DESIGN AND REGRESSION ESTIMATION IN DOUBLE SAMPLING

by

Edith Estillore Tan

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

APPROVED:

J. C. Arnold, Chairman

R. H. Myers

R. V. Foutz

C. A. Brooks

K. Hinkelmann

August, 1987

Blacksburg, Virginia

DESIGN AND REGRESSION ESTIMATION

IN DOUBLE SAMPLING

by

Edith Estillore Tan

Committee Chairman: Jesse C. Arnold

Statistics

(ABSTRACT)

Two methods developed to improve regression estimation in double sampling under the superpopulation model approach are examined. One method proposes the use of an alternative double sample regression estimator. The other method recommends the use of nonrandom, purposive subsampling plans. Both methods aim to reduce the mean squared errors of regression estimators in double sampling.

A major criticism against the superpopulation model approach is its strong dependence on the correctness of the assumed model. Thus, two purposive subsampling plans were considered. The first plan designed subsamples based on the assumption that the superpopulation model was a first order linear model. The second plan selected subsamples that guarded against the occurrence of a second order model. As expected, the designed subsamples without protection can be very sensitive to the presence of a second order linear model. On the other hand, the designed subsamples with protection rendered the double sample regression estimators robust not only to a second order superpopulation model but also fairly robust to other slight model deviations such as variance misspecification. Therefore the use of designed subsamples with protection against a second order model is suggested whenever a first order

superpopulation model is uncertain.

Under designed subsamples with or without protection, the alternative double sample regression estimator is not found to be more efficient than the usual double sample regression estimator found in most sampling textbooks. However, the alternative double sample regression estimator has shown itself to be more efficient under simple random subsampling when the correlation between variables is weak and subsamples are small.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to the members of my dissertation committee. First, I would like to thank Dr. Jesse Arnold for directing this research with patience and understanding. His suggestions and encouragement have proven invaluable. I am grateful to Dr. Ray Myers for always finding time in his busy schedule to answer questions and give encouragement. I am thankful to Dr. Robert Foutz, Dr. Camilla Brooks, and Dr. Klaus Hinkelmann for their comments and assistance.

I commend and thank for the excellent typing of this thesis. I also thank her for being a friend.

I extend my appreciation and gratitude to the rest of the faculty and staff of the Statistics Department for their warm friendship and assistance, especially . Her willingness to listen and help, and her encouraging words have contributed much to this work. Special thanks go to all my friends who have made the years in graduate school memorable.

Most of all, I am deeply indebted to my entire family who believed in me and stood by me. I am especially thankful to my husband, , for his love and support without which this project could not have been completed, and to my parents who provided me with the opportunities in making this possible, right from the start, twenty nine years ago.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

LIST OF TABLES

LIST OF TABLES (cont.)

LIST OF TABLES (cont.)

I. INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

Estimation of a population characteristic is and has always been a major goal of sample surveys. Governments of different countries worldwide conduct sample surveys to estimate agricultural production, unemployment, industrial production, size of the labor force, family income and expenditures, and other relevant attributes vital to national planning. Other areas which use sample surveys include business and industry. Market researchers use sample surveys to estimate the sizes of television and radio audiences as well as to gauge the reactions of people to new products. This is just one of the many applications of sample survey estimation in business. Although present day statisticians will all agree in describing sampling as being invaluable and standard practice, it was not until 1925 that "representative" sampling became accepted and respectable among statisticians. Before 1925, full coverage or complete enumeration was the rule. Kiaer, Director of the Norwegian Central Bureau of Statistics, first introduced the notion of "representative" sampling at the 1895 Berne Meeting of the International Statistical Institute. Kiaer was strongly opposed by many leaders in statistics who argued for full enumeration. One of the most vocal proponents of full coverage was Georg von Mayr. Kruskal and Mosteller (1980) gave a fascinating account of the history of "representative" sampling in statistics.

1.2 APPROACHES TO FINITE POPULATION INFERENCE

The widely accepted approach to sample survey theory has been randomization inference but, in recent years, it has been vigorously challenged by the model-based approach. The ongoing debate on the proper approach to

finite population inference has stimulated several papers offering arguments and counterarguments.

In the randomization approach, the probability structure used as the basis for inference is generated by the sampling plan used. If Y is the variable under study, then the values of Y are treated as unknown constants, and probabilities are introduced through random selection or artificial randomization. In the model-based approach the values of Y are treated as realizations of random variables giving rise to what is known as "superpopulation models".

The foundation for randomization inference is based on Jerzy Neyman's (1934) famous Royal Statistical Society paper. Here Neyman revealed the concept of confidence intervals. He stated,

> If we are interested in a collective character X of a population π and use methods of sampling and of estimation, allowing us to ascribe to every possible sample, Σ, a confidence interval $X_1(\Sigma)$, $X_2(\Sigma)$ such that the frequency of errors in the statement
>
> $$X_1(\Sigma) \leq X \leq X_2(\Sigma)$$
>
> does not exceed the limit $1 - \epsilon$ prescribed in advance, whatever the unknown properties of the population, I should call the method of sampling representative and the method of estimation consistent.

This statement was the forerunner of statistical inference from finite populations within the framework of repeated samples for a given randomization scheme.

Two properties of estimators that the randomization approach finds desirable are: design-unbiasedness and design-consistency. If the sampling strategy consists of a sampling plan p and an estimator of T, $\hat{T}$, then $\hat{T}$ is defined to be design-unbiased if $E_p(\hat{T}) = T$ where

$$E_p(\hat{T}) = \sum_s p(s)\hat{T} \ . \qquad (1.1)$$

The sampling plan p assigns for every possible sample s the probability of selection $p(s) \geq 0$ such that $\sum_s p(s) = 1$. If an estimate equals the quantity being estimated when the sample consists of the whole population, then the estimate is design-consistent. Large sample sizes have also led to concepts such as asymptotic unbiasedness and asymptotic consistency where asymptotics are defined by forming k replicates of the population, selecting a sample from each replicate according to the sampling plan, aggregating the samples and populations, and letting k tend to infinity. (See Brewer, 1979.)

The development of the model-based approach to finite population inference was stimulated by the nonexistence results of Godambe (1955). He proved that under the randomization principle, a uniformly minimum variance unbiased estimator does not exist for the entire class of linear estimators. Godambe (1966) also showed that, under the randomization approach, the likelihood function for the population parameter $\underline{y} = (y_1, y_2, \ldots, y_N)$ where $\underline{y}$ is a member of Y, the parameter space containing all possible populations, is uninformative. The likelihood function defined on Y partitions Y into two sets, those with zero likelihood and those with a constant likelihood greater than zero. Smith (1976) discussed how concepts in classical statistical inference, substantiated by the above results, lead to problems within the randomization

framework. These results raised questions about the randomization principle as being the basis for inference from finite populations. Royall (1983) even went as far as to call it a dubious principle.

In the model-based approach, superpopulation models characterize the actual population values, both those observed in the sample and the unobserved ones, as realizations of random variables. The superpopulation model provides the relationship between the observed units to the unobserved ones. It places the finite population estimation problem in the area of predictive statistical inference. The randomization approach, on the other hand, provides no such connection between the observed units to the unobserved ones except that the unobserved units could have been in the sample. It is this writer's opinion that inference based on such a premise has severe limitations even if it is always probabilistically valid.

Under the superpopulation model approach, there are different methods of handling finite population estimation problems as problems in prediction. Both Bayesian and fiducial prediction techniques have been applied to finite populations under superpopulation models. (See Ericson (1969); Scott and Smith (1969); Kalbfleisch and Sprott (1969).) A non-Bayesian approach was undertaken by Royall (1970). He used the linear least-squares prediction approach where he applied the linear model and the Gauss-Markov theorem to the problem. Royall and Cumberland (1981a) indicated that the linear least-squares prediction theory yields results that are comparable to those from randomization theory.

There is a consensus among statisticians that both randomization and models have important roles to play in the design and analysis of surveys. The issue is their respective roles in inference after the sample is observed.

For inferences from an observed sample, the conditionality principle (Cox and Hinckley, 1974) suggests that the relevant distribution is the conditional distribution given the sample, s. Within the randomization framework, this conditional distribution is degenerate implying that inference should be based on the model rather than on the sampling distribution. The randomization theory of inference is also a large-sample theory and as such is more limited in scope than the modeling approach. The model-based approach then has a decided advantage for drawing inferences from small samples. It also extends directly to the analysis of nonsampling errors such as nonresponses and missing values. These are not readily handled by randomization theory because the presence of nonresponses in a survey may be regarded as a form of self selection resulting in a nonrandom sampling scheme which ruins the basis for probability sampling methods. These features of the modeling approach as well as the fact that the prediction framework puts it in the mainstream of statistical analysis make it appealing to statisticians working in areas other than with sample surveys.

The main criticism against the prediction approach is its dependence on the assumption of a superpopulation model. No model can perfectly represent the real-world population that generates the data. Small deviations from the model that are undetectable in samples may lead to seriously misleading inferences. With randomization theory, for large enough samples, the sampling design provides consistent estimators and valid confidence intervals regardless of whether the presumed model is correct. The validity of the results depends only on the process of random selection which is under the sampler's control and can be fully described. Royall (1976a) pointed out that probabilistically correct results are, however, not necessarily inferentially right. Modelers argue that models provide relationships which are essential for making inferences that are

often ignored in probability sampling analyses. Randomization theorists, on the other hand, claim that making seemingly reasonable assumptions to provide a relationship is an unnecessary risk when results based on randomization are always valid without having to defend assumptions. Although they recognize models as valuable tools in their choice and evaluation of sampling designs, they are of the opinion that it is important to refrain from inferences that need to be defended as subjective judgments.

The main question in current literature has become one of robustness ——— insensitivity to failure of assumptions. In recent years, proponents of the model-based approach have concentrated on making model-based methods robust through the choice of particular sampling schemes and estimators of variance. Results hint that the gap between the two competing theories may not be as wide as some think.

1.3 RATIO AND REGRESSION ESTIMATION

Ratio estimation and regression estimation are two methods of estimation that are widely used in sample surveys when auxiliary information is available to increase the precision of estimates. Cochran (1942) provided an account of the theory behind ratio and regression estimators. Olkin (1958) extended the ratio estimator to the situation in which p auxiliary X-variables are available in the case of simple random sampling and stratified sampling. Des Raj (1965) has also used multivariate auxiliary information in constructing difference estimators.

Let $U = \{1, 2, \ldots, N\}$ denote a finite population of N units. Associated with each unit i ($i = 1, 2, \ldots, N$) are two numbers (x_i, y_i) where X and Y are correlated. Suppose one wants to estimate the population mean

$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$. When $x_1, x_2, \ldots, x_N$ are known, under simple random sampling, the ratio estimator of $\bar{Y}$ is

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X} , \qquad (1.2)$$

where

$\bar{y}$ is the sample mean of the y's,

$\bar{x}$ is the sample mean of the auxiliary variable, the x's,

and

$\bar{X}$ is equal to $\frac{\sum_{i=1}^{N} x_i}{N}$.

The ratio estimator $\hat{\bar{Y}}_R$ is the best linear unbiased estimator of $\bar{Y}$ when

i) the relationship between Y and X is linear and goes through the origin,

and

ii) the variance of Y_i is proportional to X_i .

The concept of unbiasedness being used here differs from that of randomization theory. An estimator $\hat{\bar{Y}}$ is defined to be model-unbiased for $\bar{Y}$ with respect to the model M if $E_M[\hat{\bar{Y}}] = E_M[\bar{Y}]$. Here $\hat{\bar{Y}}_R$ is defined as unbiased with respect to the model

$$Y_i = \beta x_i + \epsilon_i , \; x_i > 0, \; i = 1, 2, \ldots, N \qquad (1.3)$$

where

$$E(\epsilon_i | x_i) = 0 ,$$

and

$$E(\epsilon_i\epsilon_j|x_i, x_j) = \sigma_\epsilon^2 x_i \qquad \text{for } i = j ,$$

$$= 0 \qquad \text{for } i \neq j .$$

When the relationship between Y and X, although linear, does not go through the origin, the regression estimator would be more appropriate. The simple linear regression estimator of $\bar{Y}$, under simple random sampling, is

$$\bar{Y}_{\ell_r} = \bar{y} + b(\bar{X} - \bar{x}) , \qquad (1.4)$$

where

$\bar{y}$ is the sample mean of the y's,

$\bar{x}$ is the sample mean of the x's,

$\bar{X}$ is equal to $\frac{\sum_{i=1}^{N} x_i}{N}$, and

b is the linear least-squares regression coefficient estimated from the sample.

(There are cases when b is preassigned due to prior information.) Under the superpopulation model,

$$Y_i = \beta_o + \beta_1 x_i + \epsilon_i , \qquad (1.5)$$

where

$$E(\epsilon_i|x_i) = 0 ,$$

and

$$E(\epsilon_i\epsilon_j|x_i, x_j) = \sigma_\epsilon^2 \qquad \text{for } i = j ,$$

$$0 \qquad \text{for } i \neq j ,$$

$\bar{Y}_{\ell_r}$ is the best linear unbiased estimator of $\bar{Y}$.

Royall (1970) gave a more general result on best linear unbiased estimators (BLUE) under certain linear regression models. Given that the ϵ_i's are independent random variables with mean zero, and variance $\sigma_\epsilon^2 v(x_i)$ where $v(x_i) > 0$ for $x_i > 0$ and $v(x)$ is a known function, he gave the BLUE estimators for the population total $T = \sum_{i=1}^{N} y_i$ under two regression models:

$$\text{i) } Y_i = \beta x_i + \epsilon_i , \qquad i = 1, 2, \ldots , N , \tag{1.6}$$

$$\text{ii) } Y_i = \alpha + \beta x_i + \epsilon_i , \qquad i = 1, 2, \ldots , N . \tag{1.7}$$

For model i), the BLUE of T is of the form

$$\hat{T}^* = \sum_s y_i + \hat{\beta}^* \sum_{\bar{s}} x_i , \tag{1.8}$$

where

$$\hat{\beta}^* = \sum_s \{x_i y_i / v(x_i)\} / \sum_s \{x_i^2 / v(x_i)\} ,$$

letting s denote the set of units that are in the sample and $\bar{s}$, the set of units that are not in the sample.

For model (ii), the BLUE of T is of the form

$$\hat{T}^{**} = \sum_s y_i + (N - n(s))\hat{\alpha}^{**} + \hat{\beta}^{**} \sum_{\bar{s}} x_i , \tag{1.9}$$

where

n(s) is the number of different labels in the sample, and

$\hat{\alpha}^{**}$ and $\hat{\beta}^{**}$ are the weighted least squares estimators of α and β.

In this paper, Royall also advocated the use of a purposive sampling plan to improve mean squared error of estimates. For example, the selection of the n

largest x_i in the population would obviously minimize the mean squared error of the ratio estimator.

In an effort to make model-based methods robust to model misspecification, Royall and Herson (1973a) studied the ratio estimator's behavior under different polynomial regression models. Retaining the notation they used, let $\xi[\delta_0, \delta_1, \delta_2, \ldots, \delta_J: \nu(x)]$ denote the superpopulation model,

$$Y_i = \sum_{j=0}^{J} \delta_j \beta_j x_i^j + \epsilon_i [\nu(x_i)]^{\frac{1}{2}} , \quad i = 1, 2, \ldots, N, \qquad (1.10)$$

where the δ_j's are zeros and ones. If $\delta_j = 1$, it means that the term $\beta_j x^j$ appears in the regression function; otherwise not. The ϵ_i's are independent with mean zero and variance σ_ϵ^2, and $\nu(x)$ is some known function of x. Royall and Herson showed that the ratio estimator remains unbiased under any polynomial regression model $\xi[\delta_0, \delta_1, \delta_2, \ldots, \delta_J: \nu(x)]$ if and only if

$$\frac{\bar{x}_s^{(j)}}{\bar{x}_s} = \frac{\bar{x}^{(j)}}{\bar{x}} , \quad j = 0, 1, \ldots, J , \qquad (1.11)$$

where

$$\bar{x}_s^{(j)} = \frac{\sum_{i \in s} x_i^j}{n} \quad \text{and} \quad \bar{x}^{(j)} = \frac{\sum_{i=1}^{N} x_i^j}{N} .$$

They referred to samples for which condition (1.11) holds as balanced samples. More important, they proved that the ratio estimator is not only unbiased but optimal under the models $\xi[1, \delta_1, \delta_2, \ldots, \delta_J: 1]$, $\xi[\delta_0, 1, \delta_2, \ldots, \delta_J: x]$, $\xi[\delta_0, \delta_1, 1, \ldots, \delta_J: x^2]$, $\ldots$, $\xi[\delta_0, \delta_1, \delta_2, \ldots, 1: x^J]$ for any sequence $\delta_0, \delta_1, \delta_2, \ldots, \delta_J$ of zeros and ones as long as the sample was balanced. Royall and Herson (1973b) also considered stratification on a size

variable as another technique for protecting against model failure. Stratification and balanced sampling used together are shown to be more efficient than balanced sampling alone. Herson (1976) compared the performance of the ratio estimator using extreme, unrestricted random, and balanced sampling plans when applied to an actual population. He found the extreme samples to be inferior to both unrestricted random and balanced samples, with balanced samples being as much as 30 percent more efficient than unrestricted random samples.

Meanwhile, Royall and Eberhardt (1975) turned their attention to robust variance estimators for the ratio estimates. Their results were extended to a more general linear regression model by Royall and Cumberland (1978). Royall and Cumberland (1981a & b) then did empirical studies on both the ratio estimator and regression estimator and estimators of their variances to verify theoretical results obtained.

Scott, Brewer, and Ho (1978) continued Royall and Herson's work by extending their results to a more general estimator of T, $\hat{T}_o$, which is BLUE under the model $\xi[0, 1: v(x)]$ where $v(x)$ is a known function and $v(x) > 0$. Their estimator is given by

$$\hat{T}_o = \sum_{i \in s} y_i + \frac{\sum_{i \in s} y_i x_i / v(x_i)}{\sum_{i \in s} x_i^2 / v(x_i)} \sum_{i \in \bar{s}} x_i \, . \qquad (1.12)$$

Robustness of $\hat{T}_o$ is achieved by taking samples that they refer to as "overbalanced" samples. "Overbalanced" samples are defined to be samples for which

$$\frac{\sum_{i\in s} x_i^{j}}{\sum_{i\in s} x_i} = \frac{\sum_{i\in s} x_i^{j+1}/\nu(x_i)}{\sum_{i\in s} x_i^{2}/\nu(x_i)} \quad , \; j = 0, 1, \ldots , J \; . \qquad (1.13)$$

With "overbalanced" samples, $\hat{T}_o$ remains unbiased for any variance function $\nu(x)$ and is optimal under $\xi[\delta_o, \delta_1, \ldots , \delta_J: V^*(x)]$ where

$$V^*(x) = \nu(x) \sum_{j=0}^{J} \delta_j a_j x^{j-1} \; . \qquad (1.14)$$

Cumberland and Royall (1981) arrived at another balance condition when they studied a familiar estimator, the Horwitz-Thompson estimator ($\hat{T}_{HT}$), and different estimators of its variance under prediction models. They found that $\hat{T}_{HT}$ reamins unbiased for T under a general J^{th} order polynomial regression model provided samples are "Π-balanced". "Π-balanced" samples are samples satisfying

$$\bar{X} \sum_{i\in s} x_i^{j-1}/n - N^{-1} \sum_{i=1}^{N} x_i^{j} = 0, \; j = 0, 1, \ldots , J. \qquad (1.15)$$

Royall and his colleagues introduced some form of balanced sampling as an approach to providing robustness against certain model failures but selection of such balanced samples in practice can be a problem since exact balance will be very difficult if not impossible to achieve. Iachan (1985) addressed the problem by presenting methods for the practical implementation of robust sampling designs for ratio and regression estimators.

1.4 DOUBLE SAMPLE REGRESSION ESTIMATOR

All of the estimators that have been described above require knowledge of all the X values in the population or at least the population mean, $\bar{X}$. This

information will not always be available. When the y's are very expensive to sample and resources are limited, double sampling is a solution to this lack of information if the x's are cheaper to sample than the y's. Some of the earliest works in double sampling were done by Neyman (1938), Bose (1943), and Cox (1952).

Consider the double sample regression estimator. Take an initial sample of size n', measuring only the x values. Then measure the corresponding values of y in a subsample of size n from the initial sample. The double sample regression estimator of $\bar{Y}$ is

$$\bar{y}_C = \frac{\sum_{i=1}^{n} y_i}{n} + \hat{\beta}(\bar{x}' - \bar{x}) , \qquad (1.16)$$

where

$\hat{\beta}$ is the least squares regression coefficient estimated from the subsample,

$\bar{x}$ is the sample mean of the x's in the subsample, and

$\bar{x}'$ is the sample mean of the x's in the initial sample.

