applying the estimated

domain means directly to each some domain.The small domain estimators are then constructed by combining the domain estimators with known weights of the small domains. More elaborate synthetic estimators are obtained in the following way.With the domains of study,the sample units are partitioned (also called poststratified)into strata on the basis of information collected about them. The poststrata may be groups such as age and sex.Unbiased estimators are constructed for the actual characteristics within the poststrata.The synthetic estimators of the small domains are then obtained as a weighted sum of domain estimators,the relative sizes of the poststrata within the small domain.

As for predictive estimator,the small domain statistics is treated as a production problem.One of the advantages of the prediction approach is the fact that it may yield estimates of mean squared errors(MSE) as a measure of reliability.Thus, the approach may be an appropriate tool for comparison of different strategies for estimation within small domains.

In contrast to the prediction approach, synthetic estimators are constructed only on the basis of classical sampling procedures for finite populations.The estimators are usually biased,and there exists no good way to estimate the bias.The only evident way to compare the MSE of the estimators seems to be to establish a population in which the characteristics of the population are known,which has been done by Laake and Langva(1976) and Laake(1978).

Laake(1979) also showed in his paper that the predictors compare favorably with the synthetic estimators with respect to the MSE. The gain of the MSE is however,moderate.

2.5 Some Other Investigations

The synthetic estimator (**SYN**) has been traditionally used to estimate characteristics of small domains. Although it has the advantage of a small variance, it also has following disadvantages:

(a) it can be badly biased in some domains, and ordinarily we do not known which ones;

(b) consequently, a calculated coefficient of variation (cv), or a calculated confidence interval, is meaningless for such domains.

Särndal (1981) introduced the regression estimator (REG) in the context of domain estimation. This estimator is nearly unbiased and has additional advantage that a standard design based confidence interval is easily computed for each domain estimate.

However two drawbacks with REG are:

(a) the estimated variance can be unacceptably large in very small domains

(b) although with small probability, it can take negative values in situations where such values are unacceptable.

It is therefore desirable to strike a balance between SYN and REG. M. A. Hidiroglou and C. E. Särndal (1984) reported experiment with one such compromise estimator, the modified regression estimator (MRE). It's advantages are as follows

(a) It has a small (but noticeable) bias in those domains where the synthetic estimator is greatly biased; in other domains, the MRE is nearly unbiased. The MRE has the advantage of considerably reduced variance compared to the REG estimator.

(b) It has a smaller Mean Squared Error than the SYN estimator in

domains where the latter is badly biased.

(c) Meanihgful confidence intervals can also be easily constructed for the MRE estimator.

(1) Estimators

The following estimators of the qth domain total $T_q = \sum_{k\in U_{\bullet q}} y_k$ are proposed by M.A.Hidiroglou and C.E.Särndal (1984)

Two versions of the SYN,REG and MRE have been investigated,the "Count" versions and the "Ratio"version.

The formulas for the "Count"versions are:

Synthetic-Count estimator (SYN/C):

$$\hat{T}_{qSYN/C} = \sum_{h=1}^{H} N_{hq} \bar{y}_{s_{h\bullet}}$$

where $\bar{y}_{s_{h\bullet}}$ is the mean of y in $s_{h\bullet}$.

Regression-Count estimator (REG/C)

$$\hat{T}_{qREG/C} = \sum_{h=1}^{H} [N_{hq} y_{s_{h\bullet}} + \hat{N}_{hq} (\bar{y}_{s_{hq}} - \bar{y}_{s_{h\bullet}})]$$

where $\bar{y}_{s_{hq}}$ is the mean of y in s_{hq} , and $\hat{N}_{hq} = Nn_{s_{hq}}/n$. Here $\sum_{h=1}^{H} \hat{N}_{hq} (\bar{y}_{s_{hq}} - \bar{y}_{s_{h\bullet}})$ is a bias correction term that ordinarily carries a considerable variance contribution.

Modified Regression-Count estimator (MRE/C)

$$\hat{T}_{qMRE/C} = \sum_{h=1}^{H} [N_{hq} \bar{y}_{s_{h\bullet}} + F_q \hat{N}_{hq} (\bar{y}_{s_{hq}} - \bar{y}_{s_{h\bullet}})]$$

with

$$F_q = \begin{cases} E_q / n_{s_{\cdot q}} & \text{if } n_{s_{\cdot q}} \geq E_q \\ n_{s_{\cdot q}} / E_q & \text{if } n_{s_{\cdot q}} < E_q \end{cases}$$

where

$$E_q = E_{srs}(n_{s_{\cdot q}}) = nN_{\cdot q}/N$$

is the expected sample take, under simple random sample, from the qth domain.

The MRE/C estimator thus differs from the ordinary REG/C estimator in that the bias correction term revises a weight, F_q, which is bounded above by unity, and attains unity when the sample take equals its expectation. The theoretical justification for F_q is given in the paper of Hidiroglou and Särndal and is not presented here. Intuitively, the effect of F_q is to dampen the variance contributed by the correction term. The MRE/C estimator will have some bias, which is, however, ordinary much less than that of the SYN/C estimator.

The "Ratio" versions of the SYN, REG and MRE estimators are:

A) Synthetic-Ratio estimator (SYN/R):

$$\hat{T}_{qSYN/R} = \sum_{h=1}^{H} X_{hq} \hat{R}_h$$

with $X_{hq} = \sum_{k \in U_{hq}} x$ and $\hat{R}_h = \sum_{k \in s_{h\cdot}} y_k / \sum_{k \in s_{h\cdot}} x_k$.

B) Regression-Ratio estimator (REG/R):

$$\hat{T}_{qRFG/R} = \sum_{h=1}^{H} [X_{hq}\hat{R}_h + \hat{N}_{hq}(\bar{y}_{s_{hq}} \hat{R} \bar{x}_{s_{hq}})]$$

where $\bar{x}_{s_{hq}}$ is the mean of x in s_{hq}.

C) Modified Regression-Ratio estimator (MRE/R):

$$\hat{T}_{qMRE/R} = \sum_{h=1}^{H} [X_{hq}\hat{R}_h + F_q\hat{N}_{hq}(\bar{y}_{s_{hq}} - \hat{R}_h\bar{x}_{s_{hq}})]$$

where F_q is defined as in the MRE/C estimator above.

(2) Results From The Empirical Study

These results were supported by Mote Carlo study involving 500 samples.

The hypothesis expected to be verified was that the MRE estimator is situated, with respect to both bias and variance between the SYN and REG estimators, also, the part of MRE estimator a rather small bias and a substantial decrease in variance and Mean Squared Error as compared to the REG estimators. These hypotheses were indeed borne out by the empirical results.

For the Monte Carlo study reported in Dagum et al (1984) 500 samples had been drawn from a Nova Scotia population of N=1678 unincorated tax filers. The results are based on these same 500 samples. From these results, the following conclusions emerge:

(a) the SYN/C and SYN/R estimators are badly biased in some domains, namely, in those domains where the underlying model fits poorly. However, they consistently have an attractively low variance, compared to the other alternatives. The Mean Squared Error of the two SYN estimators will

consequently be very large in domains with large bias (poor model fit);By contrast, the Mean Squared Error is small in domains with little bias (good model fit)

(b) The REG/C and REG/R estimators are essentially unbiased. Their variance, although usually much lower than that of the EXP and POS estimators,(EXP for the straight expansion estimator, $\hat{T}_{qEXP} = \frac{N}{n} \sum_{k \in s_{\cdot q}} y_{k;}$ POS for the poststratified estimator $\hat{T}_{qPOS} = N_{\cdot q} \bar{y}_{s_{\cdot q}}$) is consistently much higher than that of the SYN/C and SYN/R estimators.

(c) The two MRE estimators,MRE/C and MRE/R, are negligibly biased when the SYN estimators happen to be nearly unbiased,otherwise the MRE estimators have a certain bias, which, however, is ordinarily much less pronounced than that of the SYN estimators.The MRE estimators have considerably smaller variance and Mean Squared Error,in all domains,than the REG estimators.This tendency is particularly pronounced in the smaller domains.In comparison with the SYN estimators,the MRE estimators(as expected) still have a large variance in virtually all domains.However, the Mean Squared Error of the MRE estimators is smaller than that of the SYN estimators in domains where the latter are badly biased.IN this study the MRE/R estimator has a smaller Mean Squared Error than that of the SYN/R in 9 out of 16 small areas.the obvious explanation is that in domains where the SYN estimator is greatly biased, the $(\text{bias})^2$ constitutes an extremely large contribution to the Mean Squared Error of the SYN, whereas for the MRE estimators,the $(\text{bias})^2$ is not very important.Since we do not know which domains create the large biases, the goal of producing reliable estimates in all domains is on the whole better served by the MRE method of estimation.

The MRE method presented here involves a simple mechanism for steering the estimates slightly in the direction of the stable STN estimators, when the sample taken is less than expected.This goal is also manifested in such other attempts as the empirical Bayes (Fay and Herriot,1979) and sample-dependent (Drew,Singh and Choudhry,1982) methods of estimation.

CHAPTER 3

GENERALIZED REGRESSION APPROACH

3.1 Generalized Regression Estimator

(1) Estimators

To virtually eliminate the major drawback of synthetic estimator, namely their design bias. The generalized regression estimator was proposed (Cassel, Särndal, and Wretman 1976, 1977, 1979; Sarndal 1980, 1984; related techniques in Isaki and Fuller 1982 and Wright 1983). For a general design, this estimator of T_q is

$$\hat{T}_{Gq} = \sum_{k\in U} {}_{\cdot q}\hat{y}_k + \sum_{k\in s} {}_{\cdot q} e_k/\pi_k, \tag{3.1}$$

where $e_k = y_k - \hat{y}_k$ is the residual and $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}$ the predicted y_k-value arising from a fit of a general model ξ, assuming that the y_k are independent and

$$\begin{aligned} \mathbb{E}_\xi (y_k) &= \mathbf{x}_k'\beta \\ &= \sum_{j=0}^{p} x_{kj}\beta_j \\ \mathbb{V}_\xi (y_k) &= v_k^{-1} \end{aligned} \tag{3.2}$$

For $k = 1, 2, \ldots N$, the vectors $\mathbf{x}_k' = (x_{ko}, \ldots x_{kp})$ are assumed to be known, and the v_k, known up to multipliers that cancel when $\hat{\mathbf{B}}$ in (3.3) is derived. The sample-based estimator $\hat{\mathbf{B}}$ is chosen as

$$\hat{\mathbf{B}} = (\sum_{k\in s} v_k \mathbf{x}_k \mathbf{x}_k'/\pi_k)^{-1} \sum_{k\in s} v_k \mathbf{x}_k y_k/\pi_k, \tag{3.3}$$

which is the usual survey sampling (approximately design-unbiased) estimator of its finite population counterpart

$$B = \left(\sum_{k\in U} v_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k\in U} v_k \mathbf{x}_k y_k, \qquad (3.4)$$

which, in turn, is the WLSQ (for Weighted LSQ)estimator of β that would be obtained if it were possible to carry out the census fit, that is, the fit of (3.2) to all population units. Thus (3.1) can be called a *design-model technique*, with emphasis on design. The specification of a design-model pair determines the form of $\hat{T}_{Gq}$, which is approximately design unbiased irrespective of model choice; Model choice is important to achieve low design variance.

