INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Nonparametric Prediction in Survey Sampling and its Application to the Nonresponse Problem

Anthony N. Crisalli

A Thesis

in

Special Individualized

Program

Presented in Partial Fulfilment of the Requirements

for the Degree of Doctor of Philosophy at

Montreal, Quebec, Canada

August 1999

ABSTRACT

Nonparametric Prediction in Survey Sampling and its Application to the Nonresponse Problem

Anthony N. Crisalli, Ph.D.

Concordia University, 1999

Nonparametric regression provides an important tool towards exploring the relationship between a dependent variable and the independent variable(s) without assuming a functional form between the variables. This thesis incorporates the nonparametric regression methodology in the context of estimation of a finite population mean and/or total. The resulting estimator is called the generalized smoothing (GSE) in contrast to the, model based generalized regression estimator (GRE), developed by Särndal (1972). The theory available for GRE is extended for the GSE. We replace model-based predictors by those based on nonparametric regression.

The first objective is to investigate the merits of GSE over GRE in the simple random samples with respect to different criteria. It is shown that GSE is design-consistent and asymptotically design-unbiased. Furthermore, Monte Carlo simulations are carried out to compare the GSE and the GRE with respect to several criteria, such as the Bias and the Mean Square Error (MSE). The second objective is to extend the above methodology in situations of non-response. In this case a probability response model for the respondents is assumed; the case of each sample unit not responding with equal probability is considered as a particular case. Monte Carlo simulations of the proposed procedures are presented so as to understand the behavior the GSE, and the probability response model.

The linear regression model is ubiquitous in the sample survey literature, an assumption that is untenable in practice. In this dissertation we illustrate that it is feasible to estimate a finite population mean (also in the presence of nonresponse) with a nonparametric regression model. The simulations demonstrate in both the cases of full response and nonresponse, that if the underlying point scatter is linear then the GRE is the best choice. On the other hand if the point scatter is not linear, a situation often met in practice, the GSE seems to outperform the GRE, in terms of the benchmark criteria of MSE and bias.

Acknowledgements

The author would like to express his deepest appreciation to Professor Yogendra P. Chaubey for his guidance, support and patience. Despite his many responsibilities he always found time to direct and encourage me in my endeavors. Working under his supervision is truly an enriching experience.

There were times in the process of writing this dissertation that I had made some critical errors and thought that it might not be possible to make substantial contributions to the field. Professor Chaubey however encouraged me to go on and "do what I can do". His confidence in my ability gave me courage to finish my dissertation.

I wish to express my gratitude to the Mathematics and Statistics Department for giving me the opportunity to pursue graduate studies. Throughout the writing of this dissertation I always had the unwavering support, and friendship of Professors José Garrido, Tariq N. Srivastava and Fassil Nebebe.

Above all, I cannot find words to express my warmest thanks to my dear wife Anna and to two people most dear to us, Claudia and Carlo. Anna and my children encouraged me to do graduate work. They stood by me, endured with me, and their love gave me the strength to complete this task.

CONTENTS

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 6

Chapter 1

Introduction

1.1 Model Assisted Survey Sampling

The last twenty five years has seen a large body of research done with respect to inferential methodologies in finite populations. The thrust of this emphasis is the non applicability of standard methods developed for infinite populations. A finite population may not be satisfactorily characterized by a few parameters, in general, in contrast to infinite populations. However, in practice, since a few parameters such as the mean or variance were of interest, probability mechanisms for generating a sample did provide valid methodology of point estimation (as in Neyman 1934). The basic concern at this juncture was the error of the point estimator, and therefore, concepts such as the bias, standard error, variance, coefficient of variation and mean square error figured prominently in the paradigm. Various sampling mechanisms were developed in order to reduce the sampling variation in the estimator.