The initial sample mean, $\bar{x}'$, is used to estimate the unknown finite population mean, $\bar{X}$. Note that the second sample need not be a subsample of the first. The case of two independent samples was first considered by Bose (1943). Khan and Tripathi (1967) extended regression estimators in double sampling to the case of p auxiliary variates.

From the point of view of regression, $\bar{y}_C$ (equation (1.16)) is simply the predicted response $\hat{y}$ at the point $\bar{x}'$. In Chapter 2, we propose an alternative estimator to $\bar{y}_C$ which is not too dependent on a prediction at a point. The idea of designing the subsamples to reduce the mean squared error of both $\bar{y}_C$ and

the proposed estimator is considered in Chapter 3. Samples designed for robustness to model misspecification are investigated in Chapter 4.

II. REGRESSION ESTIMATION IN DOUBLE SAMPLING

2.1 INTRODUCTION

Double sampling has proven to be an effective sampling technique employed in the absence of required information about an auxiliary variable X. In stratified random sampling, stratification is sometimes done according to the values of X so that there is a need to know the frequency distribution of X. The theory behind the use of double sampling in stratification to get an idea of the unknown frequency distribution of X was first given by Neyman (1938). Ratio and regression estimators also require prior knowledge about the auxiliary variate X, at least its population mean, $\bar{X}$. When one does not have knowledge of $\bar{X}$, then double sampling is a practical solution to the problem. Note, however, that when one takes a preliminary sample measuring only the auxiliary variable, the size of the sample measuring the variable of interest, Y, would normally be reduced because of increased cost. Therefore double sampling is profitable only when the x's are cheaper than the y's and the gain in precision through ratio or regression estimators, or stratification outweighs the loss in precision due to the reduced size of the sample measuring the y's.

It is also worth pointing out that double sampling, also called two-phase sampling, is not the same as two-stage sampling. In two-stage sampling, a sample of primary units is first selected from the population, then a selection of second-stage units or subunits from each chosen primary unit is made. In double sampling, one first picks an initial sample from the population, then a second sample is selected from the first sample or from the population independent of the first sample. Double sampling does not involve primary and secondary sampling units. Royall (1976b) has applied the linear prediction

approach to two-stage sampling. Scott and Smith (1969) used a Bayesian approach to two-stage sampling earlier.

2.2 PROPERTIES OF AN ALTERNATIVE DOUBLE SAMPLE REGRESSION ESTIMATOR

2.2.1 Analytical Results for Proposed Estimator

Using the notation that Royall and Herson used, assume the superpopulation model $\xi[1, 1: 1]$ to be true. (Refer to (1.10).) There is a finite population $U = \{1, 2, \ldots, i, \ldots, N\}$ consisting of N units. Associated with each unit i are two numbers (x_i, y_i) whose relationship is described by the superpopulation model $\xi[1, 1: 1]$. An estimate of $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$ is required. The double sample regression estimator of the population mean, $\bar{Y}$, as defined in (1.16) is

$$\bar{Y}_C = \frac{\sum_{i=1}^{n} Y_i}{n} + \hat{\beta}_1(\bar{x}' - \bar{x}) .$$

Under the superpopulation model $\xi[1, 1: 1]$, the conditional mean squared error of $\bar{Y}_C$ given the x's is

$$E[(\bar{Y}_C - \bar{Y})^2 | x_1, \ldots, x_{n'}]$$

$$= \sigma_\epsilon^2 \left[\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1} (x_i - \bar{x})^2} \right] + \beta_1^2(\bar{x}' - \bar{X})^2 . \qquad (2.1)$$

The first term on the right hand side of equation (2.1) is the conditional mean squared error of the simple linear regression estimator, $\bar{Y}_{\ell_r}$ (1.4), when the x's are fixed and if $\bar{x}'$ is equal to $\bar{X}$. However, since $\bar{x}'$ is only an

estimate of $\bar{X}$, the double sample regression estimator $\bar{Y}_c$, unlike $\bar{Y}_{\ell_r}$, is biased. The squared bias of $\bar{Y}_c$ is accounted for by the second term on the right hand side of equation (2.1).

From a regression standpoint, $\bar{Y}_c$ is the predicted response $\hat{y}$ at the point $\bar{x}'$. How good the estimator $\bar{Y}_c$ is clearly depends on how well the regression equation predicts at the point $\bar{x}'$. It is therefore reasonable to consider an alternative double sample regression estimator to $\bar{Y}_c$ that is not contingent on prediction at a point. We consider here a shrinkage estimator which takes into account the strength of the relationship between Y, the variable of interest, and X, the auxiliary variable.

Consider the following alternative double sample regression estimator:

$$\bar{Y}^* = (1 - k) \frac{\sum_{i=1}^{n} Y_i}{n} + k \frac{\sum_{j=1}^{n'} \hat{Y}_j}{n'}, \qquad (2.2)$$

where

$\hat{Y}_j$ is the predicted response at the point $\underline{x}_j$, $j = 1, 2, \ldots, n'$ and $\underline{x}_j$ is a vector of values $[x_{1j}, x_{2j}, \ldots, x_{pj}]$, (p = number of auxiliary variables used in prediction),

k is the weighting factor, and

$\bar{Y}^*$ is the combined estimator found by weighting the sample mean of the values of Y in the subsample, and the mean of the predicted responses at all x's in the initial sample such that MSE $(\bar{Y}^*)$ is minimized.

The estimator, $\bar{Y}^*$, as defined in equation (2.2), is the alternative double sample regression estimator in its general form. It is general in the sense that there is no restriction on the type of regression function used for prediction.

The regression function can be a linear, nonlinear, polynomial, or multiple regression function.

For the purpose of comparing $\bar{Y}^*$ with $\bar{Y}_C$, consider the following special case of $\bar{Y}^*$ where the regression function used is a simple linear regression. As in $\bar{Y}_C$ (1.16), the second sample is a subsample of the first sample. This special case of $\bar{Y}^*$ is given by

$$\bar{Y}^*_\ell = (1 - k)\frac{\sum_{i=1}^{n} Y_i}{n} + k(\hat{\beta}_0 + \hat{\beta}_1\bar{X}')$$

$$= (1 - k)\frac{\sum_{i=1}^{n} Y_i}{n} + k(\hat{Y}(\bar{X}'))$$

$$= (1 - k)\frac{\sum_{i=1}^{n} Y_i}{n} + k\left[\frac{\sum_{i=1}^{n} Y_i}{n} + \hat{\beta}_1(\bar{X}' - \bar{X})\right]$$

$$= \frac{\sum_{i=1}^{n} Y_i}{n} + k\hat{\beta}_1(\bar{X}' - \bar{X}) , \qquad (2.3)$$

where

$\hat{\beta}_1$ is the least squares regression coefficient estimated from the subsample, and

k is the weighting factor such that $MSE(\bar{Y}^*_\ell)$ is minimized.

Theorem 2.1. Assuming that the y_i's (i = 1, 2, ... , N) are randomly drawn from an infinite superpopulation in which $\xi[1, 1: 1]$ holds, the value of k that minimizes $E[(\bar{Y}^*_\ell - \bar{Y})^2 | x_1, x_2, \ldots , x_{n'}]$ is given by

$$k_C = \frac{(\bar{X} - \bar{x})}{(\bar{x}' - \bar{x})} \cdot \frac{\beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{(\sigma_\epsilon^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2)} \cdot \qquad (2.4)$$

Proof: Under the superpopulation model, $\xi[1, 1: 1]$,

$$\bar{Y}_\ell^* - \bar{Y} = \left[\frac{\sum_{i=1}^{n} Y_i}{n} + k\hat{\beta}_1(\bar{x}' - \bar{x}) \right] - \bar{Y}$$

$$= [\beta_0 + \beta_1 \bar{x} + \bar{\epsilon}_n + k\hat{\beta}_1(\bar{x}' - \bar{x})] - (\beta_0 + \beta_1 \bar{X} + \bar{\epsilon}_N)$$

$$= \beta_1(\bar{x} - \bar{X}) + (\bar{\epsilon}_n - \bar{\epsilon}_N) + k\hat{\beta}_1(\bar{x}' - \bar{x}) \, . \qquad (2.5)$$

It can be shown easily that $\hat{\beta}_1$ can be expressed as

$$\beta_1 + \frac{\sum_{i=1}^{n} \epsilon_i(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \qquad (2.6)$$

Substituting (2.6) for $\hat{\beta}_1$ in (2.5), after completing the squares in $(\bar{Y}_\ell^* - \bar{Y})^2$, equation (2.5), and taking expectations with respect to the model $\xi[1, 1: 1]$, we have the conditional mean squared error of $\bar{Y}_\ell^*$ given the initial sample. It is equal to

$$E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_n']$$

$$= \sigma_\epsilon^2\left(\frac{1}{n} - \frac{1}{N}\right) + \beta_1^2(\bar{x} - \bar{X})^2 + k^2\beta_1^2(\bar{x}' - \bar{x})^2$$

$$+ \sigma_\epsilon^2 \frac{k^2(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + 2k\beta_1^2(\bar{x}' - \bar{x})(\bar{x} - \bar{X})$$

$$= \sigma_\epsilon^2 \left[\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{k^2(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

$$+ \beta_1^2[\{(1 - k)\bar{x} + k\bar{x}'\} - \bar{X}]^2 . \qquad (2.7)$$

By applying standard calculus techniques of minimization to (2.7), result (2.4) follows.

Substituting the optimal value of k into the formula for the conditional mean squared error of $\bar{Y}_\ell^*$ given the x's, the following formula is obtained:

$$MSE(\bar{Y}_\ell^* | x_1, x_2, \ldots, x_{n'}) = \sigma_\epsilon^2 \left[\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\beta_1^2(\bar{X} - \bar{x})^2}{\sigma_\epsilon^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \qquad (2.8)$$

If X and Y are considered as jointly distributed random variables, then $\sigma_\epsilon^2 = E(\epsilon_i^2)$ is equal to $\sigma_y^2(1 - \rho_{xy}^2)$, and the weighting factor k can be expressed in terms of the correlation coefficient between X and Y as follows:

$$k_c = \frac{(\bar{X} - \bar{x})}{(\bar{x}' - \bar{x})} \frac{\rho_{xy}^2}{\rho_{xy}^2 + \frac{\sigma_x^2(1-\rho_{xy}^2)}{s_x^2(n-1)}} . \qquad (2.9)$$

Note that k approaches 1 as $n \to \infty$. When $k = 1$, $\bar{Y}_\ell^*$ reduces to $\bar{Y}_c$.

Under certain conditions, a simple result on the unconditional mean squared error of $\bar{Y}_\ell^*$ is obtained in Theorem 2.2.

Theorem 2.2. If the following conditions are true,

i) the initial sample is a simple random sample,

ii) the subsample is a random subsample of the initial sample,

iii) the variable X is normally distributed, and

iv) the model $\xi[1, 1: 1]$ holds for the entire population,

then $MSE(\bar{Y}_{\ell}^{*}) \leq MSE(\bar{Y}_{C})$.

Proof: Under condition iv), the conditional mean squared error of $\bar{Y}_{\ell}^{*}$ given the initial sample is given by equation (2.7).

Taking the expectation of equation (2.7) with respect to the distribution of X under conditions i), ii), and iii), the unconditional mean squared error of $\bar{Y}_{\ell}^{*}$ is equal to

$$E_X[E_{Y|X}[(\bar{Y}_{\ell}^{*} - \bar{Y})^2 | X_1, X_2, \ldots, X_{n'}]$$

$$= \sigma_\epsilon^2 \left\{ \left(\frac{1}{n} - \frac{1}{N}\right) + E_X\left[\frac{k^2(\bar{X}_{n'} - \bar{X}_n)^2}{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2} \right] \right\}$$

$$+ \beta_1^2 E_X[\{(1 - k)(\bar{X}_n - \bar{X}) + k(\bar{X}_{n'} - \bar{X})\}^2] ,$$

where

$$\bar{X}_{n'} = \frac{\sum_{i=1}^{n'} X_i}{n'} \quad \text{and} \quad \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} .$$

Since $\sigma_\epsilon^2 = \sigma_y^2(1 - \rho_{xy}^2)$ and $\beta_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}$,

$$MSE(\bar{Y}^*_\ell) = \sigma_y^2(1 - \rho_{xy}^2)\left[\left(\frac{1}{n} - \frac{1}{N}\right) + k^2\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{1}{n-3}\right)\right]$$

$$+ \frac{\sigma_y^2\rho_{xy}^2}{\sigma_x^2}\left[\sigma_x^2\{(1 - k)^2\left(\frac{1}{n} - \frac{1}{N}\right) + k^2\left(\frac{1}{n'} - \frac{1}{N}\right)\right.$$

$$\left. + 2(1 - k)k\left(\frac{1}{n'} - \frac{1}{N}\right)\}\right]$$

$$= \sigma_y^2(1 - \rho_{xy}^2)\left[\left(\frac{1}{n} - \frac{1}{N}\right) + k^2\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{1}{n-3}\right)\right]$$

$$+ \sigma_y^2\rho_{xy}^2\left[(1 - k)^2\left(\frac{1}{n} - \frac{1}{n'}\right) + \left(\frac{1}{n'} - \frac{1}{N}\right)\right] \quad . \qquad (2.10)$$

By applying standard calculus techniques of differentiation with respect to k to equation (2.10), the value of k that minimizes $MSE(\bar{Y}^*_\ell)$ given by equation (2.10) is given by

$$k_u = \frac{\rho_{xy}^2(n-3)}{1 + \rho_{xy}^2(n-4)} \quad . \qquad (2.11)$$

Theorem 2.2 follows from the fact that

$$\bar{Y}^*_\ell = \frac{\sum_{i=1}^{n} Y_i}{n} + k_u\hat{\beta}_1(\bar{X}_{n'} - \bar{X}_n)$$

has minimum mean squared error within a class of linear estimators belonging to $Y_S = \{\bar{Y}_S: \bar{Y}_S = \frac{\sum_{i=1}^{n} Y_i}{n} + t\beta_1(\bar{X}_{n'} - \bar{X}_n),\ 0 \leq t \leq 1\}$ of which $\bar{Y}_C$ with $t = 1$ is a member. It is obvious from (2.11) that $k_u = 1$ only if $\rho_{xy}^2 = 1$. Hence $MSE(\bar{Y}^*_\ell) \leq MSE(\bar{Y}_C)$ with equality being achieved only if $|\rho_{xy}| = 1$.

Theorem 2.2 states that under certain conditions, the double sample

regression estimator, $\bar{Y}_C$, is optimal only when there is a perfect linear relationship between the variable of interest and the auxiliary variable. Not only is this an unlikely situation but should this situation occur, then double sampling would not even be warranted.

A case which is more likely to occur in practice is the case of large subsamples. It is noted that k_u, like k_C, approaches unity as $n \to \infty$. This implies that $\bar{Y}_\ell^*$ approaches $\bar{Y}_C$ if the subsample is large enough whether one conditions or does not condition on the initial sample drawn.

Case of preassigned B_1:

There are cases when B_1 is preassigned or known due to prior information. Assuming conditions i) and ii) of Theorem 2.2 hold, it can be shown that the double sample regression estimator, $\bar{Y}_\ell'$, given by

$$\bar{Y}_\ell' = \bar{y} + B_1(\bar{x}' - \bar{x}) , \qquad (2.12)$$

under randomization theory, has minimum variance within a class of linear unbiased estimators, $Y_A = \{\bar{Y}_A: \bar{Y}_A = \bar{y} + kB_1(\bar{x}' - \bar{x})\}$.

$$\mathrm{Var}(\bar{Y}_A) = \mathrm{Var}(\bar{y}) + k^2B_1^2\mathrm{Var}(\bar{x}' - \bar{x})$$

$$+ 2k\beta_1 \mathrm{Cov}(\bar{y}, \bar{x}' - \bar{x})$$

$$= S_y^2(\frac{1}{n} - \frac{1}{N}) + k^2R_{xy}^2S_y^2(\frac{1}{n} - \frac{1}{n'})$$

$$- 2kR_{xy}^2S_y^2(\frac{1}{n} - \frac{1}{n'}) \cdot$$

Taking the derivative of $\mathrm{Var}(\bar{Y}_A)$ with respect to k and setting it equal to zero, k is found to be equal to one.

Assume conditions iii) and iv) of Theorem 2.2 in addition to conditions i) and ii) of the same theorem. Under the superpopulation model approach, it also turns out that the estimator,

$$\bar{Y}_{\ell}'' = \frac{\sum_{i=1}^{n} Y_i}{n} + k\beta_1(\bar{X}' - \bar{X}) , \qquad (2.13)$$

with the value of k equal to one has the smallest mean squared error among estimators of the form given by equation (2.13). This implies that the optimal value of the weighting factor k given in equation (2.11) no longer holds when β_1 is known.

2.2.2. Numerical Comparisons of the Estimators, $\bar{Y}_S$, $\bar{Y}_C$, and $\bar{Y}_{\ell}^*$.

The optimal value for k (2.11) for a minimum MSE($\bar{Y}_{\ell}^*$) has been shown to be a function of the correlation between X and Y and the subsample size n. It is therefore of interest to examine the relative efficiencies of $\bar{Y}_C$, $\bar{Y}_{\ell}^*$, and the simple sample mean, $\bar{Y}_S$, for different values of the correlation coefficient ρ_{xy} and the subsample size n. Assuming 1/N to be negligible, the results are tabulated in Tables 1 to 4.

Tables 1a, 1b, and 1c show the relative efficiencies of the estimators as the subsample size n varies for a given initial sample size n' and a given correlation coefficient ρ_{xy}. The tables reveal that the relative efficiency of $\bar{Y}_{\ell}^*$ to $\bar{Y}_C$ decreases steadily, converging to one, as n increases for a given n' and correlation ρ_{xy}.

Tables 2a, 2b, and 2c show the relative efficiencies of the estimators under varying initial sample sizes for a fixed subsample size n and a given correlation ρ_{xy}. The relative efficiency of $\bar{Y}_{\ell}^*$ to both $\bar{Y}_C$ and $\bar{Y}_S$ increases with

larger initial sample sizes n' for a fixed n and a fixed ρ_{xy}.

Table 3 demonstrates how the relative efficiencies of the estimators behave with changes in ρ_{xy} for fixed n and n'. The estimator $\bar{Y}_{\ell}^{*}$ gains most over $\bar{Y}_C$ with respect to mean squared error when the correlation ρ_{xy} is small. On the other hand, when the correlation ρ_{xy} is large, $\bar{Y}_{\ell}^{*}$ is just as efficient as $\bar{Y}_C$ but a great deal more efficient than the simple sample mean, $\bar{Y}_S$. Whereas, $\bar{Y}_C$ can be less efficient than $\bar{Y}_S$ for low correlations and $\bar{Y}_S$ less efficient than $\bar{Y}_C$ for strong correlations, $\bar{Y}_{\ell}^{*}$, is never less efficient than either $\bar{Y}_S$ or $\bar{Y}_C$ for any value of the correlation ρ_{xy}. This property of $\bar{Y}_{\ell}^{*}$ makes $\bar{Y}_{\ell}^{*}$ a desirable alternative double sample regression estimator. Results indicate that theoretically $\bar{Y}_{\ell}^{*}$ is a compromise between $\bar{Y}_S$ and $\bar{Y}_C$. This property is also evident in Tables 1 and 2.

Table 4 gives the values of the optimal k for different combinations of subsample size n and correlation coefficient ρ_{xy}. It indicates that as n becomes larger, the optimal value for k approaches unity faster for higher correlations than for low correlations. This is consistent with previous conclusions attained through theoretical derivations.

In Tables 1 to 4, the correlation coefficient ρ_{xy} is assumed known, and the actual values of k are obtained. However, in practice, neither β_1 nor ρ_{xy} is usually known making it imperative to study the behavior of the different estimators when only estimates of ρ_{xy} and β_1 are available. A number of simulations have been run as an aid in determining how the different estimators compare to one another in practical situations.

The IMSL routine GGNPM was used to generate normal random variates. Bivariate normal random variates were then generated from these normal variates utilizing the following procedure. Given two independent normal random variates

Z_1 and Z_2 generated by GGNPM, 2 bivariate normals, X and Y, with correlation ρ_{xy}, are generated with $X = Z_1$ and $Y = \rho_{xy}Z_1 + \sqrt{1 - \rho_{xy}^2}\, Z_2$. Through transformations, $(X, Y) \sim BVN(30, 100, 2, 10, \rho_{xy})$ is generated. The mean squared errors of the different estimators when ρ_{xy} is estimated from the subsample were then computed using a thousand samples each time. Results are shown in Tables 5 and 6.