If we specify the one-way model ξ_o such that for all $k \in U_{h\bullet}$,

$$E_{\xi_o}(y_k) = \beta_h$$

$$V_{\xi_o}(y_k) = \sigma_h^2 \qquad h=1,2,\cdots H$$

where y_k are assumed independent.

(3.1) becomes (with G for generalized regression, O for one-way)

$$\hat{T}_{GOq} = \sum [N_{hq} \bar{y}_{s_{h\bullet}} + \sum_{k\in s_{hq}} (y_k - \bar{y}_{s_{h\bullet}})/\pi_k]$$

where $\bar{y}_{s_{h\bullet}} = \hat{B}_h = (\sum_{k\in s_{h\bullet}} y_k/\pi_k)/(\sum_{k\in s_{h\bullet}} 1/\pi_k)$. Further, for the strs design with $\pi_k = n_{h\bullet}/N_{h\bullet}$ for $k \in U_{h\bullet}$. Thus with $\hat{N}_{hq} = n_{hq}N_{h\bullet}/n_{h\bullet}$,

$$\hat{T}_{GOq} = \sum [N_{hq} \bar{y}_{s_{h\bullet}} + \hat{N}_{hq}(\bar{y}_{s_{hq}} - \bar{y}_{s_{h\bullet}})] \qquad (3.5)$$

This estimator was considered in Särndal (1981).

(2) Design Variance Of $\hat{T}_q$ And Its Estimator

If $\hat{T}_q$ is a member of the generalized regression estimator family (3.1), an approximate design variance of $\hat{T}_q$ is given by Särndal(1984)

$$V_p(\hat{T}_q) = \sum_{k<l\in U}\sum \pi_{kl}\Delta_{kl}\left[\frac{c_{kq}e_k^*}{\pi_k} - \frac{c_{lq}e_l^*}{\pi_l}\right]^2 \qquad (3.6)$$

where $c_{kq} = 1$ if $k \in U_{\cdot q}$ and $c_{kq} = 0$ for all other k, $\Delta = (\pi_k\pi_l - \pi_{kl})/\pi_{kl}$, and $e_k^* = y_k - \mathbf{x}_k'\mathbf{B}$ [with $\mathbf{B}$ given by (3.4)] is the residual arising for unit k in the hypothetical census fit of the model (3.2). A Taylor expansion is used to obtain (3.6), which is the leading term of the exact design variance.

The sample-based estimator of $V_p(T_q)$ is given by the Yates-Grundy-type formula

$$\hat{V}_p(\hat{T}_q) = \sum_{k<l\in s}\sum \Delta_{kl}\left[\frac{c_{kq}e_k}{\pi_k} - \frac{c_{lq}e_l}{\pi_l}\right]^2$$

Here $e_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$, with $\hat{\mathbf{B}}$ given by (3.3), is the observed residual.

(3) A design-based confidence interval

For any given sampling design p, a design-based confidence interval is given at (approximately) level $100(1-\alpha)\%$ by

$$\hat{T}_q \pm z_{\alpha/2}\sqrt{\hat{V}_p(\hat{T}_q)} \qquad (3.7)$$

where $\hat{T}_q$ is (approximately) design unbiased for T_q, $\alpha/2$ is the unit normal deviate, and $\hat{V}_p(\hat{T}_q)$ is an (approximately) design-unbiased estimator of the theoretical variance of $\hat{T}_q$ under the sample design p. The interpretation of (3.7) is simple that in repeated draws of samples s by

the design p, the interval will cover the unknown population value T_q for roughly $100(1-\alpha)\%$ of all such samples.

3.2 Generalized Regression Estimator (GRE) Under Some Superpopulations And Its Proterties

Now we'll discuss the properties of generalized regression estimator under some superpopulations. The model is

$$y_k = \beta x_k + \xi_k \qquad k \in U \tag{3.8}$$

where ξ_k's are independent random variables

$$E_\xi(\xi_k | x_k) = 0 , \tag{3.9}$$

$$V_\xi(\xi_k | x_k) = \delta x_k^2 \tag{3.10}$$

and x_k's are independent random variables following

Case 1) Gamma distribution with probability density as given by

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases} \tag{3.11}$$

Case 2) General inverse Gaussian distribution with probability density as given by

$$f(x;\mu,\lambda) = \begin{cases} [\lambda/2\pi x^3]^{\frac{1}{2}} \exp\{-\lambda(x-\mu)^2/2\mu^2 x\}, & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases} \tag{3.12}$$

the above *pdf* will be denoted by $IG(\mu,\lambda)$.

The generalized regression estimator for qth domain total T_q is

$$\hat{T}_{Gq} = \sum_{k\in U_{\cdot q}} \hat{y}_k + \sum_{k\in s_{\cdot q}} e_k/\pi_k$$

where $e_k = y_k - \hat{y}_k$

$$\hat{y} = bx_k \quad ,$$

Here we have assumed that

(*i*) b is preassigned for β.

(*ii*) sample is obtained under a simple random sampling scheme.

We'll derive the average bias of $\hat{T}_{Gq}$: $E_M B(\hat{T}_{Gq})$ and its variance

$$Var(\hat{T}_{Gq}) = E_M Var(\hat{T}_{Gq}|y_1, y_2, \cdots y_N + Var_M E(\hat{T}_{Gq}|y_1, y_2, \cdots y_N)$$

$$= E_M V_D(\hat{T}_{Gq}) + V_M E_D(\hat{T}_{Gq})$$

where the subscript M stands for under model, and D for under design.

THEOREM 3.1

If the estimator $\hat{T}_{Gq}$(3.1) under the conditions (3.8), (3.9) and $E_M(x_k)=a$, where a is a parameter, then the average bias of $\hat{T}_{Gq}$, $E_M B(\hat{T}_{Gq})$, is zero.

Proof:

Let $c_k = \begin{cases} 1 & \text{for } k\in s_{\cdot q} \\ 0 & \text{for } k\notin s_{\cdot q} \end{cases}$

$$P(c_k = 1) = P(k\in s, k\in s_{\cdot q})$$

$$= P(k\in s_{\cdot q}|k\in s)\cdot P(k\in s)$$

$$= \frac{N_{\cdot q}}{N}\cdot\frac{n}{N}$$

$$= \frac{nN_{\cdot q}}{N^2}$$

$$\Rightarrow \quad \mathbb{E}_D c_k = \mathbb{E}_D c_k^2 = \frac{nN_{\cdot q}}{N^2} \tag{3.13}$$

$$\begin{aligned} B(\hat{T}_{Gq}) &= \mathbb{E}_D(\hat{T}_{Gq} - T_{\cdot q}) \\ &= \mathbb{E}_D(\textstyle\sum_{U_{\cdot q}} \hat{y}_k + \sum_{S_{\cdot q}} e_k/\pi_k - \sum_{U_{\cdot q}} y_k) \\ &= \mathbb{E}_D(\textstyle\sum_{S_{\cdot q}} e_k/\pi_k) - \sum_{U_{\cdot q}} e_k \\ &= \mathbb{E}_D(\textstyle\sum_U e_k c_k/\pi_k) - \sum_{U_{\cdot q}} e_k \\ &= \frac{N_{\cdot q}}{N} \textstyle\sum_U e_k - \sum_{U_{\cdot q}} e_k \end{aligned} \tag{3.14}$$

as $\quad y_k = \beta x_k + \xi_k \qquad k \in U \qquad (3.8)$

$$\mathbb{E}_\xi(\xi_k | x_k) = 0 , \tag{3.9}$$

and $\quad \mathbb{E}_M(x_k) = a$

so

$$\begin{aligned} \mathbb{E}_M B(\hat{T}_{Gq}) &= \mathbb{E}_M(\frac{N_{\cdot q}}{N} \textstyle\sum_U e_k - \sum_{U_{\cdot q}} e_k) \\ &= \frac{N_{\cdot q}}{N} \textstyle\sum_U \mathbb{E}_M[(\beta-b)x_k + \xi_k] - \sum_{U_{\cdot q}} [(\beta-b)x_k + \xi_k] \\ &= \frac{N_{\cdot q}}{N} [N(\beta-b)a] - N_{\cdot q}(\beta-b)a \\ &= 0 \end{aligned}$$

■

COROLLARY 3.2

The average bias of $\hat{T}_{Gq}$ is zero, under superpopulation with both cases, (3.11) and (3.12).

Proof:

The rth moment about the orign of the Gama distrition is given by

$$\mu_r' = \frac{\Gamma(\alpha+r)}{\Gamma(\alpha)} \tag{3.15}$$

We have

$$E_M(x_k) = \alpha \tag{3.16}$$

The first moment about origin of $IG(\mu,\lambda)$ is given below

$$\mu_1' = \mu$$

We have

$$E(x) = \mu$$

So both cases have form $E_M(x_k) = a$ .By theorem 1, the statement holds

■

THEOREM 3.3

The variance of $\hat{T}_{Gq}$ under Gamma model (3.11) is

$$\frac{NN_{\cdot q}}{n}\alpha(\alpha+1)[(\beta-b)^2+\delta] + \frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)N}{n(N-1)}(\beta-b)^2\alpha^2$$

$$- \frac{2N_{\cdot q}^2}{N}\alpha[(N\alpha+1)(\beta-b)^2+(\alpha+1)\delta]$$

$$+ N_{\cdot q}\alpha[(\beta-b)^2(N_{\cdot q}\alpha+1)+\delta(\alpha+1)+b^2]$$

$$+ \frac{N_{\cdot q}}{N}\alpha[\beta^2-b^2+\delta(\alpha+1)]$$

Proof

$$Var(\hat{T}_{Gq}) = E_M V_D(\hat{T}_{Gq}) + V_M E_D(\hat{T}_{Gq})$$

1) $E_D(\hat{T}_{Gq})$

$$E_D(\hat{T}_{Gq}) = E_D(\sum_{U\cdot q}\hat{y}_k + \sum_{S\cdot q} e_k/\pi_k)$$

$$= \sum_{U\cdot q} bx_k + E_D \frac{N}{n}\sum_{S\cdot q}(y_k-\hat{y}_k)$$

$$= \sum_{U\cdot q} bx_k + \frac{N}{n}\sum_{S\cdot q} E_D e_k$$

$$= \sum_{U\cdot q} bx_k + \frac{N}{n}\frac{nN_{\cdot q}}{N^2}\sum_U e_k$$

$$= \sum_{U\cdot q} bx_k + \frac{N_{\cdot q}}{N}\sum_U[(\beta-b)x_k+\xi_k]$$

2) $V_D(\hat{T}_{Gq})$

$$P(c_j, c_k \in s_{\cdot q}) = P(c_j \in s_{\cdot q})\cdot P(c_k \in s_{\cdot q} | c_j \in s_{\cdot q})$$