Godambe (1955) wrote a monumental paper "*A unified theory of sampling from finite populations*", in which he showed that no '*uniformly best linear unbiased estimator*' exists to estimate a finite population mean or total if one only considers a probability design. Starting with this paper a gradual shift in research activity was initiated. Many concepts that were external to survey sampling were borrowed from the general theory

of mathematical statistics and introduced to the sample survey theory. This body of knowledge has come to be known as Superpopulation model theory, Prediction theory and Model assisted designs. The Superpopulation model theory comes in many flavors. Some theoretical survey statisticians (Särndal, 1972, 1976) assume a parametric model for the underlying population and at the same time use the probability mechanism used to gather the sample to estimate the population characteristics under consideration. On the other hand, we have methods advanced by Royall (1968, 1970a, 1970b) that only assume a parametric model for the finite population under consideration and makes no use of the sampling mechanism that generated the sample. Finally, we have Bayesian methods (Ericson, 1969, Rubin, 1983) which also assume a parametric model but make their inferences by using a prior distribution on the parameters. The common theme in all these inferential procedures is the assumption of a parametric model for estimating a population characteristic.

Hansen, Madow and Tepping (1981) presented a paper entitled "Foundations of Inference in Survey Sampling", which has become a corner stone in the survey sampling literature. The thesis of this paper was that only inferences based on probability sampling protected the survey statistician '*against failures of assumed models and provide robustness for all estimators*'. The authors also encouraged research in what they termed model dependent designs. They stated '*The proper use of models has much to contribute to survey design. We urge continuing strong efforts, taking the fullest feasible advantage of models, but ordinarily within the framework of probability sampling, i. e., using designs and estimators that are not model-dependent*'.

1.2 Nonresponse

One of the major problems in all surveys is that of nonresponse. Studies by sampling statisticians, market researchers, and sociologists have attempted to identify methods (sampling and nonsampling) that elicit a high response rate. A unified theory for the

nonresponse problem has not been developed to the same extent as the theory for sampling survey designs.

Nonresponse rates vary widely from survey to survey and from survey organization to survey organization. These rates change over time and even change for repetitions of the same survey. A sample response rate is defined as

$$R = \frac{r}{n}$$

where r is the number of individuals that respond to a survey and n is the number of individuals contacted through the survey. The sample nonresponse rate is defined as

$$R^c = 1 - R.$$

This measure is used to judge the success of a survey; the smaller R^c the better the survey because then the response rate is high. The Panel on Incomplete Data in Sample Surveys (P.I.D.S.S., 1983) came to the following conclusion "*nonresponse and refusal rates have appeared to increase in some surveys, with refusal rates increasing relative to the total nonresponse rates*". As an example, they compared the nonresponse rates for two U. S. federally funded surveys for the years 1968 to 1980. In Table 1 one easily observes that refusal rates have been increasing with time.

Table 1.1

Refusal Rates as a Function of Time

Year	Current Population Survey	Household Health Survey
1968	39%	26%
1969	39%	28%
1970	40%	26%
1971	43%	31%
1972	45%	36%
1973	44%	42%
1974	49%	52%
1975	54%	52%
1976	56%	55%
1977	61%	58%
1978	57%	55%
1979	59%	56%
1980	59%	56%

Source: Published data from the Bureau of the Census for the Current Population Survey and from the National Center for Health for the Household Health Interview Survey.

Many terms and definitions are currently being used to describe various aspects of the non-response problem. The P.I.D.S.S. has addressed this problem and dressed a glossary of non-sampling error terms. We present standardized terminology as developed by P.I.D.S.S.:

Definition 1.2.1 ***Non-response*** *refers to missing or incomplete information due to any of the following;*

a) the respondent is not available when the data is being collected,

b) the respondent refuses to cooperate with the survey mechanism,

c) the respondent gives a partial answer to questions on the survey.

Definition 1.2.2 ***Unit nonresponse*** *occurs if a unit is selected for the sample and is eligible for the survey but no response is obtained for the unit or the information obtained is unusable.*

Definition 1.2.3 ***Item nonresponse*** *occurs if questions that should be answered are not answered or if answered are classified as unusable. Item nonresponse may be due to any of the following;*

a) the respondent does not have the information needed for one or more questions,

b) the respondent refuses to answer a specific question,

c) the interviewer or respondent skip the question.

1.3 Approaches to the Nonresponse Problem

Statisticians fear that nonrandom response will introduce a severe bias effect in estimates derived from incomplete survey data. One school of thought contends that approaches to the nonrandom response problem should be in the framework of the classical randomization theory traditionally used to analyze sample survey data. The second school of thought argues that the problem would be better handled by finding new methods based on probability models.