As expected, since k has to be estimated, $\bar{Y}_\ell^*$ does not do as well in Tables 5 and 6 as it did in Tables 1 and 3. There are now instances when $\bar{Y}_\ell^*$'s efficiency with respect to $\bar{Y}_C$ or $\bar{Y}_S$ has gone below one. However, there remains the same relationship between the correlation coefficient ρ_{xy} and the relative efficiencies of $\bar{Y}_\ell^*$ with respect to $\bar{Y}_C$ and $\bar{Y}_S$. Even with some loss of efficiency in $\bar{Y}_\ell^*$ due to estimating the value of k, the results from the simulation study suggest that $\bar{Y}_\ell^*$ is a viable, reasonable alternative to $\bar{Y}_C$.

Table 1a

Relative Efficiencies of the Estimators
for Increasing Subsample Sizes

$\rho_{xy} = 0.9$

n'	n	k	EFFC[1]	EFFY[2]	EFFYC[3]
50	5	0.895	1.026	2.877	2.805
	10	0.968	1.002	2.681	2.676
	15	0.981	~1.0	2.253	2.252
	20	0.986	~1.0	1.921	1.920
	25	0.989	~1.0	1.669	1.669
	30	0.991	~1.0	1.473	1.473
	40	0.994	~1.0	1.192	1.192
100	10	0.968	1.003	3.394	3.385
	20	0.986	~1.0	2.771	2.770
	30	0.991	~1.0	2.284	2.283
	40	0.994	~1.0	1.934	1.934
	50	0.995	~1.0	1.675	1.675
	60	0.996	~1.0	1.476	1.476
	80	0.997	~1.0	1.193	1.193
500	50	0.995	~1.0	3.641	3.641
	100	0.998	~1.0	2.828	2.828
	150	0.998	~1.0	2.305	2.305
	200	0.998	~1.0	1.943	1.943
	250	0.999	~1.0	1.680	1.680
	300	0.999	~1.0	1.479	1.479
	400	0.999	~1.0	1.193	1.193

$^{1}EFFC = \frac{MSE(\bar{Y}_C)}{MSE(Y^*_\ell)}$ $^{2}EFFY = \frac{MSE(\bar{Y}_S)}{MSE(Y^*_\ell)}$ $^{3}EFFYC = \frac{MSE(\bar{Y}_S)}{MSE(Y_C)}$

Table 1b

Relative Efficiencies of the Estimators
for Increasing Subsample Sizes

$\rho_{xy} = 0.5$

n'	n	k	EFFC[1]	EFFY[2]	EFFYC[3]
50	5	0.400	1.222	1.099	0.899
	10	0.700	1.030	1.163	1.129
	15	0.800	1.010	1.163	1.151
	20	0.850	1.005	1.146	1.141
	25	0.880	1.002	1.124	1.121
	30	0.900	1.001	1.099	1.099
	40	0.925	~1.0	1.048	1.048
100	10	0.700	1.034	1.187	1.148
	20	0.850	1.006	1.205	1.197
	30	0.900	1.002	1.187	1.184
	40	0.925	1.001	1.161	1.160
	50	0.940	1.001	1.133	1.133
	60	0.950	~1.0	1.105	1.105
	80	0.962	~1.0	1.050	1.050
500	50	0.940	1.001	1.268	1.267
	100	0.970	~1.0	1.241	1.240
	150	0.980	~1.0	1.207	1.207
	200	0.985	~1.0	1.173	1.173
	250	0.988	~1.0	1.141	1.141
	300	0.990	~1.0	1.110	1.110
	400	0.992	~1.0	1.052	1.052

Table 1c

Relative Efficiencies of the Estimators
for Increasing Subsample Sizes

$\rho_{xy} = 0.1$

n'	n	k	EFFC[1]	EFFY[2]	EFFYC[3]
50	5	0.020	1.437	~1.0	0.696
	10	0.066	1.106	1.001	0.905
	15	0.108	1.052	1.001	0.952
	20	0.147	1.030	1.001	0.972
	25	0.182	1.018	1.001	0.983
	30	0.214	1.012	1.001	0.989
	40	0.272	1.004	1.001	0.997
100	10	0.066	1.119	1.001	0.894
	20	0.147	1.040	1.001	0.963
	30	0.214	1.020	1.002	0.982
	40	0.272	1.012	1.002	0.990
	50	0.322	1.007	1.002	0.994
	60	0.365	1.004	1.001	0.997
	80	0.438	1.001	1.001	0.999
500	50	0.322	1.013	1.003	0.990
	100	0.495	1.004	1.004	~1.0
	150	0.598	1.002	1.004	1.002
	200	0.666	1.001	1.004	1.003
	250	0.714	1.001	1.004	1.003
	300	0.750	~1.0	1.003	1.003
	400	0.800	~1.0	1.002	1.002

Table 2a

Relative Efficiencies of the Estimators
for Increasing Sample Sizes

$\rho_{XY} = 0.9$

	n = 5		k = 0.895
n'	EFFC	EFFY	EFFYC
50	1.026	2.877	2.805
75	1.029	3.093	3.006
100	1.030	3.213	3.118

	n = 8		k = 0.955
n'	EFFC	EFFY	EFFYC
50	1.004	2.856	2.845
75	1.005	3.238	3.222
100	1.005	3.470	3.451

	n = 10		k = 0.968
n'	EFFC	EFFY	EFFYC
50	1.002	2.681	2.676
75	1.002	3.118	3.110
100	1.003	3.394	3.385
300	1.004	4.126	4.111
500	1.004	4.312	4.296
1000	1.004	4.462	4.445

	n = 20		k = 0.986
n'	EFFC	EFFY	EFFYC
50	~1.0	1.921	1.921
75	~1.0	2.415	2.415
100	~1.0	2.771	2.770
300	1.001	3.933	3.931
500	1.001	4.292	4.288
1000	1.001	4.608	4.604

Table 2b

Relative Efficiencies of the Estimators
for Increasing Sample Sizes

$\rho_{XY} = 0.5$

	n = 5		k = 0.40
n'	EFFC	EFFY	EFFYC
50	1.223	1.099	0.899
75	1.232	1.103	0.896
100	1.236	1.105	0.894

	n = 8		k = 0.625
n'	EFFC	EFFY	EFFYC
50	1.054	1.151	1.092
75	1.058	1.162	1.098
100	1.060	1.168	1.101

	n = 10		k = 0.7
n'	EFFC	EFFY	EFFYC
50	1.030	1.163	1.129
75	1.033	1.179	1.141
100	1.034	1.187	1.148
300	1.037	1.204	1.160
500	1.038	1.207	1.163
1000	1.038	1.210	1.165

	n = 20		k = 0.85
n'	EFFC	EFFY	EFFYC
50	1.005	1.146	1.141
75	1.006	1.185	1.178
100	1.006	1.205	1.197
300	1.008	1.247	1.238
500	1.008	1.256	1.246
1000	1.008	1.263	1.253

Table 2c

Relative Efficiencies of the Estimators
for Increasing Sample Sizes

$\rho_{XY} = 0.1$

	n = 5		k = 0.02
n'	EFFC	EFFY	EFFYC
50	1.437	~1.0	0.696
75	1.453	~1.0	0.688
100	1.461	~1.0	0.685

	n = 8		k = 0.048
n'	EFFC	EFFY	EFFYC
50	1.158	~1.0	0.864
75	1.168	~1.0	0.856
100	1.173	~1.0	0.853

	n = 10		k = 0.666
n'	EFFC	EFFY	EFFYC
50	1.105	1.001	0.905
75	1.115	1.001	0.898
100	1.119	1.001	0.894
300	1.128	1.001	0.887
500	1.130	1.001	0.886
1000	1.131	1.001	0.885

	n = 20		k = 0.147
n'	EFFC	EFFY	EFFYC
50	1.030	1.001	0.972
75	1.036	1.001	0.966
100	1.040	1.001	0.963
300	1.046	1.001	0.957
500	1.048	1.001	0.956
1000	1.049	1.001	0.955

Table 3

Relative Efficiencies of the Estimators with Different Correlation Coefficients for X and Y

	n = 5		n' = 50	
rho	k	EFFC	EFFY	EFFYC
0.9	0.895	1.026	2.877	2.805
0.8	0.780	1.065	1.817	1.706
0.7	0.658	1.111	1.409	1.268
0.6	0.529	1.164	1.207	1.037
0.5	0.400	1.223	1.099	0.899
0.4	0.276	1.285	1.041	0.810
0.3	0.165	1.347	1.014	0.753
0.2	0.077	1.400	1.003	0.716
0.1	0.020	1.437	~1.0	0.696

	n = 10		n' = 50	
rho	k	EFFC	EFFY	EFFYC
0.9	0.968	1.002	2.681	2.676
0.8	0.926	1.006	1.901	1.890
0.7	0.871	1.011	1.518	1.501
0.6	0.797	1.019	1.298	1.274
0.5	0.700	1.030	1.163	1.129
0.4	0.571	1.044	1.079	1.033
0.3	0.409	1.063	1.030	0.969
0.2	0.226	1.086	1.007	0.928
0.1	0.066	1.106	1.001	0.905

Table 4

Optimal Values of k for Different Values of n and Different Values of Correlation Coefficients

n/ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
5	.895	.780	.658	.529	.400	.276	.165	.077	.020
8	.955	.899	.828	.738	.625	.488	.331	.172	.048
10	.968	.926	.871	.797	.700	.571	.409	.226	.066
15	.981	.955	.920	.871	.800	.696	.543	.333	.108
20	.986	.968	.942	.905	.850	.764	.627	.415	.147
25	.989	.975	.955	.925	.880	.807	.685	.478	.182
30	.991	.980	.963	.938	.900	.837	.728	.529	.214
40	.994	.985	.973	.954	.925	.876	.785	.607	.272
50	.995	.988	.978	.964	.940	.900	.823	.662	.322
60	.996	.990	.982	.970	.950	.916	.849	.704	.365

Table 5a

Estimated Relative Efficiencies of the Estimators for Increasing Subsample Sizes (simulation)

$\rho_{XY} = 0.9$

n'	n	Ave. r	$\widehat{EFFC}^1$	$\widehat{EFFY}^2$	$\widehat{EFFYC}^3$
50	5	0.865	0.951	2.739	2.881
	10	0.892	0.989	2.718	2.749
	15	0.896	0.998	2.217	2.221
	20	0.892	~1.0	1.913	1.914
	25	0.892	~1.0	1.680	1.680
	30	0.896	~1.0	1.460	1.460
	40	0.898	~1.0	1.218	1.217
100	10	0.893	0.994	3.445	3.465
	20	0.895	~1.0	2.920	2.919
	30	0.896	~1.0	2.221	2.222
	40	0.898	~1.0	1.935	1.936
	50	0.898	~1.0	1.721	1.721
	60	0.898	~1.0	1.428	1.428
	80	0.900	~1.0	1.242	1.242
500	50	0.897	~1.0	3.824	3.824
	100	0.900	~1.0	2.688	2.688
	250	0.902	~1.0	1.680	1.680
	400	0.899	~1.0	1.166	1.166

$$^1\widehat{EFFC} = \frac{\widehat{MSE}(\bar{Y}_C)}{\widehat{MSE}(\bar{Y}^*_\ell)} \qquad ^2\widehat{EFFY} = \frac{\widehat{MSE}(\bar{Y}_S)}{\widehat{MSE}(\bar{Y}^*_\ell)} \qquad ^3\widehat{EFFYC} = \frac{\widehat{MSE}(\bar{Y}_C)}{\widehat{MSE}(\bar{Y}^*_\ell)}$$

Table 5b

Estimated Relative Efficiencies of the Estimators
for Increasing Subsample Sizes (simulation)

$\rho_{XY} = 0.5$

n'	n	Ave. r	$\widehat{EFFC}^1$	$\widehat{EFFY}^2$	$\widehat{EFFYC}^3$
50	5	0.452	1.154	1.021	0.855
	10	0.475	1.008	1.113	1.105
	15	0.482	0.984	1.134	1.152
	20	0.488	0.986	1.168	1.184
	25	0.486	0.998	1.132	1.134
	30	0.490	0.993	1.087	1.095
	40	0.493	~1.0	1.061	1.062
100	10	0.470	0.994	1.152	1.158
	20	0.482	0.995	1.182	1.188
	30	0.498	0.995	1.144	1.150
	40	0.495	0.991	1.203	1.214
	50	0.493	1.001	1.096	1.096
	60	0.498	~1.0	1.091	1.090
	80	0.495	~1.0	1.034	1.033
500	50	0.498	0.997	1.292	1.297
	100	0.499	0.998	1.263	1.265
	250	0.501	~1.0	1.162	1.162
	400	0.499	~1.0	1.042	1.042

Table 5c

Estimated Relative Efficiencies of the Estimators
for Increasing Subsample Sizes (simulation)

$\rho_{XY} = 0.1$

n'	n	Ave. r	$\widehat{EFFC}^1$	$\widehat{EFFY}^2$	$\widehat{EFFYC}^3$
50	5	0.025	1.161	0.816	0.703
	10	0.098	1.049	0.940	0.896
	15	0.110	1.017	0.993	0.977
	20	0.106	1.010	0.993	0.983
	25	0.103	1.009	0.991	0.982
	30	0.098	1.002	0.997	0.995
	40	0.096	1.001	~1.0	~1.0
100	10	0.100	1.056	0.963	0.912
	20	0.107	1.032	0.960	0.930
	30	0.101	1.006	~1.0	0.994
	40	0.100	~1.0	1.008	1.004
	50	0.101	1.004	1.004	~1.0
	60	0.104	1.006	0.993	0.987
	80	0.099	1.003	0.995	0.992
500	50	0.098	1.005	0.998	0.992
	100	0.099	0.999	0.997	0.997
	250	0.105	1.001	0.997	0.996
	400	0.101	0.999	1.004	1.005

Table 6

Estimated Relative Efficiencies of the Estimators with Different Correlation Coefficients for X and Y (simulation)

rho	ave. r	n = 5 $\hat{}$ EFFC	n' = 50 $\hat{}$ EFFY	$\hat{}$ EFFYC
0.9	0.865	0.951	2.739	2.881
0.8	0.756	1.001	1.646	1.644
0.7	0.654	1.040	1.322	1.272
0.6	0.534	1.084	1.065	0.982
0.5	0.452	1.154	1.021	0.885
0.4	0.366	1.171	0.998	0.852
0.3	0.267	1.172	0.873	0.745
0.2	0.184	1.155	0.854	0.740
0.1	0.114	1.161	0.816	0.703

rho	ave. r	n = 10 $\hat{}$ EFFC	n' = 50 $\hat{}$ EFFY	$\hat{}$ EFFYC
0.9	0.892	0.989	2.718	2.749
0.8	0.786	0.997	1.788	1.793
0.7	0.678	0.979	1.554	1.587
0.6	0.590	~1.0	1.275	1.275
0.5	0.475	1.008	1.113	1.105
0.4	0.387	1.007	1.064	1.056
0.3	0.291	1.037	0.971	0.937
0.2	0.172	1.036	0.955	0.922
0.1	0.098	1.049	0.940	0.896

2.2.3. Comparison of the Optimal Weighting Factor k for the Conditional and the Unconditional Mean Squared Error of $\bar{Y}^*_\ell$

Thus far, only the performance of $\bar{Y}^*_\ell$ using the optimal k with respect to its unconditional mean squared error has been compared to $\bar{Y}_C$ and $\bar{Y}_S$. $\bar{Y}^*_\ell$ with the unconditional k_u, estimated or not, has been seen to compare favorably to both $\bar{Y}_C$ and $\bar{Y}_S$. A question then arises. If the optimal value of k with respect to the unconditional mean squared error is different from that of the conditional mean squared error, given the initial sample, which optimal value of k would give a better estimator with respect to average squared error?

Theorem 2.3

The optimal conditional $MSE(\bar{Y}^*_\ell | x_1, x_2, \ldots, x_{n'})$ averaged over the x's is less than or equal to the optimal unconditional $MSE(\bar{Y}^*_\ell)$.

Proof: Denote $\hat{\bar{Y}}^*_\ell$ by $\hat{\bar{Y}}(k)$ where k is the weighting factor.

Let

$$\min_k E_{x_{n'}}\{E_{y|x_{n'}}[(\hat{\bar{Y}}(k) - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]\}$$

$$= E_{x_{n'}}\{E_{y|x_{n'}}[(\hat{\bar{Y}}(k_u) - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]\} .$$

Since

$$\min_k E_{y|x_{n'}}[(\hat{\bar{Y}}(k) - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$$

$$\leq E_{y|x_{n'}}[(\hat{\bar{Y}}(k_u) - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}] ,$$

it follows that

$$\int_{\underline{x}_{n'}} \min_{k} E_{y|\underline{x}_{n'}}[(\hat{\bar{Y}}(k) - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]dF(\underline{x}_{n'})$$

$$\leq \int_{\underline{x}_{n'}} E_{y|\underline{x}_{n'}}[(\hat{\bar{Y}}(k_u) - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]dF(\underline{x}_{n'})$$

$$= \min_{k} E[(\hat{\bar{Y}}(k) - \bar{Y})^2] \ .$$

For the discrete case, the same logic applies except summation is used instead of integration.

Simulations were also run to compare the optimal conditional $MSE(\bar{Y}^*_\ell | x_1, x_2, \ldots, x_{n'})$ against the optimal unconditional $MSE(\bar{Y}^*_\ell)$ for both cases: i) when the optimal value of k is known, and ii) when the optimal value of k is estimated from the subsample. Normal bivariate random variables were generated as before. The conditional $MSE(\bar{Y}^*_\ell | x_1, x_2, \ldots, x_{n'})$ was then estimated from 100 subsamples from a single initial sample. One thousand initial samples were considered for each combination of n, n', and ρ_{xy} when n' = 50 or n' = 100. Otherwise, five hundred initial samples were considered.

Simulation results in Table 7 indicate that when the true values of the optimal k can be obtained, it is clearly advantageous to use the optimal k conditioned on the x's, especially when there is a strong correlation between X and Y. The apparent advantage of the conditional optimal k over the unconditional optimal k increases as the subsample size n gets larger with respect to the initial sample size n'. It is also worth noting that $\bar{Y}^*_\ell$ conditioned on the x's can be three times more efficient than $\bar{Y}^*_\ell$ using the unconditional optimal k. However, this superiority manifested by the conditional optimal k to

the unconditional optimal k dissolves when ρ_{XY} has to be estimated. Table 8 indicates that there is practically no advantage in using one over the other with respect to mean squared error. Accordingly, the unconditional optimal k will most likely be preferred over the conditional optimal k by reason of its being easier to compute.

Table 7

Estimated Relative Efficiency[1] of $\bar{Y}^*_\ell$ using the Conditional Optimal k to $\bar{Y}^*_\ell$ using the Unconditional Optimal k (true values of k)

n'	n	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$
50	5	1.3299	1.0433	1.0004
	10	1.6861	1.0470	1.0004
	25	2.8953	1.1557	1.0004
100	10	1.3636	1.0280	1.0001
	20	1.8541	1.0509	1.0005
	50	3.1148	1.1520	1.0037
500	50	1.3900	1.0248	1.0009
	100	1.8551	1.0537	1.0002
	250	3.3537	1.2094	1.0021

[1]Est. Rel. Efficiency = $\dfrac{\widehat{MSE}(\bar{Y}^*_\ell \text{ using } k_U)}{\widehat{MSE}(\bar{Y}^*_\ell \text{ using } k_C)}$

Table 8

Estimated Relative Efficiency of $\bar{Y}_{\ell}^{*}$ using the Conditional Optimal k to $\bar{Y}_{\ell}^{*}$ using the Unconditional Optimal k (estimated values of k)

n'	n	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$
50	5	1.0078	1.0346	1.0528
	10	0.9953	1.0033	1.0032
	25	0.9999	0.9998	1.0001
100	10	0.9927	1.0044	1.0032
	20	0.9996	~1.0	1.0007
	50	~1.0	0.9999	1.0001
500	50	~1.0	0.9998	~1.0
	100	~1.0	~1.0	~1.0
	250	~1.0	~1.0	~1.0

III. PURPOSIVE SAMPLING

3.1 DESIGNING SUBSAMPLES GIVEN THE INITIAL SAMPLE

The main thrust of this dissertation is to minimize the mean squared error of a regression estimator in double sampling. Another strategy to improve the conditional mean squared error of the estimators $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ when given the initial sample will be explored in this chapter. It employs a subsampling plan that is nonrandom. This idea is not new. Royall (1970) stated in his paper, "If the sampler believes it to be important that he obtain a sample in which the x values have a certain configuration than he should choose such a sample deliberately."

We first consider $\bar{Y}_{C}$, the double sample regression estimator found in most sampling textbooks. Assuming the hypothesized superpopulation model $\xi[1, 1: 1]$ is true,

$$MSE(\bar{Y}_{C}|x_1, x_2, \ldots, x_{n'})$$

$$= \sigma_{\epsilon}^{2}\left[\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right] + \beta_1^{2}(\bar{x}' - \bar{x})^2 . \qquad (3.1)$$

Since the population size N is almost always a large number, 1/N can often be regarded as negligible.