$$= \frac{nN_{\cdot q}}{N^2}\left(\frac{n-1}{N-1}\frac{N_{\cdot q}-1}{N-1}\right)$$

$$= \frac{n(n-1)N_{\cdot q}(N_{\cdot q}-1)}{N^2(N-1)^2}$$

$$\Rightarrow \quad E_D c_j c_k = \frac{n(n-1)N_{\cdot q}(N_{\cdot q}-1)}{N^2(N-1)^2} \qquad (3.17)$$

$$E_D(\sum_{S\cdot q} e_k)^2 = E_D(\sum_U e_k c_k)^2$$

$$= E_D(\sum_U e_k^2 c_k^2 + \sum_U c_j c_k e_j e_k)$$

$$= \frac{nN_{\cdot q}}{N^2}\sum_U e_k^2 + \frac{n(n-1)N_{\cdot q}^2(N_{\cdot q}-1)}{N^2(N-1)^2}\sum_{j\neq k U} e_j e_k \qquad (3.18)$$

$$V_D(\hat{T}_{Gq}) = E_D(\hat{T}_{Gq} - T_q)^2$$

$$= E_D(\sum_{S\cdot q} e_k/\pi_k - \sum_{U\cdot q} e)^2$$

$$= \frac{N^2}{n} \mathbb{E}_D(\sum_{s\cdot q} e_k)^2 - \frac{2N}{n} \sum_{U\cdot q} e_k \mathbb{E}_D(\sum_{s\cdot q} e_k) + (\sum_{U\cdot q} e_k)^2$$

$$= \frac{N_{\cdot q}}{n}\sum_U e_k^2 + \frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)}{n(N-1)^2} \sum_{j\neq k\,U} e_j e_k - \frac{2N_{\cdot q}}{N} (\sum_{U\cdot q} e_k)(\sum_U e_k)$$

$$+(\sum_{U\cdot q} e_k)^2 \qquad \text{[see (3.13), (3.18)]} \qquad (3.19)$$

3) $\mathbb{E}_M e_k$, $\mathbb{E}_M e_j e_k$ and $\mathbb{E}_M e_k^2$

From (3.15), (3.9) and (3.10) we have

$$\mathbb{E}_M(\xi_k) = 0 \qquad (3.20)$$

$$\mathbb{E}_M(x_k^2) = \alpha^2 + \alpha \qquad (3.21)$$

$$\mathbb{E}_M(\xi_k^2 x_k) = \delta(\alpha+1)\alpha \qquad (3.22)$$

(a) $\mathbb{E}_M e_k = \mathbb{E}_M(y_k - \hat{y}_k)$

$$= \mathbb{E}_M[(\beta-b)x_k + \xi_k]$$

$$= (\beta-b)\alpha \qquad (3.23)$$

(b) $\mathbb{E}_M e_k^2 = (\beta-b)^2 \mathbb{E}_M x_k^2 + \mathbb{E}_M \xi_k^2 + 2(\beta-b)\mathbb{E}_m x_k \xi_k$

$$=(\beta-b)^2(\alpha+1)\alpha + \delta(\alpha+1)\alpha$$

$$=\alpha(\alpha+1)[(\beta-b)^2 + \delta] \qquad (3.24)$$

(c) $\mathbb{E}_M e_k e_j = \mathbb{E}_M(y_k - \hat{y}_k)(y_j - \hat{y}_j)\ j \neq k$

$$= \mathbb{E}_M[(\beta-b)x_k + \xi_k][(\beta-b)x_j + \xi_j]$$

$$= \mathbb{E}_M[(\beta-b)^2 x_j x_k]$$

$$= (\beta-b)^2\alpha^2 \qquad (3.25)$$

4) Evaluation of other expressions

(a) $$\mathbb{E}_M\Big(\sum_{U\cdot q} e_k\Big)^2 = N_{\cdot q}\alpha(\alpha+1)[(\beta-b)^2+\delta] + \mathbb{E}_M\Big(\sum_{j\neq k\,U\cdot q} e_j e_k\Big)$$

$$= N_{\cdot q}\alpha[(\beta-b)^2(1+N_{\cdot q}\alpha) + (\alpha+1)\delta] \qquad (3.26)$$

(b) $$\mathbb{E}_M\Big(\sum_{U\cdot q} e_k\Big)\Big(\sum_U e_k\Big) = \mathbb{E}_M\Big[\Big(\sum_{U\cdot q} e_k\Big)^2 + \Big(\sum_{U\cdot q} e_k\Big)\Big(\sum_{k\notin U\cdot q} e_k\Big)\Big]$$

$$= N_{\cdot q}\alpha[(\beta-b)^2(1+N_{\cdot q}\alpha)+(\alpha+1)\delta]$$

$$+N_{\cdot q}(N-N_{\cdot q})(\beta-b)^2\alpha^2$$

$$= N_{\cdot q}\alpha[(N\alpha+1)(\beta-b)^2+(\alpha+1)\delta] \qquad (3.27)$$

(c) $$\mathbb{V}_M\Big(\sum_{U\cdot q} bx_k\Big) = b^2\sum_{U\cdot q}[\mathbb{E}(x_k)^2-(\mathbb{E}x_k)^2]$$

$$= b^2\sum_{U\cdot q}\left[\frac{\Gamma(2+\alpha)}{\Gamma(\alpha)} - \alpha^2\right]$$

$$= N_{\cdot q}b^2\alpha \qquad (3.28)$$

(d) $$\mathbb{V}_M\Big\{\frac{N_{\cdot q}}{N}\sum_U[(\beta-b)x_k+\xi_k]\Big\} = \frac{N_{\cdot q}^2}{N}(\beta-b)^2\alpha + \frac{N_{\cdot q}^2}{N}\delta(\alpha+1)\alpha$$

$$= \frac{N_{\cdot q}^2}{N}[(\beta-b)+\delta(\alpha+1)]\alpha \qquad (3.29)$$

(e) $$\mathrm{Cov}\Big\{\sum_{U\cdot q} bx_k,\ \frac{N_{\cdot q}}{N}\sum_U[(\beta-b)x_k+\xi_k]\Big\}$$

$$= \frac{bN_{\cdot q}}{N}(\beta-b)\ \mathrm{Cov}\Big(\sum_{U\cdot q} x_k,\ \sum_U x_k\Big)$$

$$= \frac{N_{\cdot q}}{N}(\beta-b)b\Big[\mathbb{E}_M\Big(\sum_{U\cdot q} x_k\cdot\sum_U x_k\Big) - \mathbb{E}_M\Big(\sum_{U\cdot q} x_k\Big)\cdot\mathbb{E}_M\Big(\sum_U x_k\Big)\Big]$$

$$= \frac{N_{\cdot q}}{N}(\beta-b)b\Big[\mathbb{E}_M\Big(\sum_{U\cdot q} x_k\Big)^2 + \mathbb{E}_M\Big(\sum_{U\cdot q} x_k\cdot\sum_{k\notin U\cdot q} x_k\Big) - N_{\cdot q}N\alpha^2\Big]$$

$$= \frac{N_{\cdot q}}{N}(\beta-b)b[N_{\cdot q}(\alpha+1)\alpha + N_{\cdot q}(N_{\cdot q}-1)\alpha^2 + N_{\cdot q}(N-N_{\cdot q})\alpha^2 - N_{\cdot q}N\alpha^2]$$

$$= \frac{N_{\cdot q}^2}{N} (\beta-b)b\alpha \qquad (3.30)$$

Now we have

$$\mathrm{Var}(\hat{T}_{Gq}) = \mathbb{E}_M \mathbb{V}_D(\hat{T}_{Gq}) + \mathbb{V}_M \mathbb{E}_D(\hat{T}_{Gq})$$

$$= \mathbb{E}_M\Big[\frac{N_{\cdot q}}{n} \sum_U e_k^2 + \frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)}{n(N-1)^2} \sum_{j\neq k\in U} e_j e_k$$

$$- \frac{2N_{\cdot q}}{N} (\sum_{U_{\cdot q}} e_k)(\sum_U e_k) + (\sum_{U_{\cdot q}} e_k)^2\Big]$$

$$+ \mathbb{V}_M\Big\{\sum_{U_{\cdot q}} bx_k + \frac{N_{\cdot q}}{N} \sum_U [(\beta-b)x_k + \xi_k]\Big\}$$

$$= \frac{N_{\cdot q}N}{n} \alpha(\alpha+1)[(\beta-b)^2+\delta] + \frac{(n-1)NN_{\cdot q}(N_{\cdot q}-1)}{n(N-1)}(\beta-b)^2\alpha^2$$

$$- \frac{2N_{\cdot q}}{N} \mathbb{E}_M[(\sum_{U_{\cdot q}} e_k)^2 + (\sum_{U_{\cdot q}} e_k)(\sum_{k\notin U_{\cdot q}} e_k)]$$

$$+ \mathbb{E}_M(\sum_{U_{\cdot q}} e_k)^2 + \mathbb{V}_m(\sum_{U_{\cdot q}} bx_k) + \mathbb{V}_M\Big\{\frac{N_{\cdot q}}{N}\sum_U [(\beta-b)x_k+\xi_k]\Big\}$$

$$+ 2\mathrm{Cov}\Big\{ \sum_{U_{\cdot q}} bx_k \,,\, \frac{N_{\cdot q}}{N} \sum_U [(\beta-b)x_k + \xi_k] \Big\}$$

$$= \frac{NN_{\cdot q}}{n}\alpha(\alpha+1)[(\beta-b)^2+\delta] + \frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)N}{n(N-1)}(\beta-b)^2\alpha^2$$

$$- \frac{2N_{\cdot q}^2}{N} \alpha[(N\alpha+1)(\beta-b)^2+(\alpha+1)\delta]$$

$$+ N_{\cdot q}\alpha[(\beta-b)^2(N_{\cdot q}\alpha+1)+\delta(\alpha+1)+b^2]$$

$$+ \frac{N_{\cdot q}}{N} \alpha[\beta^2 - b^2+\delta(\alpha+1)] \qquad (3.31)$$

If $b\cong\beta$ then

$$\mathrm{Var}(\hat{T}_{\cdot q}) \cong N_{\cdot q}\alpha(\alpha+1)[(\frac{N}{n} - \frac{2N_{\cdot q}}{N}\delta)\alpha + (\frac{N_{\cdot q}}{N}+1)\delta] + N_{\cdot q}\alpha b^2$$

(3.32)