Looking at formula (3.1), an obvious way to minimize $MSE(\bar{Y}_{C}|x_1, x_2, \ldots, x_{n'})$ is to make the subsample size n as large as possible. That is precisely the problem. We assume that the y's are relatively expensive to sample, thereby restricting the subsample size n. However, one can get a handle on

the value of $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$. In order to make the mean squared error of $\bar{y}_c$ when given the initial sample small, $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ should be made as small as possible for a given initial sample.

When sampling is done without replacement for the subsample, there are $\binom{n'}{n}$ possible subsamples of size n for a given initial sample of size n'. One can only be absolutely sure of attaining the minimum $MSE(\bar{y}_c | x_1, x_2, \ldots, x_{n'})$ by looking at all possible subsamples and choosing the subsample with the smallest $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ value. In practice this turns out to be a very laborious task, even for a relatively small initial sample of size n' = 50. Such a procedure would most likely be rejected by sampling practitioners.

Herson (1976) and Iachan (1985) used a form of restricted random sampling to achieve the desired configuration of x's needed in robust designs. The method of restricted random sampling can also be applied to the problem of designing subsamples in double sampling. Restricted random sampling can be used to obtain a subsample that yields a mean squared error in the region of the minimum $MSE(\bar{y}_c | x_1, x_2, \ldots, x_{n'})$. The scheme can proceed as follows:

i) Select a subsample s by simple random sampling.

ii) Accept s if $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} < \delta$, where δ is some prespecified small number; otherwise, go back to i).

Depending upon how stringent the acceptance criterion is, the scheme will

involve repeated sampling until an acceptable sample is obtained.

This research takes designing subsamples a step further through the use of an algorithm that will yield a unique subsample having a value of $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ which approximates the smallest value of $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ possible for the given initial sample. The algorithm enables one to avoid having to sample more than once and yet be assured of a subsample with a mean squared error that is near minimum.

Algorithm:

Aim: To yield a subsample with $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ very close to the minimum value possible for the given initial sample.

Procedure:

i) Get the subsample with $\bar{x}$ as close to $\bar{x}'$ as possible, and

ii) Select the x's in the subsample with the largest variation possible.

The algorithm is translated into the FORTRAN program called DESIGN found in Appendix I. The program picks a subsample of x's at the extremes of the distribution of the initial sample, roughly symmetric about $\bar{x}'$.

For a given initial sample, the subsample obtained by the algorithm was compared with one thousand simple random subsamples from the same initial sample with respect to their values of $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$. This was done for various combinations of n and n' and different initial samples. The algorithm outperforms at least 80% of the thousand simple random subsamples obtained from the same initial sample. For example, for a given initial sample with n' = 500

and $n \leq 250$, roughly 95% of the simple random subsamples yielded values of $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ greater than for the subsample selected by the algorithm. This suggests that the algorithm is useful in a practical sense.

Assuming the same superpopulation model $\xi[1, 1: 1]$, the conditional mean squared error of $\bar{y}_\ell^*$ is of the form

$$MSE(\bar{y}_\ell^* | x_1, x_2, \ldots, x_{n'}) = \sigma_\epsilon^2 \left[\left(\frac{1}{n} - \frac{1}{N} \right) + \frac{k^2 (\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + \beta_1^2 [\{(1 - k)\bar{x} + k\bar{x}'\} - \bar{x}]^2 . \tag{3.2}$$

If k is treated as a constant, the same condition that minimizes $MSE(\bar{y}_C | x_1, x_2, \ldots, x_{n'})$ minimizes $MSE(\bar{y}_\ell^* | x_1, x_2, \ldots, x_{n'})$. That is, one should make the quantity $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ as small as possible.

From formula (3.2), the conditional squared bias of $\bar{y}_\ell^*$ is $\beta_1^2[\{(1 - k)\bar{x} + k\bar{x}'\} - \bar{x}]^2$. If $\bar{x}'$ is indeed a good estimate of $\bar{x}$, then the squared bias of $\bar{y}_\ell^*$ given $x_1, x_2, \ldots, x_{n'}$ is approximately $\{\beta_1(1 - k)(\bar{x} - \bar{x}')\}^2$. Therefore, if $\bar{x}' \cong \bar{x}$ then $E[(\bar{y}_\ell^* - \bar{y})^2 | x_1, x_2, \ldots, x_{n'}]$ is reduced by minimizing $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ with an emphasis on $|\bar{x} - \bar{x}'|$ being as small as possible.

When k is not treated as a constant and the optimal k is used in the estimator $\bar{y}_\ell^*$, then under the model $\xi[1, 1: 1]$, the optimal $MSE(\bar{y}_\ell^* | x_1, x_2, \ldots, x_{n'})$ is

$$E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_\epsilon^2 \left[\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\beta_1^2 (\bar{X} - \bar{x})^2}{\sigma_\epsilon^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2} \right] . \qquad (3.3)$$

Referring to formula (3.3), since one does not have control over the value $(\bar{X} - \bar{x})$, the only means of reducing the optimal $MSE(\bar{Y}_\ell^* | x_1, x_2, \ldots, x_{n'})$ through the design of subsamples is to select the subsample with the largest variation,

or the maximum value of $\sum_{i=1}^{n} (x_i - \bar{x})^2$.

As before, simulation runs were used to provide an indication of how designed subsamples affect the mean squared errors of both $\bar{Y}_C$ and $\bar{Y}_\ell^*$, given the initial sample. Using the same routine described in Chapter 2, X and Y are generated from a bivariate normal distribution BVN(30, 100, 2, 10, ρ_{xy}). One thousand different initial samples were considered for n' = 50 and n' = 100. Only five hundred different initial samples were considered for n' = 500 due to the amount of CPU time that would be required for the job. A unique subsample was obtained using the algorithm for each initial sample while two hundred simple random subsamples were obtained from each initial sample. The conditional mean squared errors of the estimators, $\bar{Y}_C$ and $\bar{Y}_\ell^*$, given the initial sample, were then estimated from the two hundred simple random subsamples. These estimated conditional mean squared errors of $\bar{Y}_C$ and $\bar{Y}_\ell^*$, averaged over one thousand initial samples, were then compared to the mean squared errors of $\bar{Y}_C$ and $\bar{Y}_\ell^*$ yielded by the designed subsamples. Although the criterion for the design of subsamples depends on whether or not k is treated as a constant, the simulation runs only deal with the case treating k as a constant. Because of the way the

algorithm was set up to approximate the minimum value of

$\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ for a given initial sample, the two different criteria resulted in very similar subsamples if not the same. Results comparing designed subsamples against simple random subsamples are shown in Tables 9-12. For these simulations, ρ_{XY} was not assumed known, and hence, was estimated from the subsamples.

Tables 9 and 10 show the relative efficiencies of $\bar{Y}_C$ and $\bar{Y}^*_\ell$, respectively, for designed subsamples compared to simple random subsamples. The tables indicate that there is significant gain in efficiency for both $\bar{Y}_C$ and $\bar{Y}^*_\ell$ in using designed subsamples when the initial sample size n' is small. It should also be noted that the improvement in efficiency for designed subsamples is more consistent for $\rho_{XY} = 0.1$ than for other values of ρ_{XY} where its performance is more erratic. This suggests that designed subsamples should be used over simple random subsamples for either estimator $\bar{Y}_C$ or $\bar{Y}^*_\ell$ when the initial sample size n' is small and/or when one believes the correlation to be quite weak.

Table 11 shows that when designed subsamples are used, there is no advantage in using either $\bar{Y}_C$ or $\bar{Y}^*_\ell$ over the other with respect to their conditional mean squared errors given $x_1, x_2, \ldots, x_{n'}$. However, since $\bar{Y}^*_\ell$ involves the additional computation for the optimal k, $\bar{Y}_C$ would most likely be the preferred estimator for designed subsamples.

An interesting conclusion follows from the previous results. The estimator, $\bar{Y}_C$, with designed subsamples, is a better strategy than $\bar{Y}^*_\ell$ with simple random subsamples for a small initial sample size n' or a small correlation ρ_{XY}. Relative efficiencies are tabulated in Table 12. For small correlations, if one

were to rank the four sampling plans in descending order of preference, they would be as follows: 1) $\bar{Y}_c$ with designed subsamples, 2) $\bar{Y}_\ell^*$ with designed subsamples, 3) $\bar{Y}_\ell^*$ with simple random subsamples, and last, 4) $\bar{Y}_c$ with simple random subsamples.

Table 9

Estimated Relative Efficiency of $\bar{Y}_c$ for Designed Subsamples Versus Simple Random Subsamples

n'	n	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$
50	5	1.2218	1.5434	1.3902
	10	1.0119	1.0968	1.1113
	25	1.0464	1.0499	1.0740
100	10	1.1447	1.0985	1.1709
	20	0.9743	1.0152	1.0243
	50	1.0403	0.9550	0.9984
500	50	0.9679	1.1455	1.0707
	100	0.9816	0.9585	1.0753
	250	0.9963	0.9476	1.0439

Table 10

Estimated Relative Efficiency of $\bar{Y}_{\ell}^{*}$ for Designed Subsamples Versus Simple Random Subsamples

n'	n	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$
50	5	1.2717	1.3334	1.1237
	10	1.0252	1.0936	1.0507
	25	1.0468	1.0553	1.0627
100	10	1.1653	1.0931	1.1014
	20	0.9762	1.0255	1.0024
	50	1.0412	0.9565	0.9945
500	50	0.9681	1.1502	1.0629
	100	0.9816	0.9591	1.0733
	250	0.9963	0.9476	1.0439

Table 11

Estimated Mean Squared Error of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ for Designed Subsamples

		$\rho = 0.9$		$\rho = 0.5$		$\rho = 0.1$	
n'	n	$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$
50	5	5.879	5.880	14.512	14.534	20.981	20.844
	10	3.711	3.711	8.156	8.156	10.044	10.037
	25	2.176	2.176	3.313	3.314	3.768	3.769
100	10	2.578	2.578	7.876	7.877	9.719	9.720
	20	1.897	1.897	4.053	4.053	5.069	5.072
	50	1.174	1.174	1.840	1.843	2.033	2.033
500	50	0.589	0.589	1.401	1.401	1.852	1.852
	100	0.392	0.392	0.856	0.856	0.953	0.953
	250	0.243	0.243	0.370	0.370	0.376	0.376

Table 12

Estimated Relative Efficiency of $\bar{Y}_C$ for Designed Subsamples to $\bar{Y}^*_\ell$ for Simple Random Subsamples

n'	n	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$
50	5	1.2719	1.3354	1.1212
	10	1.0252	1.0936	1.0500
	25	1.0469	1.0554	1.0629
100	10	1.1654	1.0932	1.1015
	20	0.9762	1.0255	1.0029
	50	1.0412	0.9565	0.9944
500	50	0.9681	1.1502	1.0630
	100	0.9816	0.9591	1.0733
	250	0.9963	0.9476	1.0439

3.2 DESIGNING SUBSAMPLES WITHIN A REGION OF POSSIBLE INITIAL SAMPLES

Thus far, only the conditional mean squared error of $\bar{Y}_C$ and $\bar{Y}^*_\ell$, given the initial large sample, have been discussed. Since the initial sample is a simple random sample from the population, there exists a huge set of possible initial samples indicating different values of $\bar{x}'$. It therefore seems to be appropriate to consider an average conditional mean squared error of $\bar{Y}_C$, averaged over a region of possible initial samples. A method used in response surface methodology is similarly employed here. (See Myers, 1976.)

The estimator, $\bar{Y}_C = \frac{\sum_{i=1}^{n} Y_i}{n} + \hat{\beta}_1(\bar{x}' - \bar{x})$, is simply a prediction at the point $\bar{x}'$ (i.e. $\bar{Y}_C = \hat{Y}(\bar{x}')$). Define the average $MSE(\bar{Y}_C|\bar{x}')$ to be

$$J = \frac{nc}{\sigma_\epsilon^2} \int_R E(\hat{Y}(\bar{x}') - \bar{Y})^2 \, d\bar{x}' \,, \qquad (3.4)$$

where

$$c = \frac{1}{\int_R d\bar{x}'} \,.$$

Suppose the x's are coded such that the $(\bar{x}')$'s of interest are uniformly distributed in the region R = [-1, 1], then

$$c = \frac{1}{\int_{-1}^{1} d\bar{x}'} = \frac{1}{2} \,.$$

Similarly,

$$J = \frac{nc}{\sigma_\epsilon^2} \int_R E[\hat{Y}(\underline{x}') - E\hat{Y}(\underline{x}')]^2 d\underline{x}' + \frac{nc}{\sigma_\epsilon^2} \int_R [E\hat{Y}(\underline{x}') - \bar{Y}]^2 \, d\underline{x}'$$

$$= \text{Variance} + (\text{Bias})^2 \, . \qquad (3.5)$$

where

$$\text{Variance} = \frac{nc}{\sigma_\epsilon^2} \int_R \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 \underline{x}') d\underline{x}'$$

$$= \frac{n}{2\sigma_\epsilon^2} \int_{-1}^{1} \left(\frac{\sigma_\epsilon^2}{n} + \frac{(\underline{x}')^2 \sigma_\epsilon^2}{n[11]} \right) d\underline{x}'$$

$$= 1 + \frac{1}{3[11]} \, ,$$

where

$[11] = \dfrac{\sum_{j=1}^{n} x_{cj}^2}{n}$, and x_{cj} is the coded x value of the jth element in the subsample, and

$$(\text{Bias})^2 = \frac{nc}{\sigma_\epsilon^2} \int_R \beta_1 (\underline{x}' - \bar{\underline{x}})^2 \, d\underline{x}'$$

$$= \frac{n\beta_1^2}{2\sigma_\epsilon^2} \int_{-1}^{1} (\underline{x}' - \bar{\underline{x}})^2 \, d\underline{x}'$$

$$= \frac{n\beta_1^2}{\sigma_\epsilon^2} \left(\frac{1}{3} + \bar{\underline{x}}^2 \right) .$$

Since the squared bias is a function of population parameters, only the variance portion can be reduced by designing subsamples to make the average conditional $MSE(\bar{Y}_c | \underline{x}')$ smaller. To reduce the variance, we should design so that the second moment [11] of the coded x's is as large as possible. If the weighting

factor k in $\bar{Y}_{\ell}^{*}$ is treated as a constant, the same design will minimize the average $MSE(\bar{Y}_{\ell}^{*}|\mathbf{x}')$. These results indicate that one should take a subsample of x's at the extremes for any initial sample within the region. It is worth noting that for the cases considered in this chapter, whether one considers a single initial sample or a region of initial samples, the subsamples designed to reduce the conditional mean squared errors of $\bar{Y}_{c}$ and $\bar{Y}_{\ell}^{*}$ given the initial sample when the assumed model is correct involves selecting elements with x values at the extremes.

IV. DESIGNING SUBSAMPLES WITH PROTECTION AGAINST MODEL MISSPECIFICATION

4.1 ANALYTICAL DERIVATION

Chapter three deals with subsamples deliberately chosen to minimize the conditional mean squared error of $\bar{Y}_C$ and $\bar{Y}_\ell^*$ given the initial sample under an assumed superpopulation model. One major criticism against the subsample designs developed in chapter three is that they are sensitive to model misspecification. Deviations of the assumed model from the true model could seriously inflate error. This chapter addresses the problem by developing designs for subsampling that provide some protection against model misspecification.

Let $Z_i = X_i - \bar{X}$. The assumed model is $Y_i = \beta_0 + \beta_1 Z_i + \epsilon_i$ when the true model is actually $Y_i = \beta_0 + \beta_1 Z_i + \beta_{11} Z_i^2 + \epsilon_i$, where

$$E(\epsilon_i | x_i) = 0 ,$$

$$E(\epsilon_i^2 | x_i) = \sigma_\epsilon^2 ,$$

and

$$E(\epsilon_i \epsilon_j | x_i, x_j) = 0 .$$

Then the $MSE(\bar{Y}_C | x_1, x_2, \ldots, x_{n'})$ can be shown to be

$$E[(\bar{Y}_C - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_\epsilon^2 \left[\left(\frac{1}{n} - \frac{1}{N} \right) + \frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + \beta_1^2 (\bar{x}' - \bar{X})^2$$

$$+ \beta_{11} F [\beta_{11} F + 2\beta_1 (\bar{x}' - \bar{X})] , \qquad (4.1)$$

where

$$F = \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} - \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} + (\bar{x}' - \bar{x}) \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] .$$

The estimator $\bar{Y}_C$ has a larger mean squared error because of the bias increase due to model misspecification. In order to reduce the bias portion of $E[(\bar{Y}_C - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$, the factor F in formula (4.1) can be made to approach zero by making

$$\sum_{i=1}^{n} (x_i - \bar{x})^3 \rightarrow \text{zero} ,$$

and

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} - \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} .$$

Making $\sum_{i=1}^{n} (x_i - \bar{x})^3$ close to zero implies that a subsample of x's symmetric about $\bar{x}$ is desired. The quantity $\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$ is unknown but a consistent estimator for it should be $\frac{\sum_{i=1}^{n'} (x_i - \bar{x})^2}{n'}$ since the initial sample is a simple random sample from the population. Hence, $\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$ should be made to approach $\frac{\sum_{i=1}^{n'} (x_i - \bar{x})^2}{n'}$. In order to

reduce the variance, we want to make $\frac{(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ as small as possible.

When trying to minimize the mean squared error, one needs to find a compromise between the criteria for minimizing variance and for minimizing squared bias. However if the main interest is to protect $\bar{Y}_C$ against a second-order model, subsamples should be chosen such that the factor $F \to 0$. (See equation (4.1).)

Under the same conditions, the conditional mean squared error of $\bar{Y}_\ell^*$ given the initial sample can be shown to be

$$E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_\epsilon^2 \left[\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{k^2(\bar{x}' - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] + \left[\beta_1 \{[1 - k)\bar{x} + k\bar{x}'] - \bar{X}\} \right.$$

$$\left. + \beta_{11} \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} - \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} + k(\bar{x}' - \bar{x}) \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \right]^2 . \quad (4.2)$$

If k is treated as some known constant, the conditions needed to minimize $E[(\bar{Y}_C - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$, (4.1), will be the same conditions necessary to minimize $E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$, (4.2).

When one assumes the model $Y_i = \beta_0 + \beta_1 Z_i + \epsilon_i$ to be correct and hence uses the optimal value for k (formula (2.4)) based on that assumption, then the mean squared error of $\bar{Y}_\ell^*$ given the initial sample can be written as

$$E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_\epsilon^2\left[\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\beta_1^4(\bar{X} - \bar{\mathcal{X}})^2 \sum_{i=1}^{n}(x_i - \bar{\mathcal{X}})^2}{(\sigma_\epsilon^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{\mathcal{X}})^2)^2}\right] + \tag{4.3}$$

$$\left[\beta_1(\bar{X}-\bar{\mathcal{X}})\left\{\frac{\beta_{11}\beta_1 \sum_{i=1}^{n}(x_i-\bar{\mathcal{X}})^3 - \sigma_\epsilon^2}{\sigma_\epsilon^2 + \beta_1^2 \sum_{i=1}^{n}(x_i-\bar{\mathcal{X}})^2}\right\} + \beta_{11}\left\{\frac{\sum_{i=1}^{n}(x_i-\bar{\mathcal{X}})^2}{n} - \frac{\sum_{i=1}^{N}(x_i-\bar{\mathcal{X}})^2}{N}\right\}\right]^2,$$

when the true model is actually $Y_i = \beta_0 + \beta_1 Z_i + \beta_{11} Z_i^2 + \epsilon_i$. It is interesting to note, however, that when the subsample is designed with protection against a second order polynomial, the $E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$, formula (4.3), reduces to the optimal $E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$ when the assumed model is correct. (See formula (2.8).) This indicates that when one is able to design subsamples with protection against a second order model exactly, the actual presence of a true second order model will have no effect whatsoever on the conditional mean squared error of $\bar{Y}_\ell^*$, an estimator based on a simple linear model. This result shows that even if model misspecification did exist, designing subsamples with protection can make it appear as if no model misspecification has occurred.

When the true model is a second order polynomial, the conditional mean squared error of $\bar{Y}_\ell^*$ given the initial sample can be shown to be that given by formula (4.2). The optimal value of k based on the assumption of a first order model (see formula (2.4)) no longer holds as being optimal when the true model is a second order model. A new value for k is needed to minimize $E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$ in equation (4.2). Using the usual calculus techniques, this new optimal value for k can be shown to be equal to

$$k_2 = \frac{\left[\beta_1 + \beta_{11}\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\left[\beta_1(\bar{X}-\bar{x}) + \beta_{11}\left\{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} - \dfrac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}\right\}\right]}{(\bar{x}'-\bar{x})\left[\dfrac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \left(\beta_1 + \beta_{11}\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2\right]} \qquad (4.4)$$

Again, if the subsample is selected such that the conditions that provide protection against a second order model are satisfied, k_2 reduces to the optimal value for k (formula (2.4)) as if no model misspecification has occurred. Thus, one need not worry about a different optimal value for k when the true model is a second order model as long as one designs subsamples with protection against a second order model. When one designs subsamples with protection against a second order model, the k that was optimal under the first order model will also be optimal under a second order model.

The above results, arrived at by applying the subsample design criteria for protection against a second order polynomial to equations (4.3) and (4.4), substantiate the assertion that $\bar{Y}_\ell^*$ can be rendered robust against model misspecification through the use of specific designs for subsampling. In fact, if subsamples designed with protection against a second order model can be exactly implemented, then $\bar{Y}_\ell^*$ and $\bar{Y}_c$ will be robust to the presence of a second order model.

4.2 SIMULATION STUDY COMPARING THE THREE SUBSAMPLING PROCEDURES

4.2.1 Performance under a Bivariate Normal Population

In practice, the design conditions for protection against a second order

polynomial can be approximated by a subsample obtained through the use of the subroutine called PROTEC found in Appendix I.

Simulations were run to determine how well the designs for protection against a second order model work for both $\bar{Y}_C$ and $\bar{Y}^*_{\ell}$. These results are compared to results obtained from designs without protection and to simple random subsamples when the true model is a second order polynomial.

The performances of these three different subsampling selection schemes were evaluated for three second order polynomials for each value of $\rho_{XY} = 0.9$, 0.5, and 0.1. The X variable is generated from a N(30, 2) distribution and the ϵ's are generated from a $N(0, \sigma_\epsilon)$ distribution where $\sigma_\epsilon^2 = 100(1 - \rho_{XY}^2)$. We arbitrarily chose β_{11}, the quadratic coefficient, to have the following magnitudes: 0.1, 0.5, and 1.0, depicting increasing curvature. (See Figures A.1, A.2, and A.3 in Appendix II.) We have also chosen to let the polynomial curve downward instead of upward. However, we specify β_0 and β_1 to be values such that the expected values of Y at the extremes (3 standard deviations from the mean), X = 24 and X = 36, when $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$ are equal to the expected values of Y' at the same X values when

$$Y' = \beta_0' + \beta_1' X + \epsilon \ ,$$

where

$$\beta_1' = 5\rho_{XY} \ ,$$

and

$$E[Y'|X = 30] = 100 \ .$$

For $\rho_{XY} = 0.9$, the Y variate is generated from the following second order polynomials:

i) $Y_i = -121.4 + 10.5X_i - 0.1X_i^2 + \epsilon_i$,

ii) $Y_i = -467.0 + 34.5X_i - 0.5X_i^2 + \epsilon_i$,

and

iii) $Y_i = -899.0 + 64.5X_i - 1.0X_i^2 + \epsilon_i$.

For $\rho_{xy} = 0.5$, the Y variable is generated from:

i) $Y_i = -61.4 + 8.5X_i - 0.1X_i^2 + \epsilon_i$,

ii) $Y_i = -407.0 + 32.5X_i - 0.5X_i^2 + \epsilon_i$,

and

iii) $Y_i = -839.0 + 62.5X_i - 1.0X_i^2 + \epsilon_i$.

For $\rho_{xy} = 0.1$, the Y variable is generated from:

i) $Y_i = -1.4 + 6.5X_i - 0.1X_i^2 + \epsilon_i$,

ii) $Y_i = -347.0 + 30.5X_i - 0.5X_i^2 + \epsilon_i$,

and

iii) $Y_i = -779.0 + 60.5X_i - 1.0X_i^2 + \epsilon_i$.

One thousand different initial samples and two hundred subsamples from each initial sample were considered for each combination of n and n' for every polynomial. The conditional mean squared errors were estimated from the 200 subsamples from each initial sample. Results showing the relative efficiencies of the three subsample selection schemes, (1) simple random sampling, (2) design without protection, and (3) design with protection are tabulated in Tables 13a, b, and c. Table 14 compares $\bar{Y}_\ell^*$ to $\bar{Y}_C$ under the subsampling plan which affords protection against a second order model.

It is observed from Tables 13a to c that designing subsamples to minimize the conditional mean squared errors of $\bar{Y}_\ell^*$ and $\bar{Y}_C$ when given the initial sample as discussed in Chapter 3, without regard for model misspecification, typically yields serious consequences when the assumed model is incorrect. When the true

model is a second order model, the actual conditional mean squared errors of both $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ from designed subsamples without protection can be inflated by as much as sixteen times over the conditional mean squared errors that would have been obtained through simple random subsampling. This demonstrates that the objections in the literature towards designing subsamples based on simple linear models are well founded. In fact, the sharper the curvature, the bigger the loss due to designing subsamples without protection. Results clearly support the need for some type of robust designs.

This chapter has outlined a subsampling scheme that protects $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ against a second order model. Tables 13a to c indicate that the protection subsampling plan yields subsamples that give up to 70% gain in efficiency over simple random subsamples when the true model is a second order model. Under the same model, these subsamples are also shown to be definitely superior to the subsamples designed without protection. Table 14 demonstrates that there is practically no difference in efficiency between $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ when the subsamples are chosen to protect against a second order model and the true model is in fact a second order model.

Table 13a

Estimated Relative Efficiencies of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ under Different Subsampling Schemes

$\rho = 0.9$

n/n'	Model	MSE(DES)/MSE(SRS) $\bar{Y}_C$	MSE(DES)/MSE(SRS) $\bar{Y}^*_\ell$	MSE(SRS)/MSE(PRO) $\bar{Y}_C$	MSE(SRS)/MSE(PRO) $\bar{Y}^*_\ell$	MSE(DES)/MSE(PRO) $\bar{Y}_C$	MSE(DES)/MSE(PRO) $\bar{Y}^*_\ell$
6/50	i	1.2146	1.1628	1.1289	1.1791	1.3711	1.3711
	ii	7.3097	7.1467	1.0695	1.0938	7.8178	7.8173
	iii	12.9608	13.0236	0.9626	0.9579	12.4767	12.4748
10/50	i	1.3022	1.2853	1.0381	1.0517	1.3518	1.2853
	ii	6.2630	5.7964	1.0750	1.0805	6.2647	6.2631
	iii	8.9905	9.0228	1.0296	1.0261	9.2561	9.2580
10/100	i	1.9921	1.9562	0.9880	1.0061	1.9682	1.9682
	ii	11.4713	11.4014	0.9820	0.9879	11.2643	11.2640
	iii	16.6781	16.7395	0.9350	0.9316	15.5945	15.5939
20/100	i	1.6986	1.6962	1.0356	1.0370	1.7590	1.7590
	ii	7.4514	7.4555	1.0269	1.0263	7.6520	7.6517
	iii	9.6914	9.7172	1.0651	1.0625	10.3224	10.3241
50/100	i	1.3173	1.3173	1.0871	1.0874	1.4321	1.4325
	ii	2.8997	2.9002	1.5305	1.5308	4.4380	4.4396
	iii	3.2951	3.2965	1.7015	1.7009	5.6066	5.6069

Table 13b

Estimated Relative Efficiencies of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ under Different Subsampling Schemes

$\rho = 0.5$

n/n'	Model	MSE(DES)/MSE(SRS) $\bar{Y}_C$	MSE(DES)/MSE(SRS) $\bar{Y}^*_\ell$	MSE(SRS)/MSE(PRO) $\bar{Y}_C$	MSE(SRS)/MSE(PRO) $\bar{Y}^*_\ell$	MSE(DES)/MSE(PRO) $\bar{Y}_C$	MSE(DES)/MSE(PRO) $\bar{Y}^*_\ell$
6/50	i	0.9325	1.0014	1.2837	1.1962	1.1970	1.0014
	ii	3.6924	3.9333	1.2404	1.1633	4.5801	4.5755
	iii	8.3163	8.5790	1.1157	1.0816	9.2789	9.2793
10/50	i	1.1247	1.1300	1.1394	1.1333	1.2814	1.2807
	ii	3.8979	3.9195	1.0097	1.0042	3.9358	3.9359
	iii	7.2876	7.2720	1.0805	1.0825	7.8740	7.8721
10/100	i	1.2386	1.2440	1.1860	1.1810	1.4689	1.4692
	ii	6.5433	6.5814	1.0223	1.0167	6.6894	6.6916
	iii	13.0169	12.9765	0.9628	0.9658	12.5323	12.5330
20/100	i	1.3157	1.3060	1.0641	1.0726	1.4001	1.4008
	ii	5.7614	5.7530	1.0235	1.0248	5.8966	5.8954
	iii	8.7339	8.7237	1.0498	1.0508	9.1687	9.1672
50/100	i	1.2858	1.2842	1.0996	1.1006	1.4139	1.4134
	ii	2.6318	2.6340	1.4684	1.4690	3.8646	3.8693
	iii	3.2459	3.2503	1.6262	1.6250	5.2784	5.2817

Table 13c

Estimated Relative Efficiencies of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ under Different Subsampling Schemes

$\rho = 0.1$

n/n'	Model	MSE(DES)/MSE(SRS)		MSE(SRS)/MSE(PRO)		MSE(DES)/MSE(PRO)	
		$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$
6/50	i	0.8964	1.0321	1.1609	1.0095	1.0406	1.0321
	ii	3.2067	3.5521	1.2053	1.0903	3.8649	3.8729
	iii	7.4132	7.7235	1.0638	1.0211	7.8863	7.8867
10/50	i	1.0117	1.0672	1.0516	0.9996	1.0638	1.0667
	ii	3.3983	3.4738	0.9711	0.9486	3.3000	3.2954
	iii	6.7827	6.7094	1.0499	1.0611	7.1215	7.1193
10/100	i	1.1307	1.2006	1.0879	1.0254	1.2302	1.2311
	ii	5.6086	5.7667	1.0450	1.0163	5.8612	5.8606
	iii	11.6999	11.5746	0.9821	0.9930	11.4904	11.4932
20/100	i	1.2686	1.2959	1.0449	1.0227	1.3256	1.3253
	ii	5.0326	5.0114	1.0768	1.0803	5.4192	5.4140
	iii	8.2523	8.1087	1.0063	1.0245	8.3045	8.3076
50/100	i	1.1784	1.1816	1.0579	1.0549	1.2466	1.2464
	ii	2.5516	2.5413	1.3556	1.3603	3.4590	3.4569
	iii	3.1543	3.1338	1.6730	1.6793	5.2771	5.2628

Table 14

Estimated Relative Efficiencies of $\bar{Y}_{\ell}^{*}$ to $\bar{Y}_{C}$ When Subsampling With Protection Against a Second Order Model

n/n'	Model	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$
6/50	i	≈1.0	1.0002	1.0012
	ii	0.9999	0.9991	1.0013
	iii	0.9999	1.0007	0.9993
10/50	i	≈1.0	0.9996	1.0031
	ii	0.9997	0.9999	0.9991
	iii	1.0002	0.9998	0.9998
10/100	i	≈1.0	1.0002	1.0004
	ii	≈1.0	1.0004	0.9999
	iii	≈1.0	≈1.0	1.0001
20/100	i	≈1.0	0.9999	0.9996
	ii	≈1.0	≈1.0	1.0001
	iii	≈1.0	≈1.0	≈1.0
50/100	i	1.0003	0.9998	0.9998
	ii	1.0004	1.0012	0.9994
	iii	≈1.0	1.0006	0.9973

The simulation results support the contention that robust designs are needed to protect against model misspecification. However, the consequence of designing subsamples to protect against model misspecification when the assumed or specified model is correct should also be considered.

The conditional mean squared errors of $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ when subsamples are designed with protection against a second order model are compared to their respective conditional mean squared errors under designed subsamples without protection against model misspecification when the true model is the assumed model, a first order linear model. Results are shown in Table 15. In Table 16, the conditional mean squared errors of $\bar{Y}_{C}$ and $\bar{Y}_{\ell}^{*}$ are compared under designed subsamples with protection to simple random subsamples when the simple linear model is the true model.

Table 15 demonstrates that some loss in efficiency, as much as 13%, will occur when one designs subsamples to protect against a second order model when it is unnecessary because no model misspecification has taken place. However, this loss in efficiency is far less than the loss in efficiency suffered by designing subsamples without regard to model misspecification whereupon the assumed model is wrong, as shown in Tables 13a to c. We therefore recommend that if there is the slightest doubt about the simple linear model as being the correct model, one should design subsamples with protection against model misspecification. The benefit attained by designing subsamples with protection against possible model misspecification when it exists far outweighs the loss incurred when there is no error in the assumed model.

Tables 13a to c have shown that although simple random subsamples afford some protection against model misspecification when the model is wrong, the designed subsamples with protection were more efficienct. It is of interest to

see how the designed subsamples with protection compare to simple random subsamples when protection is not required. Table 15 shows that the designed subsamples with protection against a second order model are generally inferior to those without protection when the assumed model is right. Nonetheless, Table 16 indicates that the designed subsamples with protection can be more efficient than simple random subsamples even when the assumed model is correct. This observation only strengthens the previous recommendation that one should design subsamples with protection if the assumed model is suspect.

Table 15

Estimated Relative Efficiencies of $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ Under Designed Subsamples Without Protection versus Designed Subsamples With Protection When the Assumed Model is True (MSE PROTECTION/MSE DESIGN)

		$\rho = 0.9$		$\rho = 0.5$		$\rho = 0.1$	
n'	n	$\bar{Y}_{C}$	$\bar{Y}_{\ell}^{*}$	$\bar{Y}_{C}$	$\bar{Y}_{\ell}^{*}$	$\bar{Y}_{C}$	$\bar{Y}_{\ell}^{*}$
50	6	1.0963	1.0957	1.0369	1.0353	0.9257	0.9232
	10	0.8806	0.8808	0.9623	0.9632	0.9894	0.9899
	24	1.0156	1.0171	1.0543	1.0556	1.0414	1.0354
100	10	1.0013	1.0010	0.9848	0.9857	1.0845	1.0843
	20	1.0040	1.0041	0.9773	0.9782	0.9357	0.9352
	50	1.0144	1.0140	0.9684	0.9695	1.0003	0.9981
500	50	0.9718	0.9719	0.9916	0.9919	1.1024	1.1030
	100	1.0035	1.0035	0.9537	0.9535	1.1325	1.1326
	250	1.0603	1.0603	0.9585	0.9585	1.0478	1.0486

Table 16

Estimated Relative Efficiencies of $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ Under Designed Subsamples With Protection Versus Simple Random Subsamples When the Assumed Model is True (MSE(SRS)/MSE(PROTECTION))

		$\rho = 0.9$		$\rho = 0.5$		$\rho = 0.1$	
n'	n	$\bar{Y}_{C}$	$\bar{Y}_{\ell}^{*}$	$\bar{Y}_{C}$	$\bar{Y}_{\ell}^{*}$	$\bar{Y}_{C}$	$\bar{Y}_{\ell}^{*}$
50	6	1.1682	1.2266	1.3143	1.2268	1.3036	1.1302
	10	1.1492	1.1640	1.1399	1.1354	1.1233	1.0614
	24	0.9939	0.9930	0.9713	0.9748	0.9754	0.9690
100	10	1.1433	1.1642	1.1155	1.1091	1.0797	1.0158
	20	0.9704	0.9722	1.0388	1.0484	1.0949	1.0719
	50	1.0264	1.0268	0.9862	0.9865	0.9983	0.9964
500	50	0.9959	0.9961	1.1552	1.1597	0.9713	0.9636
	100	0.9781	0.9781	1.0051	1.0060	0.9493	0.9476
	250	0.9395	0.9395	0.9884	0.9884	0.9128	0.9122

4.2.2 Performance Under a Bivariate Lognormal Population

The previous discussions evaluated the performance of simple random subsamples versus subsampling plans with protection and without protection against model misspecification. In these previous discussions, the variables studied came from normally distributed populations. Although the normality of distributions of random variables is not an uncommon assumption in statistics, the robustness of the designed subsampling procedures to departures from normality could be an issue.

The particular case which we examined was the case where a linear model is used to describe the relationship between the variables X and Y when X and Y are bivariate lognormal random variables. The rationale for choosing the bivariate lognormal distribution over other nonnormal bivariate distributions hinges on the reasoning that bivariate lognormal and bivariate normal distributions are similar and yet different. Bivariate lognormal distributions are skewed while bivariate normal distributions are symmetric. In practice, bivariate lognormal variables could easily be mistaken for bivariate normal variables when one makes a judgment on the distribution of two random variables based on a sample, especially when the sample is small. It is for this reason that simulation runs were done for the different subsampling procedures under bivariate lognormal distributions.

Bivariate lognormal random variable values are generated from bivariate normal distributions through the following transformation. Given the bivariate normal random variables $(Z_1, Z_2) \sim BVN(\mu_1, \mu_2, \sigma, \sigma_2, \rho)$, then $Y = \exp(Z_1)$ and $X = \exp(Z_2)$ are bivariate lognormal random variables with correlation coefficient,

$$\zeta = \frac{\exp(\rho\sigma_1\sigma_2) - 1}{\sqrt{[\exp(\sigma_1^2)-1][\exp(\sigma_2^2)-1]}} . \qquad (4.6)$$

The joint density of Y and X is given by

$$\phi(Y,X) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}YX} \exp\left[-\frac{1}{2(1-\rho^2)}(A^2 - 2\rho AB + B^2)\right] , \quad (4.7)$$

where

$$A = \frac{\log Y - \mu_1}{\sigma_1} ,$$

$$B = \frac{\log X - \mu_2}{\sigma_2} ,$$

and $\sigma_1, \sigma_2 > 0$, $-\infty < \mu_1, \mu_2 < \infty$, $-1 < \rho < 1$, $0 < Y < \infty$, $0 < X < \infty$.
(See Mostafa and Mahmoud, 1964.)

Simulation results shown in Tables 17 to 19 compare the different subsampling procedures with regard to mean squared error when the variables involved are bivariate lognormals. Estimated relative efficiencies indicate that the algorithm for designed subsampling without protection to model misspecification can be sensitive to departures from the normal distribution. In this case, the designed subsamples without protection performed dismally in comparison to both simple random subsamples and designed subsamples with protection when X and Y are bivariate lognormal random variables with a strong correlation, ζ. The designed subsamples without protection yielded mean squared errors that were as much as 6 times larger than those given by designed subsamples with protection or by simple random subsamples. However, the designed subsamples without protection do not do badly in comparison to the two other subsampling procedures for low correlations. More than 20% efficiency

can be gained by using designed subsamples without protection over simple random subsamples when ζ is low.

There appears to be no uniformly best procedure among the three subsampling methods being compared. However, the designed subsamples with protection seem most stable. Whereas the designed subsamples without protection and simple random subsamples exhibit up to 80% loss in efficiency depending on the value of ζ, the designed subsamples with protection is never more than 10% less efficient than either of the two other procedures.

Results from Table 20 indicate no difference between $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ with respect to mean squared error when using designed subsamples with protection or without protection. However, $\bar{Y}_{\ell}^{*}$ is more efficient than $\bar{Y}_{C}$ for simple random subsamples with small n' and when the correlation is low. The same observation holds when X and Y are normally distributed. (Refer to Table 21.)

When the random variables of interest have a bivariate lognormal distribution instead of a bivariate normal distribution, designed subsamples without protection lose part of their effectiveness over simple random subsamples. On the other hand, although designed subsamples with protection were developed specifically to guard only against model term misspecification, they are less sensitive to X and Y being lognormally distributed compared to simple random subsamples and designed subsamples without protection. These observations suggest that the performance of $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ in conjunction with the subsampling method used is affected by the distribution of the variables of interest. In this study, only the lognormal distribution has been considered in addition to the normal distribution. $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$ together with the different subsampling methods could behave differently under other population distributions. Further research is needed to determine the relative efficiencies

of the three subsampling procedures discussed under other population distributions. Means of rendering the estimators, $\bar{Y}_{\ell}^{*}$ and $\bar{Y}_{C}$, robust to deviations from normality should also be explored. Under a bivariate lognormal distribution, the subsampling plan that happens to be the most stable among the three being considered is designed subsampling with protection against a second order model.