THEOREM 3.4

The variance of $\hat{T}_{Gq}$ under inverse Gaussian model (3.12) is given by

$$\left(\frac{NN_{\cdot q}}{n} - \frac{2N_{\cdot q}}{N}\right)\left(\mu^2 + \frac{\mu^3}{\lambda}\right)[(\beta-b)^2+\delta]$$

$$+ \left[\frac{(n-1)NN_{\cdot q}(N_{\cdot q}-1)}{n(N-1)} + N_{\cdot q}(N_{\cdot q}-1) - \frac{2N_{\cdot q}^2(N-1)}{N}\right](\beta-b)^2\mu^2$$

$$+ \frac{\mu^3}{\lambda}\left[\frac{N_{\cdot q}^2}{N}(\beta-b)^2 + \frac{\delta N_{\cdot q}^2}{N} + \frac{bN_{\cdot q}^2}{N}(\beta-b) + b^2N_{\cdot q}\right] + \frac{\delta\mu^2N_{\cdot q}^2}{N}$$

Proof:

The first two moments about origin of $IG(\mu,\lambda)$ are given below

$$\mu_1' = \mu$$

$$\mu_2' = \mu^2 + \frac{\mu^3}{\lambda}$$

$$\Rightarrow \quad E_M(\xi_k) = 0$$

$$E_M(\xi_k | x_k) = \delta\left(\mu^2 + \frac{\mu^3}{\lambda}\right)$$

$$E_m(x_k) = \mu$$

$$E_m(x_k^2) = \left(\mu^2 + \frac{\mu^3}{\lambda}\right)$$

1. $E_M B(\hat{T}_{Gq})$

From (3.14) we have

$$E_M B(\hat{T}_{Gq}) = E_M\left(\frac{N_{\cdot q}}{N}\sum_U e_k - \sum_{U_{\cdot q}} e_k\right)$$

$$= \frac{N_{\cdot q}}{N}[N(\beta-b)\mu] - N_{\cdot q}(\beta-b)$$

$$= 0 \qquad (3.33)$$

So the average bias of $\hat{T}_{Gq}$ is zero, under superpopulation $IG(\mu,\lambda)$.

2. $\mathrm{Var}(\hat{T}_{Gq})$

1) $\mathbb{E}_M e_k = \mathbb{E}_M[(\beta-b)x_k+\xi_k]$

$$= (\beta-b)\mu \tag{3.34}$$

2) $\mathbb{E}_M e_k^2 = (\beta-b)^2\mathbb{E}_M x_k^2 + \mathbb{E}_M \xi_k^2 + 2(\beta-b)\mathbb{E}_M x_k \xi_k$

$$= (\beta-b)^2(\mu^2+\frac{\mu^3}{\lambda}) + \delta\mathbb{E}\, x^2$$

$$= (\mu^2 + \frac{\mu^3}{\lambda})[(\beta-b)^2+\delta] \tag{3.35}$$

3) $\mathbb{E}_M e_j e_k = \mathbb{E}_M[(\beta-b)x_j + \xi_j][(\beta-b)x_k + \xi_k]$

$$= \mathbb{E}_M[(b-b)^2 x_j x_k]$$

$$= (b-b)^2\mu^2 \tag{3.36}$$

4) $\mathrm{Cov}\left(\sum_{U\cdot q} x_k\ ,\ \sum_U x_k\right)$

$$= \mathbb{E}_M\left(\sum_{U\cdot q} x_k \cdot \sum_U x_k\right) - \mathbb{E}_M\left(\sum_{U\cdot q} x_k\right)\mathbb{E}_M\left(\sum_U x_k\right)$$

$$= N_{\cdot q}(\mu^2 + \frac{\mu^3}{\lambda}) + N_{\cdot q}(N_{\cdot q}-1)\mu^2 + N_{\cdot q}(N-N_{\cdot q})\mu^2 - N_{\cdot q}N\mu^2$$

$$= N_{\cdot q}\frac{\mu^3}{\lambda} \tag{3.37}$$

$$\mathrm{Var}(\hat{T}_{Gq}) = \mathbb{E}_M \mathbb{V}_D(\hat{T}_{Gq}) + \mathbb{V}_M \mathbb{E}_D(\hat{T}_{Gq})$$

$$= \mathbb{E}_M\Big[\frac{N_{\cdot q}}{n}\sum_U e_k^2 + \frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)}{n(N-1)}\sum_{j\neq k\,U} e_j e_k$$

$$- \frac{2N_{\cdot q}}{N}\left(\sum_{U\cdot q} e_k\right)\left(\sum_U e_k\right) + \left(\sum_{U\cdot q} e_k\right)^2\Big]$$

$$+ \mathbb{V}_M\Big\{\sum_{U\cdot q} bx_k + \frac{N_{\cdot q}}{N}\sum_U[(\beta-b)x_k + \xi_k]\Big\}$$

$$= \frac{N_{\cdot q}N}{n}(\mu^2 + \frac{\mu^3}{\lambda})[(\beta-b)^2+\delta] + \frac{(n-1)NN_{\cdot q}(N_{\cdot q}-1)}{n(N-1)}(\beta-b)^2\mu^2$$

$$- \frac{2N_{\cdot q}}{N}\{N_{\cdot q}(\mu^2 + \frac{\mu^3}{\lambda})[(\beta-b)^2+ \delta] + N_{\cdot q}(N_{\cdot q}-1)(\beta-b)^2\mu^2$$

$$+ N_{\cdot q}(N-N_{\cdot q})(\beta-b)^2\mu^2\} + N_{\cdot q}(\mu^2+ \frac{\mu^3}{\lambda})[(\beta-b)^2+\delta]$$

$$+ N_{\cdot q}(N_{\cdot q}-1)(\beta-b)^2\mu^2 + b^2N_{\cdot q}\frac{\mu^3}{\lambda} + \frac{\mu^3N^2_{\cdot q}}{\lambda N}(\beta-b)^2$$

$$+\frac{N^2_{\cdot q}}{N}\delta(\mu^2 + \frac{\mu^3}{\lambda}) + b\frac{\mu^3N^2_{\cdot q}}{\lambda N}(\beta-b)$$

$$= (\frac{NN_{\cdot q}}{n} - \frac{2N_{\cdot q}}{N})(\mu^2 + \frac{\mu^3}{\lambda})[(\beta-b)^2+\delta]$$

$$+ [\frac{(n-1)NN_{\cdot q}(N_{\cdot q}-1)}{n(N-1)} + N_{\cdot q}(N_{\cdot q}-1) - \frac{2N^2_{\cdot q}(N-1)}{N}](\beta-b)^2\mu^2$$

$$+ \frac{\mu^3}{\lambda}[\frac{N^2_{\cdot q}}{N}(\beta-b)^2 + \frac{\delta N^2_{\cdot q}}{N} + \frac{bN^2_{\cdot q}}{N}(\beta-b) + b^2N_{\cdot q}] + \frac{\delta\mu^2N^2_{\cdot q}}{N}$$

.....(3.38)

3.3 Generalized Regression Approach With Random Coefficients

3.3.1 Introduction

We shall take the example of section 1.1 as the introductory example in this section. The following general regression model is considered

$$y_{ih} = \beta_{0ih} + \beta_{1ih} x_{1ih} + \cdots + \beta_{pih} x_{pih} + \xi_{ih} ,$$

$$i = 1, 2, \cdots I, \qquad h = 1, 2, \ldots, H, \tag{3.39}$$

where the independent variables $\{x_{pih}\}$ are assumed to be fixed. The regression coefficients $\{\beta_{pih}\}$ are random with

$$\mathbb{E}(\beta_{pih}) = \theta_p \tag{3.40}$$

and

$$\mathrm{Cov}(\beta_{pih}, \beta_{rih}) = \lambda_{pr}. \tag{3.41}$$

Two β_{pih} with different values of the pair (i,h) are independent. We denote the covariance matrix of $(\beta_{0ih}, \ldots, \beta_{pih})$ by Λ. The error terms $\{\xi_{ih}\}$ are assumed to be independent with expectation 0 and variance σ^2. They are also assumed to be independent of the $\{\beta_{pih}\}$.

The introductory example corresponds to the special case P=1, I=5, G=3 and $\beta_{01h} = \cdots = \beta_{05h}$, $\beta_{11h} = \cdots = \beta_{15h}$ $(h=1,2,3)$.

Swamy(1970) considers the following model

$$y_{ih} = \beta_{0h} + \beta_{1h} x_{1ih} + \cdots + \beta_{ph} x_{pih} + \xi_{ih}$$

$$i = 1, \cdots I, \qquad h = 1, \cdots H. \tag{3.42}$$

This model will be applied in this paper.

With matrix notion Swamy's model(3.42) is written in the form

$$y_h = X_h\beta_h + \xi_h \qquad h=1,2,\cdots H \qquad (3.43)$$

where $y_h' = (y_{1h},\cdots y_{Ih})$, $\xi_h' = (\xi_{1h},\cdots \xi_{Ih})$, $\beta_h' = (\beta_{0h},\cdots \beta_{Ph})$ and X_h is the $I \times (P+1)$ matrix with ith row being $(1, x_{1ih},\cdots, x_{Pih})$. The vectors $y_1,\cdots y_H$ are independent with expectation $X_h\theta$, where $\theta' = (\theta_0,\theta_1,\cdots\theta_P)$, and covariance matrix $X_h\Lambda X_h' + \sigma^2 I$ where I is the $I \times I$ identity matrix.

Further we shall assume that both the $\{\beta_h\}$ and the $\{\xi_h\}$ have a normal distribution. So we have that y is $N(X\theta, X\Lambda X' + \sigma I)$ and

$$\hat{\beta}_h = (X_h'X_h)^{-1}X_h'y_h \qquad (3.44)$$

is $N(\theta, \Lambda + \sigma^2(X_h'X_h)^{-1})$.

$$\hat{\sigma}^2 = \frac{1}{H(I-P-1)} \sum_{h=1}^{H} Q(\hat{\beta}_h). \qquad (3.45)$$

where $Q(\hat{\beta}_h) = (y_h - X_h\beta_h)'(y_h - X_h\hat{\beta}_h)$ is the residual sum of squares when fitting the regression for the hth item. In this situation, however, it is difficult to find a best estimate of θ. If Λ and σ were known the generalized least squares estimate of θ would be

$$\theta^* = \left[\sum_{h=1}^{H} X_h'(X_h\Lambda X_h' + \sigma^2 I)^{-1}X_h\right]^{-1} \sum_{h=1}^{H} X_h'(X_h\Lambda X_h' + \sigma^2 I)^{-1}y_h \qquad \cdots (3.46)$$

with covairance matrix

$$\left[\sum_{h-1}^{H} X_h'(X_h\Lambda X_h' + \sigma^2 I)^{-1}X_h\right]^{-1}$$

$$= \left[\sum_{h-1}^{H} (\Lambda + \sigma^2(X_h'X_h)^{-1})^{-1}\right]^{-1}$$

An alternative and more straightforard estimate of θ is

$$\hat{\theta} = \frac{1}{H} \sum_{h=1}^{H} \hat{\beta}_h = \bar{\hat{\beta}} \qquad (3.47)$$

This estimate is also unbiased and has covariance matrix

Hence

$$\frac{1}{H}\Lambda + \frac{\sigma^2}{H^2}\sum_{h=1}^{H}(X_h'X_h)^{-1} \quad . \tag{3.48}$$

Hence

$$\hat{\Lambda} = \frac{1}{H-1}\sum_{h=1}^{H}(\hat{\beta}_h-\bar{\hat{\beta}})(\hat{\beta}_h-\bar{\hat{\beta}}) - \frac{\hat{\sigma}^2}{H}\sum_{h=1}^{H}(X_h'X_h)^{-1} \tag{3.49}$$

is an unbiased estimate of Λ. Swamy (1970) suggested to use the estimate (3.44) with the estimates $\hat{\sigma}^2$ and $\hat{\Lambda}$ inserted for σ^2 and Λ.