Table 17

Estimated Relative Efficiency[1] of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ using Designed Subsamples Without Protection Versus Simple Random Subsamples (Bivariate Lognormal Distribution)

n'	n	$\zeta = 0.9$		$\zeta = 0.5$		$\zeta = 0.1$	
		$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$
50	6	0.7017	0.7326	1.2309	1.1372	1.2769	1.1032
	10	0.7477	0.7566	1.0026	0.9992	1.0176	0.9615
	24	0.9385	0.9388	1.0379	1.0428	1.1181	1.1054
100	10	0.5136	0.5236	1.0049	0.9976	1.2111	1.1346
	20	0.5289	0.5300	1.0080	1.0171	1.0260	1.0040
	50	0.8491	0.8491	1.0183	1.0199	1.0032	0.9990
500	50	0.1626	0.1627	0.8855	0.8884	1.0445	1.0368
	100	0.2076	0.2076	0.9328	0.9335	1.0105	1.0084
	250	0.5521	0.5521	0.9282	0.9283	1.0430	1.0432

[1] $\dfrac{\text{MSE(SRS)}}{\text{MSE(w/o PROTECTION)}}$

Table 18

Estimated Relative Efficiency[1] of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ using Designed Subsamples With Protection Versus Simple Random Subsamples (Bivariate Lognormal Distribution)

n'	n	$\zeta = 0.9$		$\zeta = 0.5$		$\zeta = 0.1$	
		$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$
50	6	1.2087	1.2616	1.1794	1.0901	1.2455	1.0768
	10	1.1312	1.1444	1.1049	1.1033	1.0899	1.0288
	24	1.0173	1.0173	1.0222	1.0269	1.0487	1.0409
100	10	1.0277	1.0478	1.1507	1.1507	1.1860	1.1122
	20	1.0715	1.0735	1.0496	1.0586	1.0663	1.0432
	50	0.9967	0.9967	1.0378	1.0399	0.9454	0.9455
500	50	1.0453	1.0458	1.0688	1.0723	1.0468	1.0380
	100	1.0020	1.0021	0.9777	0.9785	1.0402	1.0383
	250	0.9641	0.9639	0.9066	0.9063	0.9966	0.9980

[1] $\dfrac{\text{MSE(SRS)}}{\text{MSE(with PROTECTION)}}$

Table 19

Estimated Relative Efficiency[1] of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ using Designed Subsamples With Protection Against Model Misspecification Versus Designed Subsamples Without Protection (Bivariate Lognormal Distribution)

n'	n	$\zeta = 0.9$		$\zeta = 0.5$		$\zeta = 0.1$	
		$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$	$\bar{Y}_C$	$\bar{Y}^*_\ell$
50	6	1.7225	1.7221	0.9582	0.9586	0.9754	0.9761
	10	1.5130	1.5127	1.1021	1.1041	1.0710	1.0700
	24	1.0840	1.0837	0.9849	0.9847	0.9379	0.9417
100	10	2.0008	2.0012	1.1528	1.1534	0.9793	0.9802
	20	2.0259	2.0256	1.0414	1.0408	1.0393	1.0390
	50	1.1739	1.1739	1.0192	1.0196	0.9424	0.9464
500	50	6.4278	6.4279	1.2071	1.2071	1.0021	1.0012
	100	4.8257	4.8259	1.0481	1.0482	1.0294	1.0296
	250	1.7462	1.7458	0.9767	0.9762	0.9555	0.9567

[1] $\frac{\text{MSE (w/o PROTECTION)}}{\text{MSE (with PROTECTION)}}$

Table 20

Estimated Relative Efficiency of $\bar{Y}_{\ell}^{*}$ to $\bar{Y}_{C}$ Under the Different Subsampling Schemes for a Bivariate Lognormal Distribution

n'	n	ζ	SRS	DESIGN	PROTECTION
50	6	0.9	0.9576	0.9998	0.9996
		0.5	1.0816	0.9992	0.9997
		0.1	1.1595	1.0018	1.0024
	10	0.9	0.9882	0.9999	0.9997
		0.5	1.0031	0.9998	1.0016
		0.1	1.0582	0.9998	0.9989
	24	0.9	0.9997	≈1.0	0.9997
		0.5	0.9952	≈1.0	0.9999
		0.1	1.0116	1.0002	1.0042
100	10	0.9	0.9808	0.9999	1.0001
		0.5	1.0070	0.9997	1.0003
		0.1	1.0674	≈1.0	1.0009
	20	0.9	0.9979	≈1.0	0.9998
		0.5	0.9910	≈1.0	0.9995
		0.1	1.0218	≈1.0	0.9997
	50	0.9	≈1.0	≈1.0	0.9999
		0.5	0.9983	≈1.0	1.0003
		0.1	1.0042	1.0001	1.0043
500	50	0.9	0.9996	≈1.0	≈1.0
		0.5	0.9967	≈1.0	≈1.0
		0.1	1.0076	1.0001	0.9991
	100	0.9	≈1.0	≈1.0	≈1.0
		0.5	0.9993	≈1.0	1.0001
		0.1	1.0020	≈1.0	1.0002
	250	0.9	≈1.0	≈1.0	0.9998
		0.5	0.9999	≈1.0	0.9995
		0.1	0.9998	≈1.0	1.0012

Table 21

Estimated Relative Efficiency of $\bar{Y}_{\ell}^{*}$ to $\bar{Y}_{C}$ Under Simple Random Subsampling (Bivariate Normal Distribution)

n'	n	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$
50	5	0.9792	1.1567	1.2824
	10	0.9869	1.0031	1.0576
	25	0.9994	0.9952	1.0096
100	10	0.9806	1.0050	1.0636
	20	0.9984	0.9898	1.0211
	50	~1.0	0.9982	1.0039
500	50	0.9996	0.9960	1.0060
	100	0.9999	0.9994	1.0020
	250	~1.0	~1.0	1.0002

4.3 MISSPECIFICATION OF THE VARIANCE STRUCTURE

4.3.1 Analytical Results

An additional source of error is the misspecification of the variance function. In Chapter 3, the superpopulation model is assumed to be $\xi[1, 1: 1]$. $\mathrm{Var}(\epsilon_i | X_i)$ is assumed equal to σ_ϵ^2 for all i. Instead of being constant for all values of X, the true variances of the errors could vary as a function of X. This research also explores the implications of the misspecification of the variance structure on both the estimators, $\bar{Y}_C$ and $\bar{Y}_\ell^*$, and the subsampling procedures being discussed.

Let us consider the case where the true $\mathrm{Var}(\epsilon_i | x_i) = \sigma_C^2 \nu(x_i)$ and $\nu(x)$ is a known function with $\nu(x) > 0$. It is a well known result that when the variances of the errors are not equal at different points, the ordinary least squares estimator $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}$ for $\underline{\beta}$ in the linear model $\underline{y} = x\underline{\beta} + \underline{\epsilon}$ is no longer optimal with respect to variance. By the Gauss-Markov Theorem, the generalized least squares estimator, $\hat{\underline{\beta}}^* = (X'V^{-1}X)^{-1}X'V^{-1}\underline{y}$ is BLUE when $\mathrm{Var}(\underline{\epsilon}) = \sigma_C^2 V$. Therefore,

$$\mathrm{Var}(\hat{\underline{\beta}}^*) \leq \mathrm{Var}(\hat{\underline{\beta}}) \qquad (4.8)$$

when $\mathrm{Var}(\underline{\epsilon}) = \sigma_C^2 \mathrm{diag}[\nu(x_1), \nu(x_2), \ldots, \nu(x_N)]$.

Assuming the superpopulation model is $\xi[1, 1: 1]$, then the proposed alternative double sample regression estimator is

$$\bar{Y}_\ell^* = \frac{\sum_{i=1}^{n} y_i}{n} + k\hat{\beta}_1 (\bar{X}' - \bar{X}), \text{ equation (2.3)},$$

where $\hat{\beta}_1$ is the OLS estimator of β_1.

If the true $\text{Var}(\epsilon_i|x_i) = \sigma_c^2\nu(x_i)$ instead of being constant as assumed, then the $E[(\bar{Y}_\ell^* - \bar{Y})^2|x_1, x_2, \ldots, x_{n'}]$ is no longer equal to equation (2.7). When $\text{Var}(\epsilon_i|x_i) = \sigma_c^2\nu(x_i)$ and $\nu(x) > 0$ is a known function, then

$$E[(\bar{Y}_\ell^* - \bar{Y})^2|x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_c^2\left[\frac{\sum_{i=1}^{n}\nu(x_i)}{n}\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\sum_{i=1}^{N}\nu(x_i)}{N^2}\right.$$

$$+ 2k(\bar{x}' - \bar{x})\left\{\frac{\sum_{i=1}^{n}(x_i - \bar{x})\nu(x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\left(\frac{1}{n} - \frac{1}{N}\right)\right\}$$

$$\left. + k^2(\bar{x}' - \bar{x})^2\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\nu(x_i)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}\right]$$

$$+ \beta_1^2[(1-k)\bar{x} + k\bar{x}' - \bar{X}]^2 \,. \qquad (4.9)$$

Similarly,

$$E[(\bar{Y}_C - \bar{Y})^2|x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_c^2\left[\frac{\sum_{i=1}^{n}\nu(x_i)}{n}\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\sum_{i=1}^{N}\nu(x_i)}{N^2}\right.$$

$$+ \frac{(\bar{X}' - \bar{X})}{\sum_{i=1}^{n} (x_i - \bar{X})^2} \left\{ 2 \sum_{i=1}^{n} (x_i - \bar{X})\nu(x_i)\left(\frac{1}{n} - \frac{1}{N}\right) \right.$$

$$\left. + (\bar{X}' - \bar{X}) \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2 \nu(x_i)}{\sum_{i=1}^{n} (x_i - \bar{X})^2} \right\} + \beta_1^2(\bar{X}' - \bar{X})^2 . \qquad (4.10)$$

The value of k that minimizes equation (4.9) is also a function of $\nu(x)$ and is given by

$$k_{\nu(x)} = \frac{\beta_1^2(\bar{X} - \bar{X}) - \sigma_c^2 \left\{ \frac{\sum_{i=1}^{n} (x_i - \bar{X})\nu(x_i)}{\sum_{i=1}^{n} (x_i - \bar{X})^2} \left(\frac{1}{n} - \frac{1}{N}\right) \right\}}{\beta_1^2(\bar{X}' - \bar{X}) + \sigma_c^2(\bar{X}' - \bar{X}) \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2 \nu(x_i)}{\left(\sum_{i=1}^{n} (x_i - \bar{X})^2\right)^2}} . \qquad (4.11)$$

It is observed from equations (4.9), (4.10), and (4.11) that the conditional mean squared errors of $\bar{Y}_c$ and $\bar{Y}_\ell^*$ are affected by a nonconstant error variance. The above observation coupled with the knowledge that ordinary least squares estimators are not the best estimators of regression coefficients when the assumption of constant variance of errors is violated leads one to consider using weighted least squares estimators for regression coefficients in the estimators, $\bar{Y}_c$ and $\bar{Y}_\ell^*$.

Given $Y_i = \beta_o + \beta_1 X_i + \epsilon_i$,

where

$$E(\epsilon_i|x_i) = 0 \ ,$$

$$E(\epsilon_i^2|x_i) = \sigma_c^2 \nu(x_i), \ \nu(x_i) > 0 \ ,$$

and

$$E(\epsilon_i\epsilon_j|x_i, x_j) = 0 \ ,$$

the BLUE of $\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ is

$$\hat{\underline{\beta}}^* = (X'V^{-1}X)^{-1}X'V^{-1}\underline{Y}$$

where

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \ , \quad \underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \ ,$$

and

$$V = \begin{bmatrix} \nu(x_1) & 0 & 0 & \cdots & 0 \\ 0 & \nu(x_2) & 0 & \cdots & 0 \\ 0 & 0 & \nu(x_3) & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \nu(x_n) \end{bmatrix} \ .$$

Letting $\hat{\underline{\beta}}^*$ be $\begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_1^* \end{bmatrix}$ then

$$\hat{\beta}_o^* = \frac{\sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} \sum_{i=1}^{n} \frac{Y_i}{\nu(x_i)} - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \sum_{i=1}^{n} \frac{x_i Y_i}{\nu(x_i)}}{\sum_{i=1}^{n} \frac{1}{\nu(x_i)} \sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} - \left(\sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \right)^2} , \quad (4.12)$$

$$\hat{\beta}_i^* = \frac{\sum_{i=1}^{n} \frac{x_i Y_i}{\nu(x_i)} \sum_{i=1}^{n} \frac{1}{\nu(x_i)} - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \sum_{i=1}^{n} \frac{Y_i}{\nu(x_i)}}{\sum_{i=1}^{n} \frac{1}{\nu(x_i)} \sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} - \left(\sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \right)^2} ,$$

and

$$\mathrm{Var}(\hat{\underline{\beta}}^*) = \sigma_c^2 a \begin{bmatrix} \sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} & - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \\ - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} & \sum_{i=1}^{n} \frac{1}{\nu(x_i)} \end{bmatrix} , \quad (4.13)$$

where

$$a = \left[\sum_{i=1}^{n} \frac{1}{\nu(x_i)} \sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} - \left(\sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \right)^2 \right]^{-1} .$$

Let

$$\bar{Y}_c{}^w = \hat{\beta}_o^* + \hat{\beta}_i^* \bar{x}' ,$$

and

$$\bar{Y}_\ell{}^w = (1-k) \frac{\sum_{i=1}^{n} Y_i}{n} + k(\hat{\beta}_o^* + \hat{\beta}_i^* \bar{x}') .$$

Under the superpopulation model $\xi[1, 1: \nu(x)]$, the

$$E[(\bar{Y}_C^W - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_C^2 \left[a \sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} + a(\bar{x}')^2 \sum_{i=1}^{n} \frac{1}{\nu(x_i)} - 2a\bar{x}' \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \right.$$

$$- \frac{2a}{N} \left\{ n \sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \frac{1}{\nu(x_i)} \sum_{i=1}^{n} x_i \right.$$

$$\left. \left. - n \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \right\} + \frac{\sum_{i=1}^{N} \nu(x_i)}{N} \right] + \beta_1^2 (\bar{x}' - \bar{X})^2 , \qquad (4.14)$$

and

$$E[(\bar{Y}_\ell^W - \bar{Y}]^2 | x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_C^2 \left[(1-k)^2 \sum_{i=1}^{n} \frac{\nu(x_i)}{n^2} - 2(1-k) \sum_{i=1}^{n} \frac{\nu(x_i)}{nN} + \frac{\sum_{i=1}^{N} \nu(x_i)}{N^2} \right.$$

$$+ 2(1-k)ka \left\{ \left(\sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} - \bar{x}' \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \right)\left(1 - \frac{n}{N}\right) \right.$$

$$\left. \left. + \left(\bar{x}' \sum_{i=1}^{n} \frac{1}{\nu(x_i)} - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \right)\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{n} x_i \right\} \right] \qquad (4.15)$$

$$- 2k^2 a \left\{ \frac{1}{N} \sum_{i=1}^{n} x_i \left(\bar{x}' \sum_{i=1}^{n} \frac{1}{\nu(x_i)} - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \right) \right.$$

$$+ \frac{n}{N} \sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} + \bar{X}' \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \left(1 - \frac{n}{N}\right)\Bigg\}\Bigg]$$

$$+ k^2 a \left\{\sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} + (\bar{X}')^2 \sum_{i=1}^{n} \frac{1}{\nu(x_i)}\right\}\Bigg] + \beta_1^2[(1-k)\bar{X} + k\bar{X}' - \bar{X}]^2 \ .$$

The optimal value of k for $\bar{Y}_R^W$, the proposed double sample regression estimator using weighted least squares regression coefficients, is given by

$$k^W = \frac{\beta_1^2(\bar{X}'-\bar{X})(\bar{X}-\bar{X}) + \left[\sum_{i=1}^{n} \frac{\nu(x_i)}{n}\left(\frac{1}{n} - \frac{1}{N}\right) - aQ\right]\sigma_c^2}{\beta_1^2(\bar{X}'-\bar{X})^2 + \left[\sum_{i=1}^{n} \frac{\nu(x_i)}{n^2} + a\{R - 2(P+Q)\}\right]\sigma_c^2} \ , \qquad (4.16)$$

where

$$Q = \Bigg\{\Bigg(\sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} - \bar{X}' \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)}\Bigg)\left(1 - \frac{n}{N}\right) + \left(\bar{X}' \sum_{i=1}^{n} \frac{1}{\nu(x_i)}\right.$$

$$\left. - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)}\right)\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{n} x_i\Bigg\} \ ,$$

$$P = \Bigg\{\frac{1}{N} \sum_{i=1}^{n} x_i \left(\bar{X}' \sum_{i=1}^{n} \frac{1}{\nu(x_i)} - \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)}\right) + \frac{n}{N} \sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)}$$

$$+ \bar{X}' \sum_{i=1}^{n} \frac{x_i}{\nu(x_i)} \left(1 - \frac{n}{N}\right)\Bigg\} \ ,$$

and

$$R = \left[\sum_{i=1}^{n} \frac{x_i^2}{\nu(x_i)} + (\bar{x}')^2 \sum_{i=1}^{n} \frac{1}{\nu(x_i)} \right] .$$

The optimal value of k given in (4.16) is complicated enough to make $\bar{Y}_\ell^W$ seem impracticable.

It was then decided to look at a specific function $\nu(x)$ for further study. The variance function $\nu(x) = x$ where $x > 0$ was selected due to its common occurrence in practice. In this case the variances of the errors are directly proportional to the values of x.

Theoretically, when $\mathrm{Var}(\epsilon_i) = \sigma_c^2 x_i$ and ordinary least squares estimators are used to estimate the regression coefficients in $\bar{Y}_c$ and $\bar{Y}_\ell^*$, then the respective mean squared errors of $\bar{Y}_c$ and $\bar{Y}_\ell^*$ are as follows:

$$E[(\bar{Y}_c - \bar{Y})^2 | x_1, x_2, \ldots, x_n']$$

$$= \sigma_\epsilon^2 \left[\frac{\bar{x}}{N} - \frac{\bar{x}}{n} + 2\bar{x}'\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{(\bar{x}'-\bar{x})^2}{\sum_{i=1}^{n} (x_i-\bar{x})^2} \right]$$

$$+ \beta_1^2 (\bar{x}' - \bar{x})^2 , \qquad (4.17)$$

and

$$E[(\bar{Y}_\ell^* - \bar{Y})^2 | x_1, x_2, \ldots, x_n']$$

$$= \sigma_\epsilon^2 \left[\bar{x}\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{1}{N}(\bar{x}-\bar{x}) + k^2(\bar{x}'-\bar{x})^2 \frac{\sum_{i=1}^{n} x_i(x_i-\bar{x})^2}{\left(\sum_{i=1}^{n} (x_i-\bar{x})^2\right)^2} \right.$$

$$+ 2k(\bar{x}'-\bar{x})(\frac{1}{n} - \frac{1}{N})] + \beta_1^2[(1-k)\bar{x} + k\bar{x}' - \bar{X}]^2 . \qquad (4.18)$$

It is evident from equations (4.17) and (4.18) that the conditional mean squared errors of $\bar{Y}_C$ and $\bar{Y}_\ell^*$ using ordinary least squares depend upon the actual error variance. (Compare equation (4.17) to (2.1) and equation (4.18) to (2.7).)

Since $MSE(\bar{Y}_\ell^* | x_1, x_2, \ldots, x_{n'})$ is strongly influenced by the error variance, the value for k that was optimal when the error variance was constant (see equation (2.4)) is no longer optimal when $Var(\epsilon_i) = \sigma_C^2 x_i$. The new value for k that minimizes $MSE(\bar{Y}_\ell^* | x_1, x_2, \ldots, x_{n'})$ in equation (4.18) is given by

$$k_x = \frac{\beta_1^2(\bar{X}-\bar{x}) - \sigma_\epsilon^2(\frac{1}{n} - \frac{1}{N})}{\beta_1^2(\bar{x}'-\bar{x}) + \sigma_\epsilon^2(\bar{x}'-\bar{x}) \dfrac{\sum_{i=1}^{n} (x_i-\bar{x})^2 x_i}{\left(\sum_{i=1}^{n} (x_i-\bar{x})^2\right)^2}} . \qquad (4.19)$$

If one were to use weighted least squares estimates of coefficients in the estimators, $\bar{Y}_C$ and $\bar{Y}_\ell^*$, in the presence of an error variance proportional to the values of X, i.e., $\sigma_\epsilon^2 = \sigma_C^2 x_i$, then the

$$E[(\bar{Y}_\ell^W - \bar{Y})^2 | x_1, x_2, \ldots, x_{n'}]$$

$$= \sigma_C^2 \left[\bar{x}(\frac{1}{n} - \frac{1}{N}) + \frac{\bar{X}-\bar{x}}{N} + 2k(\bar{x}' - \bar{x})(\frac{1}{n} - \frac{1}{N}) \right.$$

$$- k^2(\bar{x}'-\bar{x})(\frac{1}{n}) - k^2\frac{\bar{x}'}{n} + k^2\left(\frac{\bar{x}+(\bar{x}')^2\bar{w}-2\bar{x}'}{n\ \bar{x}\ \sum_{i=1} w_i - n}\right)\Bigg]$$

$$+ \beta_1^2[\{(1-k)\bar{x} + k\bar{x}'\} - \bar{X}]^2 \ . \qquad (4.20)$$

The corresponding optimal value for k was found to be

$$k_x^w = \frac{(\bar{x}'-\bar{x})[\beta_1^2(\bar{X}-\bar{x}) - \sigma_c^2(\frac{1}{n} - \frac{1}{N})}{(\bar{x}'-\bar{x})[\beta_1^2(\bar{x}'-\bar{x}) - \frac{\sigma_c^2}{n}] + \sigma_c^2\left[\frac{\bar{x} + (\bar{x}')^2\bar{w} - 2\bar{x}'}{n\ \bar{x}\ \sum_{i=1} w_i - n} - \frac{\bar{x}'}{n}\right]}, \qquad (4.21)$$

where

$$w_i = \frac{1}{x_i} \ .$$

It should be noted that when the variances of the errors are not constant but of the form $\sigma_c^2\nu(x)$, the optimal shrinkage factors involved in the alternative double sample regression estimator using either ordinary or weighted least squares coefficients are functions of the unknown parameters, β_1 and σ_c. In practice, these parameters need to be estimated making the alternative double sample regression estimator more complicated computationally.

4.3.2 Simulation Results

Having shown the consequences of a nonconstant error variance on the conditional mean squared errors of $\bar{Y}_c$ and $\bar{Y}_\ell^*$ theoretically, it is of interest to investigate the actual behavior of $\bar{Y}_c$ and $\bar{Y}_\ell^*$ under the three subsampling procedures, discussed in previous sections, through simulations.

As in section 4.3.1, we let $\nu(x)$ be equal to x, $x > 0$. Since the main purpose of the simulation study was to observe the effect of a nonhomogeneous error variance on $\bar{Y}_C$ and $\bar{Y}_\ell^*$, the variance of the ϵ_i's was set equal to $\sigma_C^2 x_i$ such that $\sigma_C^2 \bar{x}$ is equal to the value of the constant error variance, σ_ϵ^2, used in previous simulation runs. In this manner, one is able to look at a different variance structure of comparable magnitude to our previous simulation results.

From the results in Tables 22a, b, and c, it is observed that the designed subsamples are robust to this particular misspecification of the variance function. The designed subsamples with and without protection against a second order model maintained the same advantage they had over simple random subsamples when the error variance was constant. As in our previous simulation results, there is again no practical difference in mean squared errors between $\bar{Y}_C$ and $\bar{Y}_\ell^*$ when using designed subsamples. However, under simple random subsampling, the estimator, $\bar{Y}_\ell^*$, is preferable to $\bar{Y}_C$ with respect to mean squared error when the correlation is weak and when the subsample sizes are small, maintaining the same level of efficiency it had when the error variance was constant. (Refer to Table 23.) It is worth noting that if σ_C were allowed to get larger, the superiority of designed subsamples to simple random subsamples becomes more emphasized in terms of size. This also holds true for $\bar{Y}_\ell^*$'s performance compared to $\bar{Y}_C$.

Simulations were also done to explore any improvement in the mean squared error of $\bar{Y}_C^W$ under simple random subsampling attributable to weighted coefficients. Results indicated no significant improvement in $MSE(\bar{Y}_C^W | x_1, x_2, \ldots, x_{n'})$ over $MSE(\bar{Y}_C | x_1, x_2, \ldots, x_{n'})$. This is not surprising since the nonconstant error variance structure, $Var[\epsilon_i] = \sigma_C^2 x_i$, $x_i > 0$, does not have an apparent detrimental effect on either $MSE(\bar{Y}_C | x_1, x_2, \ldots, x_{n'})$ or

$MSE(\bar{Y}^*_\ell | x_1, x_2, \ldots, x_n')$ compared to the situation when errors have a homogeneous variance. One may then conclude that the double sample regression estimators, $\bar{Y}_C$ and $\bar{Y}^*_\ell$, are not sensitive to a nonhomogeneous error variance structure of this type. Therefore, weighted regression coefficients are not warranted in practice for such situations.

Table 22a

Estimated Relative Efficiencies of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ Under Different Subsampling Schemes When $Var[\epsilon_i] = \sigma_C^2 x_i$

$\rho = 0.9$

n'	n	MSE(SRS)/MSE(DES) $\bar{Y}_C$	MSE(SRS)/MSE(DES) $\bar{Y}^*_\ell$	MSE(SRS)/MSE(PRO) $\bar{Y}_C$	MSE(SRS)/MSE(PRO) $\bar{Y}^*_\ell$	MSE(DES)/MSE(PRO) $\bar{Y}_C$	MSE(DES)/MSE(PRO) $\bar{Y}^*_\ell$
50	6	1.1531	1.2031	1.2128	1.2642	1.0517	1.0508
	10	1.0314	1.0473	1.1339	1.1520	1.0994	1.1000
	24	1.0064	1.0070	0.9828	0.9823	0.9765	0.9754
100	10	1.1596	1.1825	1.0164	1.0362	0.8765	0.8763
	20	0.9745	0.9765	1.0873	1.0892	1.1157	1.1154
	50	0.9624	0.9625	0.9738	0.9740	1.0119	1.0120
500	50	0.9863	0.9865	0.8985	0.8987	0.9110	0.9109
	100	0.9886	0.9887	1.0437	1.0439	1.0558	1.0439
	250	0.9885	0.9886	1.0575	1.0577	1.0698	1.0699

Table 22b

Estimated Relative Efficiencies of $\bar{Y}_C$ and $\bar{Y}^*_\ell$ Under Different Subsampling Schemes When $Var[\epsilon_i] = \sigma_C^2 x_i$

$\rho = 0.5$

n'	n	MSE(SRS)/MSE(DES) $\bar{Y}_C$	MSE(SRS)/MSE(DES) $\bar{Y}^*_\ell$	MSE(SRS)/MSE(PRO) $\bar{Y}_C$	MSE(SRS)/MSE(PRO) $\bar{Y}^*_\ell$	MSE(DES)/MSE(PRO) $\bar{Y}_C$	MSE(DES)/MSE(PRO) $\bar{Y}^*_\ell$
50	6	1.2515	1.1641	1.2510	1.1645	0.9996	1.0004
	10	1.1130	1.1082	1.0826	1.0753	0.9727	0.9703
	24	1.0652	1.0705	1.0176	1.0205	0.9553	0.9533
100	10	1.1219	1.1160	1.0799	1.0753	0.9625	0.9635
	20	0.9454	0.9542	1.0361	1.0442	1.0960	1.0943
	50	1.0460	1.0479	0.9718	0.9739	0.9291	0.9294
500	50	1.0310	1.0345	1.0119	1.0154	0.9815	0.9815
	100	0.9357	0.9362	0.9818	0.9820	1.0493	1.0489
	250	1.0347	1.0347	1.0225	1.0220	0.9883	0.9877

Table 22c

Estimated Relative Efficiencies of $\bar{Y}_c$ and $\bar{Y}^*_\ell$ Under Different Subsampling Schemes When $Var[\epsilon_i] = \sigma_c^2 x_i$

$\rho = 0.1$

n'	n	MSE(SRS)/MSE(DES) $\bar{Y}_c$	MSE(SRS)/MSE(DES) $\bar{Y}^*_\ell$	MSE(SRS)/MSE(PRO) $\bar{Y}_c$	MSE(SRS)/MSE(PRO) $\bar{Y}^*_\ell$	MSE(DES)/MSE(PRO) $\bar{Y}_c$	MSE(DES)/MSE(PRO) $\bar{Y}^*_\ell$
50	6	1.2408	1.0709	1.2577	1.0873	1.0136	1.0153
	10	1.0498	0.9911	1.0735	1.0136	1.0226	1.0227
	24	1.0846	1.0721	1.0607	1.0523	0.9780	0.9816
100	10	1.1292	1.0616	1.0883	1.0230	0.9638	0.9636
	20	1.0156	0.9932	1.0623	1.0378	1.0460	1.0448
	50	1.0023	0.9986	0.9818	0.9810	0.9796	0.9824
500	50	0.9602	0.9539	1.0404	1.0330	1.0835	1.0829
	100	1.0733	1.0713	0.9506	0.9492	0.8856	0.8861
	250	1.0920	1.0925	1.0124	1.0123	0.9271	0.9266

Table 23

Estimated Relative Efficiency of $\bar{Y}_{\ell}^{*}$ to $\bar{Y}_C$ Under Simple Random Subsampling When $Var[\epsilon_i] = \sigma_C^2 x_i$ and When $Var[\epsilon_i]$ is Constant

		$Var[\epsilon_i] = \sigma_C^2 x_i$			$Var[\epsilon_i] = \sigma_\epsilon^2$		
		$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.1$
n'	n						
50	6	0.9584	1.0750	1.1580	0.9529	1.0718	1.1565
	10	0.9848	1.0042	1.0600	0.9870	1.0029	1.0584
	24	0.9994	0.9950	1.0115	0.9995	0.9951	1.0122
100	10	0.9807	1.0052	1.0648	0.9822	1.0049	1.0630
	20	0.9980	0.9906	1.0228	0.9980	0.9899	1.0215
	50	0.9999	0.9982	1.0037	~1.0	0.9984	1.0040
500	50	0.9997	0.9965	1.0071	0.9998	0.9959	1.0072
	100	0.9999	0.9994	1.0020	~1.0	0.9993	1.0018
	250	~1.0	0.9999	0.9996	~1.0	0.9999	0.9998

V. DISCUSSION AND CONCLUSIONS

5.1 ALTERNATIVE DOUBLE SAMPLE REGRESSION ESTIMATOR

This study seeks to improve estimation in double sampling in two ways. One is by means of an alternative double sample regression estimator, $\bar{Y}^*_\ell$, and the other, by means of sampling plans designed to reduce mean squared error. The two methods can be used together or separately. Each method has its own merit.

$\bar{Y}^*_\ell$ is a shrinkage type estimator that is theoretically superior to both the standard double sample regression estimator, $\bar{Y}_C$ (Cochran, 1977), and the simple sample mean, $\bar{Y}_S = \frac{\sum_{i=1}^{n} y_i}{n}$. In practice, one decides to use or not to use regression estimators based upon whether one believes a relationship exists between the variable of interest, Y, and a secondary variable, X. A presumption of no relationship between X and Y leads to the estimator, $\bar{Y}_S$. On the other hand, a presumption of a significant linear relationship, under certain conditions, leads to $\bar{Y}_C$ when the X's are readily available. It should be noted that the decisions to use either $\bar{Y}_S$ or $\bar{Y}_C$ are based on fallible presumptions, resulting in loss of efficiency when presumptions do not coincide with reality. The estimator, $\bar{Y}^*_\ell$ eliminates this risk. Its most appealing feature, one that is absent in both $\bar{Y}_C$ and $\bar{Y}_S$, is its ability to take into account the strength of the linear relationship between X and Y. $\bar{Y}^*_\ell$ adjusts itself depending upon the correlation coefficient between X and Y, becoming $\bar{Y}_S$ when $\rho_{XY} = 0$ and becoming $\bar{Y}_C$ when $|\rho_{XY}| = 1$. We have shown $\bar{Y}^*_\ell$ to be superior to $\bar{Y}_C$ unless $|\rho_{XY}| = 1$. Since the true strength of the linear relationship between X and Y is typically unknown to the researcher, he or she can overcome that uncertainty

by using $\bar{Y}_{\ell}^{*}$. Under simple random subsampling, $\bar{Y}_{\ell}^{*}$ is generally most efficient when n is small or when n is small relative to n'.

Our results have been obtained under the assumption of a "superpopulation model", specifically a simple linear model. Using such an approach, one can work with either conditional or unconditional inferences. It turns out that the optimal weighting factor k in $\bar{Y}_{\ell}^{*}$ derived when one conditions on the initial sample actually drawn gives a better average squared error than the unconditional optimal k. Both the conditional k (2.9) and the unconditional k (2.11) depend on the actual value of ρ_{XY}. When ρ_{XY} is unknown and has to be estimated, simulation results indicate that the apparent superiority of the conditional k over the unconditional k disappears.

5.2 PURPOSIVE SUBSAMPLING PLANS

The second method formulated to improve double sampling estimation is by means of designing subsamples when given the initial sample. Two nonrandom subsampling plans are utilized to minimize mean squared errors of $\bar{Y}_{C}$ and $\bar{Y}_{\ell}^{*}$. The first subsampling scheme assumes no model misspecification while the second subsampling plan affords some protection against model form misspecification. Both subsampling plans can be implemented exactly only through an exhaustive look at all possible subsamples from an initial sample. Although such a procedure is no longer formidable with high speed computers, it may not be cost effective. For this reason, corresponding algorithms were developed to approximate each subsampling plan.

Each algorithm yields a unique subsample for each initial sample. This is in contrast to previous applications of purposive, nonrandom sampling by Iachan (1985), Herson (1976), and Royall and Cumberland (1981a & b). They all

employed some form of restricted random sampling based on a preliminary choice of a fixed constant, δ. For example, Royall and Cumberland (1981a) fixed δ so that 90% of all possible samples would be rejected. In all three schemes, more than one possible sample had a positive probability of being the selected sample. This makes the subsampling algorithms being applied in this study seem very restrictive in comparison.

One might argue against this study's algorithms by saying they totally ignore subsamples that yield better mean squared errors than the algorithm-selected subsample. The rejoinder to such an argument would be a restricted random subsampling plan with the constant, δ, set equal to the value of the quantitative criterion when computed from the algorithm-selected subsample. This form of restricted random subsampling would be more rigorous because it avoids an arbitrary choice of δ.

Simulation results in this study compare only the algorithm-selected subsamples with simple random subsamples. This does not make the results any more tenuous than if we had compared subsamples obtained through restricted random subsampling described in the preceding paragraph. After all, examining the algorithm-selected subsample is equivalent to looking at the worst subsample possible under the restricted random subsampling scheme subject to the specified criterion. Further empirical or simulation study is needed to determine whether there is an actual substantial gain with respect to mean squared error in using restricted random subsamples described in the preceding paragraph over the algorithm-selected subsamples. Although the algorithms presented in this study were the "best" approximations among several considered at the time, they should be reviewed for possible improvements.

When there is no model misspecification and the model is

$$Y_i = \beta_o + \beta_i x_i + \epsilon_i \ ,$$

where

$$E(\epsilon_i | x_i) = 0 \ ,$$

$$E(\epsilon_i^2 | x_i^2) = \sigma_\epsilon^2 \ ,$$

and

$$E(\epsilon_i \epsilon_j | x_i, x_j) = 0 \ ,$$

the criterion needed to minimize the mean squared error of both $\bar{Y}_c$ and $\bar{Y}_\ell^*$ is to make the quantity $\frac{(\bar{X}' - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}$ as small as possible. Simulation results using the FORTRAN algorithm DESIGN indicate that the designed subsamples should be favored over simple random subsamples when the initial sample size n' is small. Although $\bar{Y}_\ell^*$ can outperform $\bar{Y}_c$ using simple random subsamples, no such advantage exists for $\bar{Y}_\ell^*$ when using designed subsamples. We have shown that with simple random subsamples $\bar{Y}_\ell^*$ is definitely more efficient than $\bar{Y}_c$ when X and Y are weakly correlated. In these cases of low correlations, designing subsamples, however, was shown to be a more effective strategy in reducing the mean squared error of double sample regression estimates than using the alternative double sample regression estimator, $\bar{Y}_\ell^*$.

In practice, assumed "superpopulation models" may be incorrect. This leads to the question of robustness of the subsampling procedure that deliberately chooses a subsample that seeks to keep $\frac{(\bar{X}' - \bar{X})^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}$ near

minimum for a given initial sample. If one assumes a first order model as the "superpopulation model" when it actually is a second order model, designed subsamples without protection result in estimates with higher mean squared errors than those from simple random subsamples. This disparity gets worse with a sharper curvature in the second order model. Simulations revealed a definite loss in efficiency among designed subsamples without protection compared to simple random subsamples when there is model misspecification. In this study, we addressed this problem by developing an alternate FORTRAN algorithm to select subsamples that provide protection for $\bar{Y}_C$ and $\bar{Y}^*_{\ell}$ against a second order model. The protection furnished by the algorithm PROTEC is clearly limited since it protects only against second order models. Additional research should explore means of rendering $\bar{Y}_C$ and $\bar{Y}^*_{\ell}$ robust to higher degree polynomials or other forms of model misspecification through purposive sampling.

Through simulation, the performance of designed subsamples with protection was compared to those of subsamples without protection and to simple random subsamples when the true model is a second order model. Whereas designed subsamples without protection grossly inflated the mean squared errors of $\bar{Y}_C$ and $\bar{Y}^*_{\ell}$ compared to simple random subsamples, the designed subsamples with protection generally yielded lower mean squared errors than simple random subsamples. In subsamples with protection against a second order model, the estimators, $\bar{Y}_C$ and $\bar{Y}^*_{\ell}$, are comparable. Although the designed subsamples without protection are not robust in the presence of a second order model, allowing the subsample size n to approach the initial sample size n' reduces the difference between mean squared errors from designed subsamples without protection and those from simple random subsamples and subsamples with protection.

Given that designed subsamples with protection work well in the presence of a second order model while designed subsamples without protection can perform very badly, we recommend designed subsamples with protection if one decides to use a purposive subsampling plan. This recommendation is strengthened by the fact that although the designed subsamples with protection may lose some efficiency compared to designed subsamples without protection when the true model is a first order model, they are repeatedly more efficient than simple random subsamples.

Additional research is required to determine what protection designed subsamples safeguarding against a second order model afford the estimators, $\bar{Y}_C$ and $\bar{Y}_{\ell}^{*}$, when the true model is a higher order polynomial, e.g., a third order model.

The preceding results comparing the three subsampling procedures have been obtained under the usual conditions of normality of variables and homogeneity of variances of error. In these results, designed subsamples with protection against a second order model was the best procedure among the three subsampling procedures investigated in this study. It was of interest in this study to find out whether designed subsamples with protection against a second order model would still prevail under nonordinary conditions. Two such situations were considered. The three subsampling schemes were compared when the variates were bivariate lognormal and when the variances of the errors were proportional to the value of the auxiliary variable.

When the variables were bivariate lognormals, although designed subsamples with protection against a second order model do not always outperform simple random subsamples or designed subsamples without protection, all in all, they are a better strategy than either simple random subsamples or designed

subsamples without protection. When the variances of the errors are proportional to the auxiliary variable, both designed subsamples with and without protection performed favorably compared to simple random subsamples with respect to mean squared errors. Neither purposive sampling nor $\bar{Y}_{\ell}^{*}$ was adversely affected by the nonconstant error variance. Although theory suggests the use of weighted regression coefficients in $\bar{Y}_{C}$ and $\bar{Y}_{\ell}^{*}$ in the presence of heterogeneous error variances, simulations indicate that the resulting gain in efficiency is of no practical value.

The purposive subsampling plans have shown themselves to be also effective under the two nonstandard conditions considered in this study. However, further research is needed to determine how the designed subsampling strategy behaves under more extreme conditions.

5.3 CONCLUSIONS AND RECOMMENDATIONS

Based on a model-dependent approach, our research leads us to the following conclusions. Designed subsamples without protection based on the assumption of a first order linear model render $\bar{Y}_{C}$ and $\bar{Y}_{\ell}^{*}$ very susceptible to the presence of a second order linear model with respect to their mean squared errors. This confirms a major argument of randomization sample theorists that model-dependent designs are very sensitive to model misspecifications. However, this by no means indicates the dissolution of purposive sampling because subsamples can be deliberately chosen to guard against model misspecification. In fact, modelers have concentrated their efforts in making model-dependent designs more robust as discussed in Chapter 1. This study contributes to this effort by presenting a subsampling design that protects $\bar{Y}_{C}$ and $\bar{Y}_{\ell}^{*}$ against a second order model. Simulation results have shown it to be not

only robust to the presence of a second order model but also fairly robust to other forms of deviations such as variance misspecification and lognormal distribution of the variates. Thus, we recommend that when using either of the double sample regression estimators, $\bar{Y}_C$ or $\bar{Y}^*_\ell$, one should design subsamples with protection against a second order model unless one is positive that the true model is a first order linear model, in which case, designed subsamples without protection would be more efficient.

Under simple random subsampling, we recommend the use of $\bar{Y}^*_\ell$ for small subsample sizes when one is uncertain of the strength of the linear relationship.

A limitation of this study is that only one auxiliary variable has been considered. In actuality, sample surveys usually include several variables wherein a variable of interest may involve more than one auxiliary variable. However, there are countless situations when analysts given limited resources are required to give estimates of attributes within a specified brief period of time. If an attribute is very expensive to measure or rare, and an auxiliary variable is readily available at a much lesser cost, then the results of this study would be of great use to analysts in such situations.

REFERENCES

Bose, Chameli (1943). Note on the sampling error in the method of double sampling. *Sankhya*, *6*, 329-330.

Brewer, K. R. W. (1979). A class of robust sampling designs for large-scale surveys. *Jour. Amer. Statist. Assoc.*, *74*, 911-915.