3.3.2. Generalized Regression Approach With One Regression Variable.

In this section, we'll discuss generalized regression approach with one regression variable. The estimate of qth domain total, $\hat{T}_{Gq}$,

$$\hat{T}_{Gq} = \sum_{U\cdot q}\hat{y}_k + \sum_{s\cdot q} e_k/\pi_k$$

where $e_k = y_k - \hat{y}_k$

$$y_k = \beta_{0k} + \beta_{1k}x_{1k} + \xi_k \quad k \in U$$

β_{0k}, β_{1k} and ξ_k are random variables with

$$E(\beta_{pk}) = \theta_p \qquad p = 0, 1.$$

$$\mathrm{Cov}(\beta_{pl}, \beta_{rk}) = \begin{cases} \lambda_{pr} & l, k \in \text{samegroup} \\ 0 & \text{otherwise} \end{cases}$$

$$E(\xi_k) = 0 \qquad k \in U$$

$$\mathrm{Cov}(\xi_l, \xi_k) = \begin{cases} \sigma^2 & l = k \\ 0 & \text{otherwise} \end{cases}$$

and ξ_k are independent of β_{pk}.

$$\hat{y}_k = \hat{\theta}_0 + \hat{\theta}_1 x_{1k}$$

define

$$\hat{\theta}' = (\hat{\theta}_0, \hat{\theta}_1)$$

$$= (\bar{\hat{\beta}}_0, \bar{\hat{\beta}}_1)$$

$$= \frac{1}{H} \sum_{h=1}^{H} \hat{\beta}'_h \qquad h\text{: hth group}$$

where $\hat{\beta}_h = (X'_h X_h)^{-1} X' y_h$

$$X_h = \begin{bmatrix} 1 & x_{11h} \\ 1 & x_{12h} \end{bmatrix} \qquad h = 1, 2, \cdots H$$

1. The average bias of $\hat{T}_{Gq}$: $E_M B(\hat{T}_{Gq})$

THEOREM 3.5

The average bias of $\hat{T}_{Gq}$, under above model, is zero.

Proof:

From (3 14)

$$B(\hat{T}_{Gq}) = \frac{N_{\cdot q}}{N} \sum_U e_k - \sum_{U_{\cdot q}} e_k$$

$$E_M B(\hat{T}_{Gq}) = E_M \{ \frac{N_{\cdot q}}{N} \sum_U [(\beta_{0k} - \hat{\theta}_0) + (\beta_{1k} - \hat{\theta}_1) x_{1k} + \xi_k]$$

$$- \sum_{\cdot q} [(\beta \quad -\hat{\theta}\ ) + (\beta \quad -\hat{\theta}\ )x \quad +\xi\ ]\}$$

$= 0$ because $\hat{\theta}$ is unbiased of β_h ■

2. Variance of $\hat{T}_{Gq}$: $\mathrm{Var}(\hat{T}_{Gq})$

THEOREM 3.6

The variance of $\hat{T}_{Gq}$, under above model, is

$$= \sum_{U \cdot q} \frac{1}{H} [(\lambda_{00}+\sigma^2 a_{00}) + x_{1k}(\lambda_{11}+\sigma^2 a_{11}) + 2x_{1k}(\lambda_{01}+\sigma^2 a_{01})]$$

$$+ \sum_{k \neq m \, U \cdot q} \frac{1}{H} [(\lambda_{00}+\sigma^2 a_{00}) + (x_{1k}+x_{1m})(\lambda_{01}+\sigma^2 a_{01}) + x_{1k}x_{1m}(\lambda_{11}+\sigma^2 a_{11})]$$

$$+ \left(\frac{N_{\cdot q}}{N}\right)^2 N\left[\left(1-\frac{1}{H}\right)(\lambda_{00}+\lambda_{01}+\lambda_{11}) + \frac{\sigma^2}{H}(a_{00}+a_{01}+a_{11})\right] + \frac{N_{\cdot q}^2}{N}\sigma^2$$

$$+ \left(\frac{N_{\cdot q}}{N}\right)^2 \{\sum_{h=1}^{H}(N_{h\cdot}-1)N_{h\cdot}\left[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01}) + \left(1-\frac{1}{H}\right)(\lambda_{00}+\lambda_{11}2\lambda_{01})\right]$$

$$+ [N(N-1) - \sum_{h=1}^{H}(N_{h\cdot}-1)N_{h\cdot}]\left[\frac{\sigma^2}{H}(a_{00}+a_{11}2a_{01}) - \frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})\right]$$

$$+ N(N-1)\sigma^2\} + \frac{2N_{\cdot q}}{N}\sum_{\substack{k \in U \cdot q \\ m \in U}} \{\frac{1}{H}[\lambda_{00} + (x_{1k}+x_{1m})\lambda_{01} + x_{1k}x_{1m}\lambda_{11}]$$

$$- [\frac{1}{H}(\lambda_{00}+\sigma^2 a_{00}) + \frac{1}{H}(\lambda_{01}+\sigma^2 a_{01})(x_{1k}+x_{1m}) + \frac{1}{H}(\lambda_{11}+\sigma^2 a_{11})x_{1k}x_{1m}]\}$$

$$+\frac{NN_{\cdot q}}{n}[(1-\frac{1}{H})(\lambda_{00}+\lambda_{01}+\lambda_{11})+\frac{\sigma^2}{H}(a_{00}+a_{01}+a_{11})]+\frac{NN_{\cdot q}\sigma^2}{n}$$

$$+\frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)}{n(N-1)^2}\{\sum_{h=1}^{H}(N_{h\cdot}-1)N_{h\cdot}[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})$$

$$+(1-\frac{1}{H})(\lambda_{00}+\lambda_{11}+2\lambda_{01})]+[N(N-1)-\sum_{h=1}^{H}(N_{h\cdot}-1)N_{h\cdot}][\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})$$

$$-\frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})]+N(N-1)\sigma^2\}-\frac{2N_{\cdot q}}{N}\{\sum_{h=1}^{H}N_{hq}(N_{h\cdot}-N_{hq})$$

$$\cdot[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})+(1-\frac{1}{H})(\lambda_{00}+\lambda_{11}+2\lambda_{01})]+[N_{\cdot q}(N-N_{\cdot q})$$

$$-\sum_{h=1}^{H}N_{hq}(N_{h\cdot}-N_{hq})][\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})-\frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})]$$

$$+N_{\cdot q}(N-N_{\cdot q}\sigma^2\}+(1-\frac{2N_{\cdot q}}{N})\{(N_{\cdot q}-\frac{N_{\cdot q}}{H})(\lambda_{00}+\lambda_{11}+\lambda_{01})$$

$$+\frac{N_{\cdot q}\sigma^2}{H}(a_{00}+a_{01}+a_{11})+N_{\cdot q}\sigma^2+\sum_{h=1}^{H}(N_{hk}-1)N_{hk}[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})$$

$$+(1-\frac{1}{H})(\lambda_{00}+\lambda_{11}2\lambda_{01})]+[N_{\cdot q}(N_{\cdot q}-1)-\sum_{h=1}^{H}(N_{hk}-1)N_{hk}]$$

$$\cdot[\frac{\sigma^2}{H}(a_{00}+a_{11}2a_{01})-\frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})]+N_{\cdot q}(N_{\cdot q}-1)\sigma^2\}$$

$$\cdots(3.50)$$

Proof:

$$\mathrm{Var}(\hat{T}_{GQ}) = \mathbb{E}_M\mathbb{V}_D(\hat{T}_{GQ}) + \mathbb{V}_M\mathbb{E}_D(\hat{T}_{GQ})$$

1) $$\mathbb{E}_D(\hat{T}_{GQ}) = \mathbb{E}_D(\sum_{U\cdot q}\hat{y}_k + \sum_{S\cdot q}e_k/\pi_k)$$

$$= \sum_{U\cdot q}\hat{y}_k + \frac{N_{\cdot q}}{N}\sum_U e_k \qquad (3.51)$$

2) $$\mathbb{V}_D(\hat{T}_{GQ})$$

From (3.22) we have

$$V_D(\hat{T}_{GQ}) = \frac{N_{\cdot q}}{n}\sum_U e_k^2 + \frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)}{n(N-1)^2}\sum_U e_k e_m$$

$$- \frac{2N_{\cdot q}}{N}(\sum_{U_{\cdot q}} e_k)(\sum_U e_k) + (\sum_{U_{\cdot q}} e_k)^2$$

$$= \frac{N_{\cdot q}}{n}\sum_U e_k^2 + \frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)}{n(N-1)^2}\sum_{k\neq m\,U} e_k e_m$$

$$- \frac{2N_{\cdot q}}{N}(\sum_{\substack{m\notin U_{\cdot q}\\ k\in U_{\cdot q}}} e_k e_m) + (1-\frac{2N_{\cdot q}}{N})(\sum_{U_{\cdot q}} e_k^2 + \sum_{k\neq m\,U_{\cdot q}} e_k e_m)$$

$\cdots(3.52)$

3) $E_M e_k$, $E_M e_k$ and $E_M e_k e_m$

(a) $E_M e_k$

$$E_M e_k = E_M(y_k - \hat{y}_k)$$

$$= E_M[\sum_{p=0}^{P}(\beta_{pk} - \hat{\theta}_p)x_{pk} + \xi_k]$$

$$= 0 \qquad (3.53)$$

note, let $x_{0k}=1$; $\hat{\theta}$ is unbiased of β_h.

(b) $E_M e_k^2$ and $E_M e_k e_m$

We have $\mathrm{Cov}(\beta_{ph}, \beta_{rh}) = \lambda_{pr}$ $\qquad p, r = 0, 1, \cdots P$

$$\Lambda = (\lambda_{pr})_{P+1, P+1}$$

Let $\mathrm{Cov}(\hat{\theta}_p, \hat{\theta}_r) = \lambda^*_{pr}$ $\qquad (3.54)$

$$\Lambda^* = (\lambda^*_{pr})$$

$$= \frac{1}{H}\Lambda + \frac{\sigma^2}{H^2}\sum_{h=1}^{H}(X_h'X_h)^{-1} \qquad \text{from (3.48)}$$

Let $$A = \frac{1}{H}\sum_{h=1}^{H}(X_h'X_h)^{-1} = (a_{pr})_{P+1,P+1} \qquad (3.55)$$

so $$\lambda_{pr}^{*} = \frac{1}{H}(\lambda_{pr} + \sigma^2 a_{pr}) \qquad (3.56)$$

Now we have P = 1 and let I = 2 , then

$$X_h = \begin{pmatrix} 1 & x_{1h} \\ 1 & x_{2h} \end{pmatrix}$$

$$X_h'X_h = \begin{pmatrix} 1 & x_{1h}+x_{2h} \\ x_{1h}+x_{2h} & x_{1h}^2+x_{2h}^2 \end{pmatrix}$$

$$(X_hX_h)^{-1} = \frac{-1}{2x_{1h}x_{2h}} \begin{pmatrix} x_{1h}^2+x_{2h}^2 & -x_{1h}-x_{2h} \\ -x_{1h}-x_{2h} & 1 \end{pmatrix} \qquad (3.57)$$

$$A = \frac{1}{H}\begin{pmatrix} \sum_{h=1}^{H}[(x_{1h}^2+x_{2h}^2)/2x_{1h}x_{2h}] & \sum_{h=1}[(x_{1h}+x_{2h})/2x_{1h}x_{2h}] \\ \sum_{h=1}[(x_{1h}+x_{2h})/2x_{1h}x_{2h}] & 1 \end{pmatrix}$$

$\cdots(3.58)$

From (3.41) we have

$$E(\beta_{ph}\cdot\beta_{rh}) = \lambda_{pr} + \theta_p\theta_r \qquad (3.59)$$

$$E(\beta_{ph}\cdot\beta_{rl}) = \theta_p\theta_r \qquad \text{for } h \neq l \qquad (3.60)$$

$$E(\hat{\theta}_p\cdot\hat{\theta}_r) = \lambda_{pr}^{*} + \theta_p\theta_r \qquad (3.61)$$

From (3.44) and (3.43)

$$\hat{\beta}_h = (X_h'X_h)^{-1}X_h'y_h$$

$$= (X_h'X_h)^{-1}X_h'(X_h\beta_h+\xi_h)$$

$$= \beta_h + (X_h'X_h)^{-1}X_h'\xi_h$$

$$\Rightarrow \quad \hat{\theta}' = \frac{1}{H}\sum_{h=1}^{H}\beta_h + \frac{1}{H}\sum_{h=1}^{H}(X_h'X_h)^{-1}X_h\xi_h \qquad \text{from (3.47)}$$

We have $E_M(\xi_k)=0$ and $Cov(\xi_{k'},\beta_{pk})=0$,then

$$E_M(\beta_{pk}\cdot\hat{\theta}_r) = E_M(\frac{1}{H}\beta_{pk}\sum_{h=1}^{H}\beta_{rh})$$

$$= \frac{1}{H}\lambda_{pr} + \theta_p\theta_r \qquad (3.62)$$

$$E_M e_k^2 = E_M(y_k-\hat{y}_k)^2$$

$$= E_M[(\beta_{0k}-\hat{\theta}_0)+(\beta_{1k}-\hat{\theta}_1)x_{1k}+\xi_k]^2$$

$$= E_M[(\beta_{0k}-\hat{\theta}_0)^2+(\beta_{1k}-\hat{\theta}_1)^2+(\beta_{0k}-\hat{\theta}_0)(\beta_{1k}-\hat{\theta}_1)x_{1k}] + E_M\xi_k^2$$

$$= (\lambda_{00}+\theta_0^2) - 2(\frac{1}{H}\lambda_{00}+\theta_0^2) + (\lambda_{00}^{*}+\theta_0^2) +$$

$$(\lambda_{11}+\theta_1^2) - 2(\frac{1}{H}\lambda_{11}+\theta_1^2) + (\lambda_{11}^{*}+\theta_1^2) +$$

$$(\lambda_{01}+\theta_0\theta_1) + (\lambda_{01}^{*}+\theta_0\theta_1) - 2(\frac{1}{H}\lambda_{01}+\theta_0\theta_1)+\sigma^2$$

$$= \lambda_{00} - \frac{2}{H}\lambda_{00} + \frac{1}{H}(\lambda_{00}+\sigma^2 a_{00}) + \lambda_{11} - \frac{2}{H}\lambda_{11} + \frac{1}{H}(\lambda_{11}+\sigma^2 a_{11}) +$$

$$\lambda_{01} + \frac{1}{H}(\lambda_{01}+\sigma^2 a_{01}) - \frac{2}{H}\lambda_{01} + \sigma^2$$

$$= (1-\frac{1}{H})(\lambda_{00}+\lambda_{01}+\lambda_{11}) + \frac{\sigma^2}{H}(a_{00}+a_{01}+a_{11}) + \sigma^2 \qquad (3.63)$$

$$E_M e_k e_m = E_M(y_k-\hat{y}_k)(y_m-\hat{y}_m)$$

$$= E_M[(\beta_{0k}-\hat{\theta}_0)+(\beta_{1k}-\hat{\theta}_1)x_{1k}+\xi_k][(\beta_{0m}-\hat{\theta}_0)+(\beta_{1m}-\hat{\theta}_1)x_{1m}+\xi_m]$$

$$= E_M[(\beta_{0k}\beta_{0m}+\hat{\theta}_0^2-\beta_{0k}\hat{\theta}_0-\beta_{0m}\hat{\theta}_0)+(\beta_{0k}\beta_{1m}+\hat{\theta}_0\hat{\theta}_1-\beta_{0k}\hat{\theta}_1-\beta_{1m}\hat{\theta}_0)+$$

$$(\beta_{1k}\beta_{1m}+\hat{\theta}_1^2-\beta_{1k}\hat{\theta}_1-\beta_{1m}\hat{\theta}_1)+(\beta_{1k}\beta_{0m}+\hat{\theta}_0\hat{\theta}_1-\beta_{1k}\hat{\theta}_0-\beta_{0m}\hat{\theta}_1)]+\sigma^2$$

From (3.58),(3.59),(3.60) and (3.61)

$$=\begin{cases}\lambda^*_{00}+\lambda^*_{11}+2\lambda^*_{01}-\frac{2}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})+\sigma^2 & \text{if } k,m \notin \text{samegroup}\\ \lambda^*_{00}+\lambda^*_{11}+2\lambda^*_{01}+(1-\frac{2}{H})(\lambda_{00}+\lambda_{11}+2\lambda_{01})+\sigma^2 & \text{otherwise}\end{cases}$$

$$=\begin{cases}\frac{\sigma}{H}(a_{00}+a_{11}+2a_{01})-\frac{}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})+\sigma^2 & \text{if } k,m \notin \text{samegroup}\\ \frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})+(1-\frac{1}{H})(\lambda_{00}+\lambda_{11}+2\lambda_{01})+\sigma^2 & \text{otherwise}\end{cases}$$

$\cdots$(3.63)

Because of $\mathbb{E}_M e_k = 0$, then

$$\mathbf{V}_M e_k = \mathbb{E}_M e_k^2$$

$$=(1-\frac{1}{H})(\lambda_{00}+\lambda_{01}+\lambda_{11})+\frac{\sigma^2}{H}(a_{00}+a_{01}+a_{11})+\sigma^2 \qquad (3.64)$$

$$\mathrm{Cov}(e_k,e_m) = \mathbb{E}_M e_k e_m$$

$$=\begin{cases}\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})-\frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})+\sigma^2 & \text{if } k,m \notin \text{samegroup}\\ \frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})+(1-\frac{1}{H})(\lambda_{00}+\lambda_{11}+2\lambda_{01})+\sigma^2 & \text{otherwise}\end{cases}$$

4) $\mathbf{V}_M(\hat{y}_k)$, $\mathrm{Cov}(\hat{y}_k,\hat{y}_m)$ and $\mathrm{Cov}(\hat{y}_k,e_m)$

(a) $\mathbf{V}_M(\hat{y}_k) = \mathbf{V}_M(\hat{\theta}_0 + \hat{\theta}_1 x_{1k})$

$$= \mathbf{V}_M(\hat{\theta}_0)+x_{1k}\mathbf{V}_M(\hat{\theta}_1)+2x_{1k}\mathrm{Cov}(\hat{\theta}_0,\hat{\theta}_1)$$

$$= \lambda^{*}_{00}+x^{2}_{1k}\lambda^{*}_{11}+2x_{1k}\lambda^{*}_{01}$$

$$= \frac{1}{H}\left[(\lambda_{00}+\sigma^{2}a_{00})+x_{1k}(\lambda_{11}+\sigma^{2}a_{11})+2x_{1k}(\lambda_{01}+\sigma^{2}a_{01})\right]$$

$$\cdots(3.65)$$

(b) $$\mathrm{Cov}(\hat{y}_k,\hat{y}_m) = \mathrm{Cov}[(\hat{\theta}_0+\hat{\theta}_1 x_{1k}),(\hat{\theta}_0+\hat{\theta}_1 x_{1k})]$$

$$= \mathrm{Cov}(\hat{\theta}_0,\hat{\theta}_0)+x_{1k}\mathrm{Cov}(\hat{\theta}_0,\hat{\theta}_1)+x_{1m}\mathrm{Cov}(\hat{\theta}_0,\hat{\theta}_1)+$$

$$x_{1k}x_{1m}\mathrm{Cov}(\hat{\theta}_1,\hat{\theta}_1)$$

$$= \lambda^{*}_{00}+(x_{1k}+x_{1m})\lambda^{*}_{01}+x_{1k}x_{1m}\lambda^{*}_{11}$$

$$= \frac{1}{H}\left[(\lambda_{00}+\sigma^{2}a_{00})+(x_{1k}+x_{1m})(\lambda_{01}+\sigma^{2}a_{01})\right.$$

$$\left.+\ x_{1k}x_{1m}(\lambda_{11}+\sigma^{2}a_{11})\right] \qquad (3.66)$$

(c) $$\mathrm{Cov}(\hat{y}_k,e_m) = \mathrm{Cov}(\hat{y}_k,y_m-\hat{y}_m)$$

$$= \mathrm{Cov}(\hat{y}_k,y_m)-\mathrm{Cov}(\hat{y}_k,\hat{y}_m)$$

$$= \mathrm{Cov}[(\hat{\theta}_0+\hat{\theta}_1 x_{1k}),(\beta_{0m}+\beta_{1m}x_{1m}+\xi_m)]-\mathrm{Cov}(\hat{y}_k,\hat{y}_m)$$

$$= \mathrm{Cov}(\hat{\theta}_0,\beta_{0m})+\mathrm{Cov}(\hat{\theta}_1,\beta_{0m})x_{1k}+\mathrm{Cov}(\hat{\theta}_0,\beta_{1m})x_{1m}+$$

$$\mathrm{Cov}(\hat{\theta}_1,\beta_{1m})x_{1k}x_{1m}-\lambda^{*}_{00}-(x_{1k}+x_{1m})\lambda^{*}_{01}-x_{1k}x_{1m}\lambda^{*}_{11}$$

$$= \frac{1}{H}\left[\lambda_{00}+(x_{1k}+x_{1m})\lambda_{01}+x_{1k}x_{1m}\lambda_{11}\right]-\left\{\frac{1}{H}\left[(\lambda_{00}+\sigma^{2}a_{00})\right.\right.$$

$$\left.\left.+(\lambda_{01}+\sigma^{2}a_{01})(x_{1k}+x_{1m})+(\lambda_{11}+\sigma^{2}a_{11})x_{1k}x_{1m}\right]\right\}$$

$$= -\frac{\sigma^{2}}{H}\left[a_{00}+(x_{1k}+x_{1m})a_{01}+x_{1k}x_{1m}a_{11}\right] \qquad (3.67)$$