Cochran, W. G. (1942). Sampling theory when the sampling-units are of unequal sizes. *Jour. Amer. Statist. Assoc.*, *37*, 199-212.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. New York: John Wiley and Sons.

Cox, D. R. (1952). Estimation by double sampling. *Biometrika*, *39*, 217-227.

Cox, D. R. and Hinckley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cumberland, W. G. and Royall, R. M. (1981). Prediction models and unequal probability sampling. *J. Royal Statist. Soc. B.*, *43*, 353-367.

Des Raj (1965). On a method of using multiauxiliary information in sample surveys. *Jour. Amer. Statist. Assoc.*, *60*, 270-277.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *J. Royal Statist. Soc. B*, *31*, 195-234.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J. Royal Statist. Soc. B*, *17*, 269-278.

Godambe, V. P. (1966). A new approach to sampling from finite populations. *J. Royal Statist. Soc. B*, *28*, 310-328.

Herson, Jay (1976). An investigation of relative efficiency of least squares prediction to conventional probability sampling plans. *Jour. Amer. Statist. Assoc.*, *71*, 700-703.

Iachan, R. (1985). Robust designs for ratio and regression estimation. *J. of Statist. Planning and Inference*, *11:2*, 149-161.

Kalbfleisch, J. D. and Sprott, D. A. (1969). Applications of likelihood and fiducial probability to sampling finite populations. In N. L. Johnson & H. Smith, Jr. (Eds.), *New Developments in Survey Sampling*. New York: Wiley Interscience.

Khan, S. and Tripathi, T. P. (1967). The use of multivariate auxiliary information in double sampling. *J. Indian Statist. Assoc.*, *5*, 42-48.

Kruskal, W. and Mosteller, F. (1980). Representative sampling, IV: the history of the concept in statistics, 1895-1939. *International Statistical Review*, *48*, 169-195.

Mostafa, M. D. and Mahmoud, M. W. (1964). On the problem of estimation for the bivariate lognormal distribution. *Biometrika*, *51*, 522-527.

Myers, R. H. (1976). *Response Surface Methodology*. Distributed by Edwards Brothers, Inc., Ann Arbor, Michigan.

Neyman, Jerzy (1934). On the two different aspects of the respresentative method: the method of stratified sampling and the method of purposive selection. *J. Royal Statist. Soc.*, *97*, 558-606.

Neyman, Jerzy (1938). Contribution to the theory of sampling human populations. *Jour. Amer. Statist. Assoc.*, *33*, 101-116.

Olkin, I. (1958). Multivariate ratio estimation for finite population. *Biometrika*, *45*, 154-165.

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

Royall, R. M. (1976a). Current advances in sampling theory: implications for human observational studies. *Amer. Jour. of Epidemiology*, *104*, 463-474.

Royall, R. M. (1976b). Linear least-squares prediction approach to two-stage sampling. *Jour. Amer. Statist. Assoc.*, *71*, 657-664.

Royall, R. M. (1983). "Comment" on An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Jour. Amer. Statist. Assoc.*, *78*, 794-796.

Royall, R. M. and Cumberland, W. G. (1978). Variance estimation in finite population sampling. *Jour. Amer. Statist. Assoc.*, *73*, 351-358.

Royall, R. M. and Cumberland, W. G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Jour. Amer. Statist. Assoc.*, *76*, 66-77.

Royall, R. M. and Cumberland, W. G. (1981b). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Jour. Amer. Statist. Assoc.*, *76*, 924-930.

Royall, R. M. and Eberhardt, K. R. (1975). Variance estimates for the ratio estimator. *Sankhya C*, *37*, 43-52.

Royall, R. M. and Herson, J. (1973a). Robust estimation in finite populations I. *Jour. Amer. Statist. Assoc.*, *68*, 880-889.

Royall, R. M. and Herson, J. (1973b). Robust estimation in finite populations II: stratification on a size variable. *Jour. Amer. Statist. Assoc., 68*, 890-893.

Scott, A. J., Brewer, K. R. W., and Ho, E. W. H. (1978). Finite population sampling and robust estimation. *Jour. Amer. Statist. Assoc., 73*, 359-361.

Scott, A. J. and Smith, T. M. F. (1969). Estimation in multi-stage surveys. *Jour. Amer. Statist. Assoc., 64*, 830-840.

Smith, T. M. F. (1976). The foundations of survey sampling: a review. *J. Royal Statist. Soc. A, 139*, 183-204.

APPENDIX I

FORTRAN SUBROUTINES FOR DESIGNING SUBSAMPLES

A. *****DESIGN SUBROUTINE*****

```
      SUBROUTINE DESIGN(N,NX,A,B,XPMEAN,SAMP1,SAMP2)
      DIMENSION A(n'),B(n'),SAMP1(n),SAMP2(n)
      DOUBLE PRECISION SS1,SS2,XM1,XM2,C1,C2,XPMEAN,TXN1,TXN2,SXN1,SXN2
      ICTR=1
      JCTR=NX
      CO1=ABS(A(1)-XPMEAN)
      CO2=ABS(A(NX)-XPMEAN)
      IF(CO1.GE.CO2) GO TO 60
      SAMP1(1)=A(NX)
      SAMP2(1)=B(NX)
      JCTR=JCTR-1
      GO TO 61
60    SAMP1(1)=A(1)
      SAMP2(1)=B(1)
      ICTR=ICTR+1
61    TXN1=SAMP1(1)
      TXN2=SAMP1(1)
      SXN1=TXN1**2
      SXN2=TXN2**2
      DO 53 J=2,N
      SX=A(ICTR)
      BX=A(JCTR)
      TXN1=TXN1+SX
      TXN2=TXN2+BX
      SXN1=SXN1+SX**2
      SXN2=SXN2+BX**2
      FJ=FLOAT(J)
      XM1=TXN1/FJ
      XM2=TXN2/FJ
      SS1=SXN1-TXN1**2/FJ
      SS2-SXN2-TXN2**2/FJ
      C1=(XM1-XPMEAN)**2/SS1
      C2=(XM2-XPMEAN)**2/SS2
      IF(C1.LE.C2) TO TO 70
      SAMP1(J)=BX
      SAMP2(J)=B(JCTR)
      JCTR=JCTR-1
      TXN1=TXN2
      SXN1=SXN2
      GO TO 53
70    SAMP1(J)=SX
      SAMP2(J)=B(ICTR)
      ICTR=ICTR+1
      TXN2=TXN1
      SXN2=SXN1
53    CONTINUE
      RETURN
      END
```

ARGUMENTS:

N	input variable containing the number of elements in the subsample
NX	input variable containing the number of elements in the initial sample
A	input vector, vector of X values of elements in the initial sample sorted from the smallest to largest value
B	input vector, vector of ID numbers corresponding to the X values found in vector A
XPMEAN	input variable containing the initial sample mean, X'
SAMP1	output vector, vector of X values of elements included in the designed subsample
SAMP2	output vector, vector of ID numbers of elements included in the designed subsample

B. *****PROTECTION SUBROUTINE*****

```
 1          SUBROUTINE PROTEC(NS,NP,X,ARY,XPBAR,SAMP1,SAMP2)
 2          DIMENSION X(n'),ARY(n'),SAMP1(n),SAMP2(n),Z(n'),Z2(n'),
           FIRZ(n'),AX(n'),AY(n')
 3          TZ=0.0
 4          DO 21 J=1,NP
 5          DIF=X(J)-XPBAR
 6          Z2(J)=DIF*DIF
 7       21 TZ=TZ+Z2(J)
 8          ZBAR=TZ/FLOAT(NP)
 9          DO 42 K=1,NP
10          IRZ(K)=K
11       42 Z(K)=ABS(Z2(K)-ZBAR)
12          CALL VSRTR(Z,NP,IRZ)
13          DO 23 L=1,NP
14          AY(L)=ARY(IRZ(L))
15       23 AX(L)=X(IRZ(L))
16          ICTR=1
17          MS=NS/2
18          MS1=MS+1
19          DO 49 JK=1,MS
20       47 IF(AX(ICTR).LE.XPBAR) GO TO 48
21          ICTR=ICTR+1
22          GO TO 47
23       48 SAMP1(JK)=AX(ICTR)
24          SAMP2(JK)=AY(ICTR)
25          ICTR=ICTR+1
26       49 CONTINUE
27          NCTR=1
28          DO 51 ML=MS1,NS
29       50 IF(AX(NCTR).GT.XPBAR) GO TO 53
30          NCTR=NCTR+1
31          GO TO 50
32       53 SAMP1(ML)=AX(NCTR)
33          SAMP2(ML)=AY(NCTR)
34          NCTR=NCTR+1
35       51 CONTINUE
36       88 RETURN
37          END
```

ARGUMENTS:

NS input variable containing the number of elements in the subsample

NP input variable containing the number of elements in the initial sample

X input vector, vector of X values of elements in the initial sample

ARY input vector, vector of ID numbers corresponding to the X values found in input vector X

XPBAR input variable containing the initial sample mean, $\bar{X}'$

SAMP1 output vector, vector of X values of elements included in the designed subsample with protection against a second order model

SAMP2 output vector, vector of ID numbers of elements included in the designed subsample with protection against a second order model

Note: The subroutine PROTEC uses the IMSL subroutine VSRTR.

APPENDIX II

SECOND ORDER POLYNOMIALS

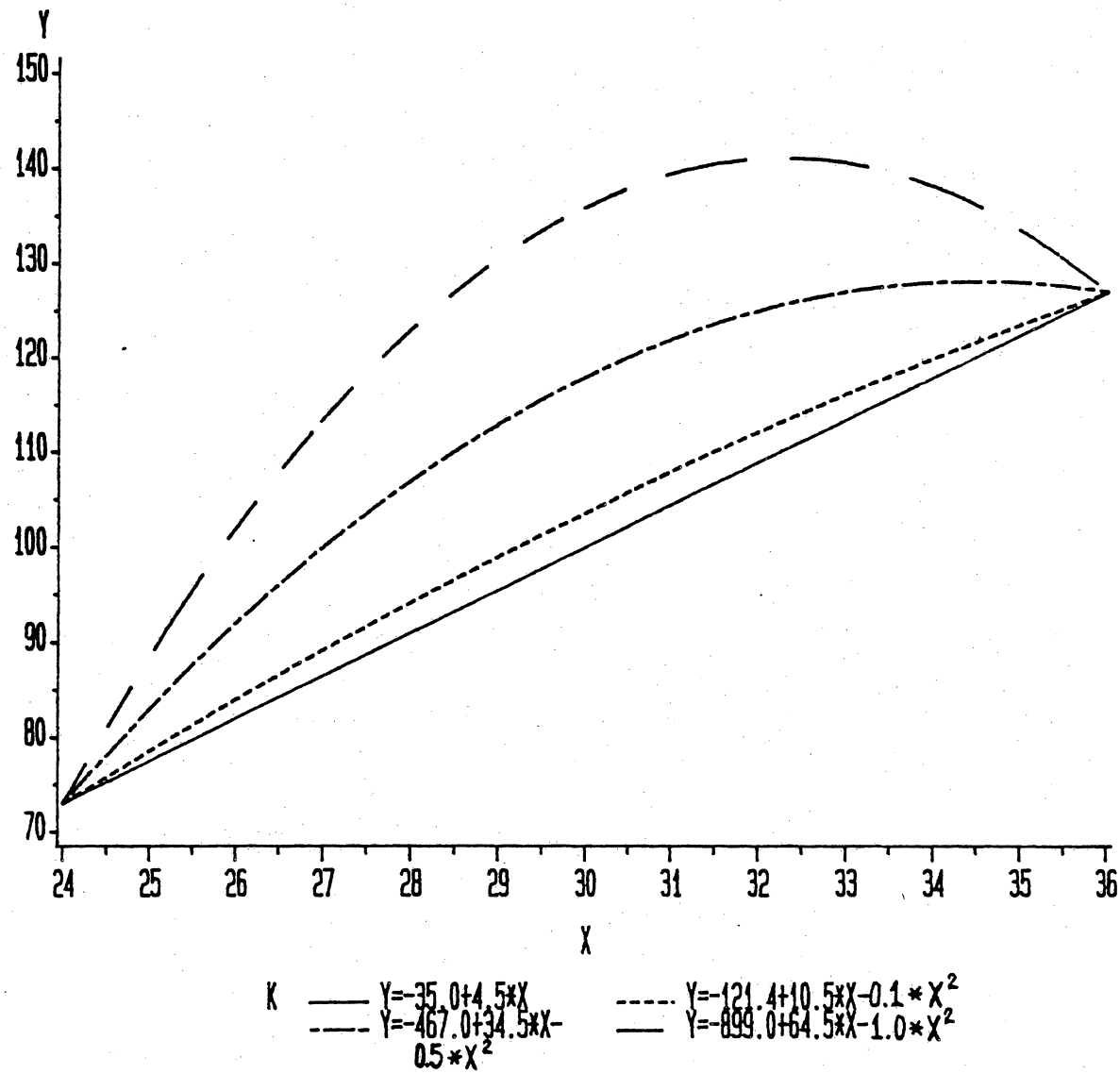

Figure A.1 Model Deviations from Y = -35.0 + 4.5X

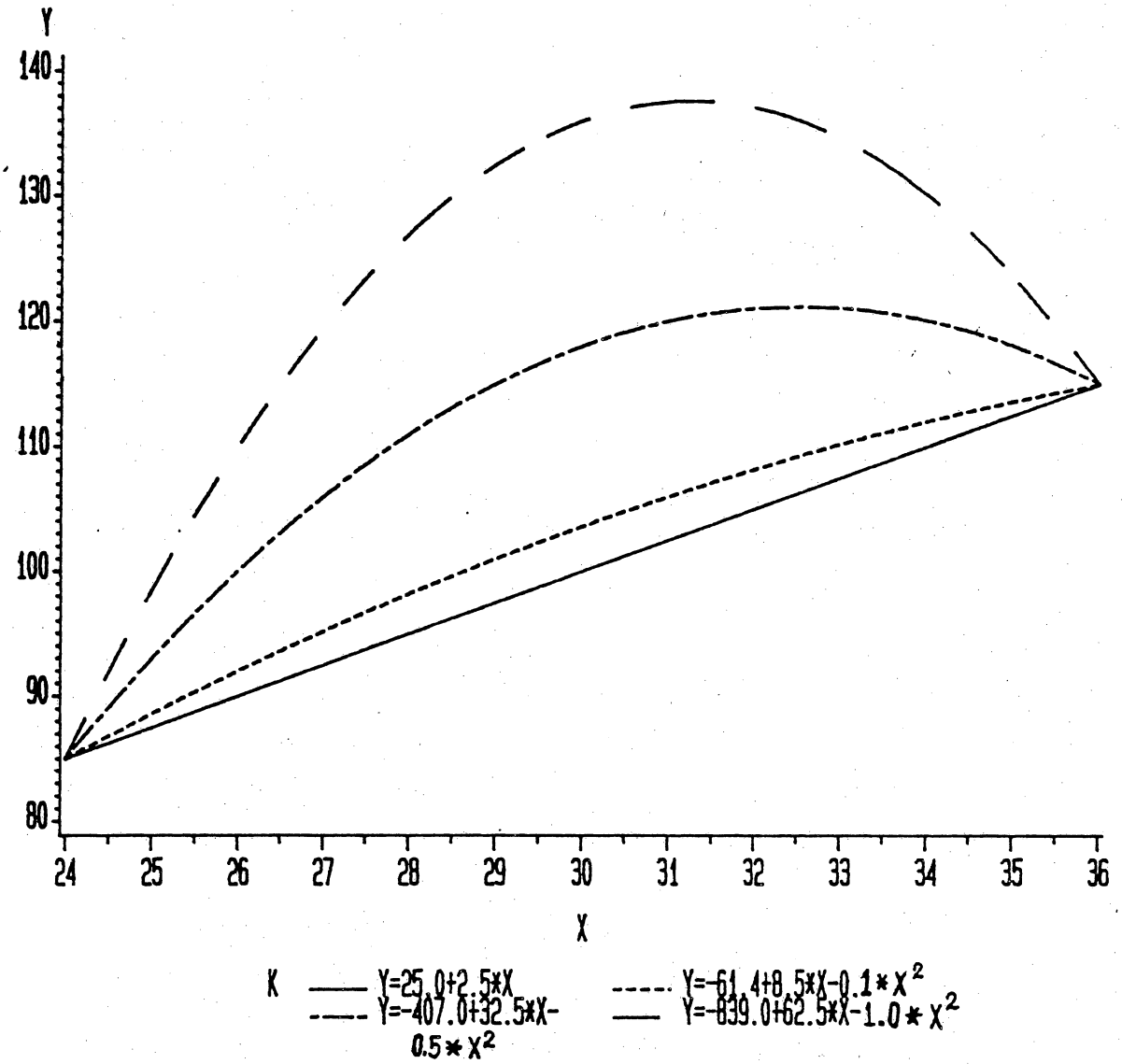

Figure A.2 Model Deviations from Y = 25.0 + 2.5X

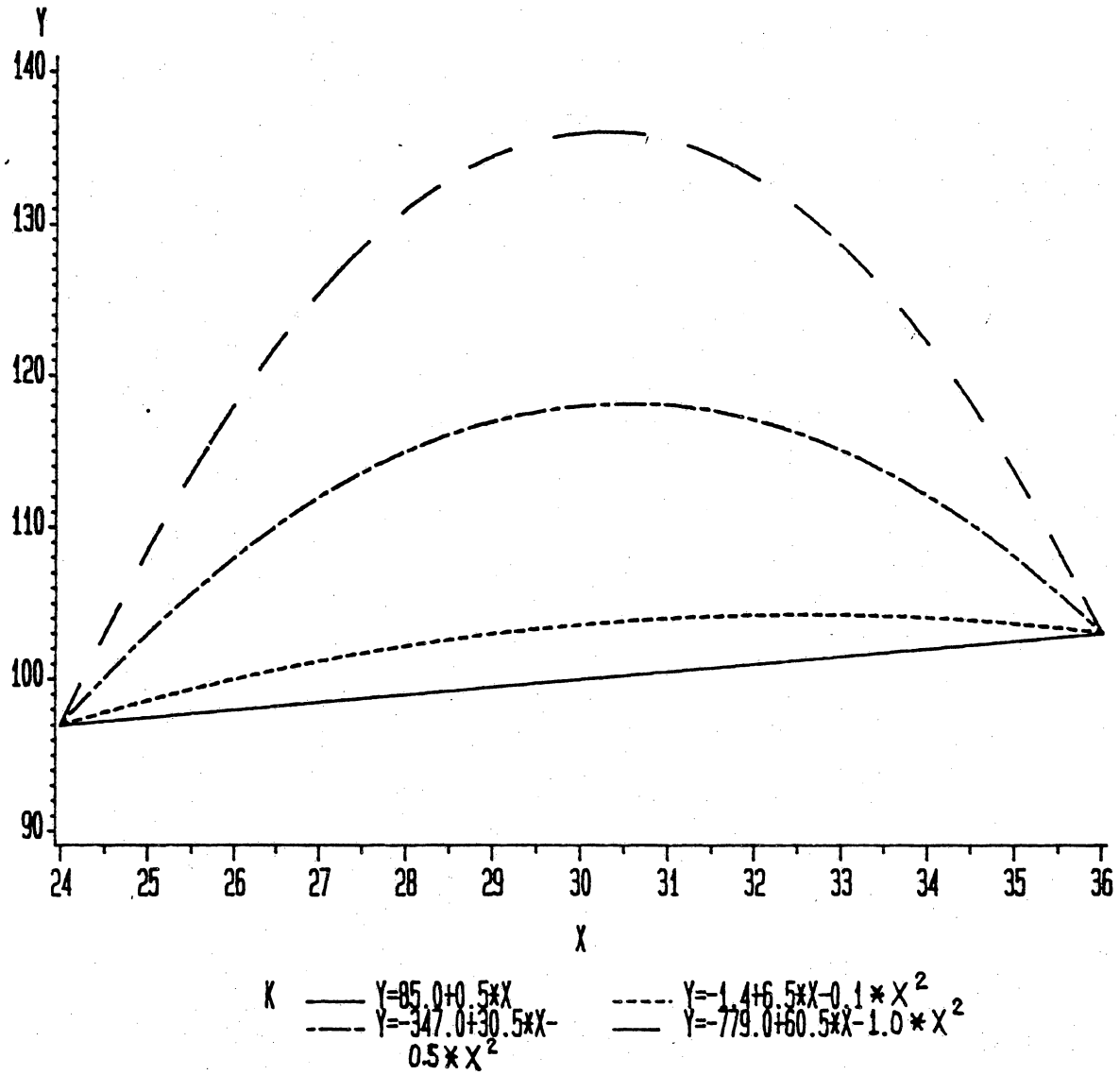

Figure A.3 Model Deviations from Y = 85.0 + 0.5X

The vita has been removed from the scanned document