5) $V_M E_D(\hat{T}_{Cq})$

$$= V_M(\sum_{U\cdot q}\hat{y}_k)+V_M(\frac{N_{\cdot q}}{N}\sum_U e_k)+2\mathrm{Cov}(\sum_{U\cdot q}\hat{y}_k,\frac{N_{\cdot q}}{N}\sum_U e_k)$$

$$= \sum_{U\cdot q}V_M(\hat{y}_k)+\sum_{k\neq m\ U\cdot q}\mathrm{Cov}(\hat{y}_k,\hat{y}_m)+(\frac{N_{\cdot q}}{N})^2\sum_U V_M(e_k)$$

$$+(\frac{N_{\cdot q}}{N})^2\sum_{k\neq m\ U}\mathrm{Cov}(e_k,e_m)+2\frac{N_{\cdot q}}{N}\sum_{k\in U\cdot q,\ m\in U}\mathrm{Cov}(\hat{y}_k,e_m)$$

$$=\sum_{U\cdot q}\frac{1}{H}[(\lambda_{00}+\sigma^2 a_{00})+x_{1k}(\lambda_{11}+\sigma^2 a_{11})+2x_{1k}(\lambda_{01}+\sigma^2 a_{01})]$$

$$+\sum_{\substack{U\\ k\neq m\cdot q}}\frac{1}{H}[(\lambda_{00}+\sigma^2 a_{00})+(x_{1k}+x_{1m})(\lambda_{01}+\sigma^2 a_{01})+x_{1k}x_{1m}(\lambda_{11}+\sigma^2 a_{11})]$$

$$+(\frac{N_{\cdot q}}{N})^2 N[(1-\frac{1}{H})(\lambda_{00}+\lambda_{01}+\lambda_{11})+\frac{\sigma^2}{H}(a_{00}+a_{01}+a_{11})]+\frac{N_{\cdot q}^2}{N}\sigma^2$$

$$+(\frac{N_{\cdot q}}{N})^2\{\sum_{h=1}^{H}(N_{h\cdot}-1)N_{h\cdot}[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})+(1-\frac{1}{H})(\lambda_{00}+\lambda_{11}2\lambda_{01})]$$

$$+[N(N-1)-\sum_{h=1}^{H}(N_{h\cdot}-1)N_{h\cdot}][\frac{\sigma^2}{H}(a_{00}+a_{11}2a_{01})-\frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})]$$

$$+N(N-1)\sigma^2\}+\frac{2N_{\cdot q}}{N}\sum_{\substack{k\in U\cdot q\\ m\in U}}\{\frac{1}{H}[\lambda_{00}+(x_{1k}+x_{1m})\lambda_{01}+x_{1k}x_{1m}\lambda_{11}]$$

$$-[\frac{1}{H}(\lambda_{00}+\sigma^2 a_{00})+\frac{1}{H}(\lambda_{01}+\sigma^2 a_{01})(x_{1k}+x_{1m})+\frac{1}{H}(\lambda_{11}+\sigma^2 a_{11})x_{1k}x_{1m}]\}$$

$$\cdots(3.68)$$

From (3.51)

$$E_M V_D(\hat{T}_{Gq})$$

$$=E_M\Big[\ \frac{N_{\cdot q}}{n}\sum_U e_k^2+\frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)}{n(N-1)^2}\sum_{k\neq m\in U} e_k e_m$$

$$-\frac{2N_{\cdot q}}{N}\Big(\sum_{\substack{m\notin U_{\cdot q}\\ k\in U_{\cdot q}}} e_k e_m\Big)+\Big(1-\frac{2N_{\cdot q}}{N}\Big)\Big(\sum_{U_{\cdot q}} e_k^2+\sum_{k\neq m\in U_{\cdot q}} e_k e_m\Big)\Big]$$

$$=\frac{NN_{\cdot q}}{n}\Big[\Big(1-\frac{1}{H}\Big)(\lambda_{00}+\lambda_{01}+\lambda_{11})+\frac{\sigma^2}{H}(a_{00}+a_{01}+a_{11})\Big]+\frac{NN_{\cdot q}}{n}\sigma^2$$

$$+\frac{(n-1)N_{\cdot q}(N_{\cdot q}-1)}{n(N-1)^2}\Big\{\sum_{h=1}^{H}(N_{h\cdot}-1)N_{h\cdot}\Big[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})$$

$$+\Big(1-\frac{1}{H}\Big)(\lambda_{00}\ \lambda_{11}+2\lambda_{01})\Big]+\Big[N(N-1)-\sum_{h=1}^{H}(N_{h\cdot}-1)N_{h\cdot}\Big]\Big[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})$$

$$-\frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})\Big]+N(N-1)\sigma^2\Big\}-\frac{2N_{\cdot q}}{N}\Big\{\sum_{h=1}^{H}N_{hq}(N_{h\cdot}-N_{hq})$$

$$\cdot\Big[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})+\Big(1-\frac{1}{H}\Big)(\lambda_{00}+\lambda_{11}+2\lambda_{01})\Big]+\Big[N_{\cdot q}(N-N_{\cdot q})$$

$$-\sum_{h=1}^{H}N_{hq}(N_{h\cdot}-N_{hq})\Big]\Big[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})\ \frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})\Big]$$

$$+N_{\cdot q}(N-N_{\cdot q}\sigma^2\Big\}+\Big(1-\frac{2N_{\cdot q}}{N}\Big)\Big\{\Big(N_{\cdot q}-\frac{N_{\cdot q}}{H}\Big)(\lambda_{00}+\lambda_{11}+\lambda_{01})$$

$$+\frac{N_{\cdot q}\sigma^2}{H}(a_{00}+a_{01}+a_{11})+N_{\cdot q}\sigma^2+\sum_{h=1}^{H}(N_{hk}-1)N_{hk}\Big[\frac{\sigma^2}{H}(a_{00}+a_{11}+2a_{01})$$

$$+\Big(1-\frac{1}{H}\Big)(\lambda_{00}+\lambda_{11}2\lambda_{01})\Big]+\Big[N_{\cdot q}(N_{\cdot q}-1)-\sum_{h=1}^{H}(N_{hk}-1)N_{hk}\Big]$$

$$\cdot\Big[\frac{\sigma^2}{H}(a_{00}+a_{11}2a_{01})-\frac{1}{H}(\lambda_{00}+\lambda_{11}+2\lambda_{01})\Big]+N_{\cdot q}(N_{\cdot q}-1)\sigma^2\Big\} \qquad (3.68)$$

Adding (3.68) and (3.69), we get the expression in (3.50) which proves the theroem. ■

Even though, the formula for the variance of the estimator is obtained for one independent variable explicitly and it can be easily

generalized to the general case of more than one independent variable, these formulas are not very convenient for estimating the variance of the estimator.

The general technique of jackknifing for this purpose is useful which will be demonstrated in the next chapter.

CHAPTER 4

A NUMERICAL STUDY

—— THE ESTIMATION OF AVERAGE INCOME OF A FAMILY IN QUÉBEC

4.1 Description Of The Data

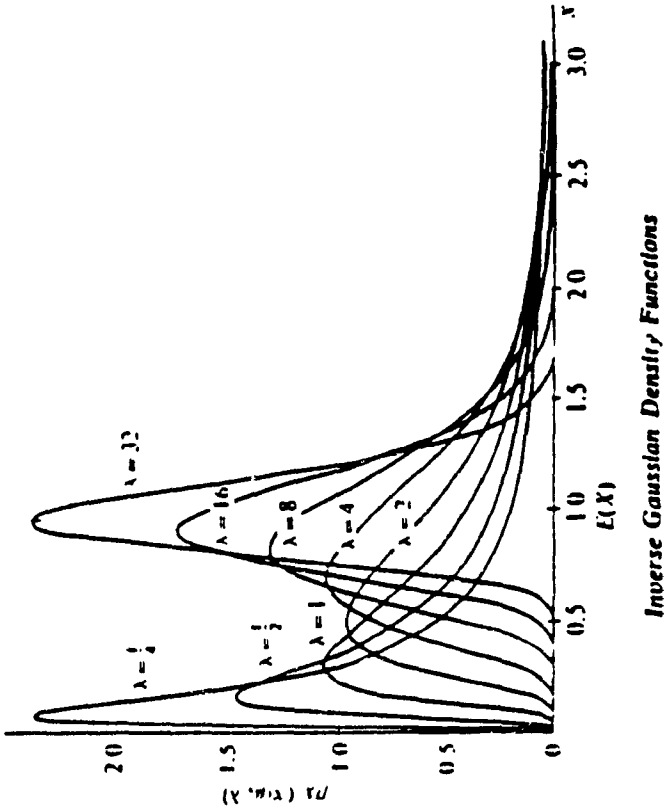
In this chapter, we'll give a numerical study of the generalized regression approach with random coefficients

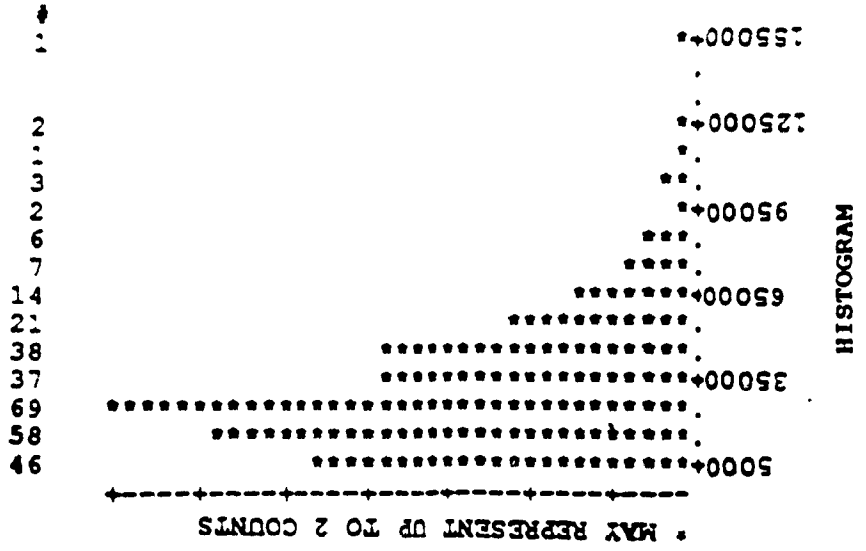
The data is given by Statistics Canada, Household Surveys Division, a microdata file named "Housed Income (1987), facilities and Equipment (1987)" It is from the survey of consumer finances (1987) and survey of household facilities(1987) it contains 30841 records selected randomly from all over Canada

The purpose of my work is to estimate the average income of a family in Québec and the variance of the estimation.

We regard 30841 records as superpopulation and take ten percent of 30841 records, say popu , as the population of Canada and take ten percent of popu. as the sample making estimation, say sam2..In this example the size of popu. is 3099, and 466 of the 3099 records are belong to Québec province, the size of sam2 is 305, and 37 of the 305 records are belong to Québec province.

We have a histogram for sam2..The number of the horizontal axis is the average income of a family, the quantity of the vertical axis is the number of the family.Comparing with the distribution of inverse Gaussian, we can see the distribution of sam2.fits in with the distribution of inverse Gaussian with a certain parameter.


Inverse Gaussian Density Functions
E(X)
λ = 16
λ = 8
λ = 4
λ = 2
λ = 1


HISTOGRAM
* MAX REPRESENT UP TO 2 COUNTS

4.2 Computation Of Estimator And Estimate Of Its Variance(Jackknife)

In this example, we take the whole country is the Domain of study and Québec province is the small domain. We divide the country into ten domains (ten provinces).

Consider the estimation as a general regression with random coefficients model.

Making this issue simple, we only take one auxiliary varible. Regarding the number of persons with income in the family as being high correlativity with the average income of a family, it is reasonble to select the number as an auxiliary variable x. The another factor affecting the average income of the family is the household head's education. As like in the example in section(1.1), the different groups of eduction have the different straght lines of regression, we classify the families into six groups by the household head's education.

Let y be the income of the family. We have model

$$y_{hqi_q} = \beta_{ho} + \beta_{h1} x_{hqi_q} + \xi_{hqi_q} \tag{4.1}$$

$$h = 1,2,\cdots 6, \qquad q = 1,2,\cdots 10, \qquad i_q = 1,2,\cdots i_Q$$

where i_Q is the size of qth domain.

The estimator of y_k given by

$$\hat{y}_k = \hat{\theta}_0 + \hat{\theta}_1 x_k \tag{4.2}$$

where $(\hat{\theta}_0 \ \hat{\theta}_1) = (\bar{\hat{\beta}}_0 \ \bar{\hat{\beta}}_1) = \frac{1}{6}\,(\sum_{h=1}^{6}\hat{\beta}_{ho} \ \ \sum_{h=1}^{6}\hat{\beta}_{h1})$, and $(\hat{\beta}_{ho}, \ \hat{\beta}_{ho})$ are the general regression estimator of (β_{ho}, β_{h1}).

The estimator of average income of the family is given by

$$\bar{\hat{T}}_{GQ} = \frac{1}{466}\left(\sum_{U\cdot q}\hat{y}_k + \sum_{S\cdot q} e_k/\pi_k\right) \qquad (4.3)$$

We use the jackknifing technique to estimate the variance of estimator (4.3). The steps of the technique are:

(1) Order the sample sam2., eliminate the first one of sam2. and obtain the new estimator based on (4.3), say $\bar{\hat{T}}_{(1)}$.

. Eliminate the second one of sam2. and obtain $\bar{\hat{T}}_{(2)}$. And so on we obtain $\bar{\hat{T}}_{(3)}, \bar{\hat{T}}_{(4)}, \cdots \bar{\hat{T}}_{(305)}$.

(3) Take the mean of 305 new estimators, we obtain $\bar{\hat{T}}_{(\cdot)}$.

(4) The estimator of variance of (4.3) is given by

$$\mathrm{Var}(\bar{\hat{T}}_{GQ}) = \frac{304}{305}\sum_{i=1}^{305}(\bar{\hat{T}}_{(i)} - \bar{\hat{T}}_{(\cdot)})^2 \qquad (4.4)$$

4.3 Summary Of Results

The results are :

(1) The real average income of a family in Québec obtained from data popu. is 32247.3.

(2) The estimation of average income of a family in Québec applying (4.3) is 33691.6

(3) If we only use the first term of (4.3) to estimate the average income of a family in Québec, we get 35115.2.

(4) the estimation of variance of $\bar{\hat{T}}_{GQ}$, say $\hat{\sigma}^2$, is 5955390.4, so σ is 2440

Summary·

(1) The relative error ,(33691.6 - 32247.3)/32247.3 = 0.0448, is small,so the estimation is good.

(2) The error term in (4.3) play a very important role, it reduces the error a lot.

(3) The relative error of estimation is, 2440/33691.6 = 0.0724, which is small. The real value is in the interval of one σ error, so the estimation is good.

REFERENCES

1. Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

2. Baker, G.A (1941). Linear regression when the standard deviation of arrays are not equal. *J. Amer. Statist.* 36, 500-506.

3. Battese, G.E. and Fuller, W.A. (1981). Predicition of county crop areas using survey and Satellite data, *Proceedings of the Survey research Methods Section, Amer. Statist. Assoc.* 500-505.

4. Box, [illegible] and [illegible], A. (1973), *bayesian Inference in [illegible] Analysis*. Addison Wesley, Cambridge, Massachusetts.

5. Bunke, H. and Glanditz, J. (1974). Empirical linear Bayes decision rules for a sequence of linear models with different regressor matrices. *Math. Operationsforsch. Statist.* 5, 235-244.

6. Cassel, C.M., Sarndal, C.E., and Wretman, J.H. (1976), Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika*, 63, 615-620.

7. Chaubey, Y.P., Dwivedi, T.D. and Singh, M. (1981). A note on the product estimator *Metro*, 34(3,4), 142-145.

8. _____ (1984 a). A note on the optimality of the regression estimator. *Biom. Journal*, 26(4), 465-467.

9. _____ (1984 b). An efficiency comparison of product and ratio estimator. *Commun. Statist.-Theor. Meth.*, 13(6)699-709.

10. Cochran, W.G. (1977) *Sampling Techniques (3rd ed.)*, New York: John Wiley.

11. Cohen, S.B. and Kalsbeek, W.D. (1977). An alternative strategy for

estimating the parameters of local areas. *1977 proceedings of the Social Statistics Section*, American Statistical Association, 781-785.

12. Drew, J.H., Singh, M.P., and Choudry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian labour force survey, *Survey Methodology*, 8(1), 17-47.

13. Ericksen, E.P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas, *Demography*, 10, 137-159.

14. Gonalez, M E. (1973). Use and evaluation of synthetic estimates, *Proceedings of the American Statistical Association, Social Statistics Section*, 867-875

15. Fay, R.E (1978). Some recent Census Bureau applications of regression techniques to estimation. Presented at the NIDA/NCHS Workshop on Synthetic Estimates, Princetion, New Jersey, April 13-14 Proceedings to be published by NIDA, 1979.

16 Fay, R.E, and Herriot R A. (1979), Estimation of income for small places. An application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74, 269-277

17. Fisk, P.R. (1967). Models of the second kind in regression analysis *J. Roy. Statist. Soc.* B 29, 266-281.

18 Froellich, B R (1973). Some estimators for a random coefficient regression model *J. Amer. Statist. Assoc.* 68, 329-335.

19. Gonzalez, M E (1973). Use and evaluation of synthetic estimates, *Proceedings of the Social Statistics Section. American Statistical Association*, 82-107.

20. Gonzalez, M.E. and Hoza, C. (1978). Small area with application to

unemployment and housing estimates. *Journal of American Statistical Association* 73,7-15.

21. Holt,D.,Smith,T.M.F., and Tombrlin,T.J. (1979). A model based approach to estimation for small subgroups of a population.*Journal of the American Statistical Association,*74,405-410.

22. Isaki,C.T.,and Fuller,W.A.(1982).Survey design under a regression superpopulation model,*Journal of the American Statistical Association,*77,89-96.

23. Kish,L. (1965).*Survey Sampling.*John Wiley and Sons,Inc.,New York.

24. [illegible] A prediction approach to subdomain estimation in finite population,*Journal of the American Statistical Association,* 74,355-358

25. Levy,P.S (1971).The use of mortality data in evaluating synthetic estimates,*Proceedings of the Social Statistics Section,American Statistical Association,*328-331.

26. Little,R.J.A (1983a). Estimating a finite population mean from unequal probability samples,*Journal of the American Statistical Association,*78,596-604.

27. Li Yong Shing,L.S.W. (1985).The use of auxiliary information for estimation in finite population sampling,Unpublished M.Sc. thesis, Concordia University,Montreal Canada.

28. Purcell,Noel J. and Kish,L.(1979).Estimation for small domains,*Biometrics* June 1979.

29. Raj,B (1975) Linear regression with random coefficients: The finite sample and convergence properties.*J.Amer. Statist.Assoc.*70,127-137.

30. Rao,C.R (1965). The theory of least squares when the parameters are

stochastic and its application to the analysis of growthcurves. *Biometrika* 52,447-458.

31. Rosenberg,B.(1973). Linear regression with randomly dispersed parameters. *Biometrika* 60 65-72.

32. Royall,R. (1970).On finite population sampling theory under certain linear regression models,*Biometrika*,57,377-387

33. Royall,R.(1977).Statistical theory of small area estimates use of predictor models.Unpublished technical report prepared under contract from the National Center for Health Statistics

34 Särndal, . , and Raback, (1983). Var. and unbiasedness for small domain estimators, *Statistical Review*, 21,5, (Essays in honour of T.E. dalenius),33-40.

35. Särndal,C.E.(1983). Estimation for small domains, *Journal of the American Statistical Association*,VOlume 79 624-631

36 Swamy,P.A.V.B (1970). Efficient inference in a random Coefficient Regression Model. *Econometrica*38,311-323.

37. Swamy,A.V.B.and Mehta,J.S.(1975).Bayesian and non-Bayesian analysis of switching regressions and of random coefficients regression models. *J.Amer.Statist.Assoc.*70, 593-702.

38 Statistics Canada(1987).Household Surveys Division, *household Income,facilities and Equipment*

39 U.S.National Center for Public Health Statistics(1968) *Synthetic Estimates of Disability*,PHS Publication No. 1759

40 Woodruff,R.S.(1966) Use of a regression technique to produce area breakdowns of the monthly national estimates of retail trade *Journal of the American Statistical Association* 61,496-504