Many authors have considered the merits and demerits of each approach and have offered their own suggestions. Bailar, Bailey and Corby (1977) deplore the lack of '*a sound statistical basis for the adjustment procedures*' currently being used for analyses of nonresponse data and welcome '*new simplified methods based on statistical models.*' Rubin (1977) uses a Bayesian argument to predict results for nonrepondents given the respondents' data. Methods based on randomization are due to Ernst (1978) and Bailar and Bailar (1978). These authors have used bias adjustments techniques, namely imputation. The two approaches have been combined by Cassel, Särndal and Wretman (1979). These authors use models to represent the point scatter and then base their analyses of these models on traditional foundations. Griliches, Hall and Hausman (1977), Hausman and

Spence (1977), Little (1983) and Nordheim (1979) have all used probability models for the respondents. These authors regard nonrandom nonresponse as a stochastic censoring process and have used the probit model to represent the censoring mechanism.

1.4 Nonparametric Statistics

While survey statisticians were debating their philosophical foundations, mathematical statisticians were creating a body of knowledge called '*Density estimation, Nonparametric Regression, and General Additive Models*'. All are nonparametric inferential procedures. The basic philosophy as stated by Eubank (1988) for all these inferential procedures is to let '*the data speak for itself*' or as stated by Hastie and Tibshirani (1995) '*let the data show us the appropriate functional form*' without making parametric model assumptions.

Traditional simple regression analysis has its foundations on a known parametric model for the relationship between a response variable y and a predictor variable x. The simplest form of the relationship is the classical linear model

$$y = \alpha + \beta x + \varepsilon,$$

where ε is an error term. The inferential procedure then involves finding estimates of the unknown parameters α and β. The modern regression analysis makes no assumptions on the parametric form of the model. In particular one assumes that

$$y = \mu(x) + \varepsilon,$$

where $\mu(.)$ is an unknown smooth function. The inferential procedure then involves estimating the functional form of the relationship between the response and explanatory variable. The basic idea behind all of these methods is to fit a model to the data points locally. The model will only depend on the observations at a given point and on some specified neighboring points. The fitted model produces estimates of the response variable

that are less variable than the known responses, the fitted values are known as *smooth predictions* and the methods used to create such fits are called *scatterplot smoothers*. Even if the underlying model is linear, smoothing methods are still useful because they enhance the underlying structure of the data without reference to a parametric model. Missing data is a problem encountered in all statistical endeavors. Some response variables may not have been recorded or recorded incorrectly. Smoothing interpolates the missing data between adjacent points, whereas parametric methods would interpolate all the observations.

Some of the more popular nonparametric regression methods are those based on kernel functions, spline functions and wavelets. Each of these methods have their own strengths and weaknesses but kernel functions have the advantage of mathematical simplicity. For multiple regression models Hastie and Tibshirani (1985) developed a multivariate version of the *scatterplot smoothers* which are called *'Generalized Additive models'*.

1.5 Overview of the Dissertation

This thesis introduces the nonparametric regression method to the survey sampling literature. The generalized smoothing estimator is described in Chapter 2, which is a modification of the generalized regression estimator (see Särndal, 1972, 1976). The bias, variance and asymptotic normality of the estimator are developed in this chapter. We demonstrate that the estimator is design-consistent and asymptotically design-unbiased. Chapters 3 and 4 consist of Monte Carlo simulations of the proposed estimators. Hence we will compare estimates of the population mean by contrasting parametric and nonparametric models. The sampling design used in Chapter 3 is *simple random sampling without replacement,* while in Chapter 4 we consider *stratified simple random sampling without replacement.*

The thrust of Chapter 5 is to develop methods that reduce the bias incurred because of nonresponse. The theory developed in Chapter 2 is now adapted for the nonresponse

problem. The response probabilities are estimated by binary regression and these estimates are then used to estimate the population mean. Chapter 6 will compare estimates of the population mean when nonresponse occurs in a sample by contrasting parametric and nonparametric models . The sampling design used throughout this chapter is *simple random sampling without replacement.* In order to understand the behavior of these different estimates the populations under investigation will have known point scatters. The final chapter, discusses some topics for further research.

Chapter 2

The Generalized Smoothing Estimator

2.1 Introduction

Let the characteristic y of a population be related to the characteristic x, through a smooth function $\mu(.)$, i.e.

$$y = \mu(x) + \varepsilon, \tag{2.1}$$

where ε represents the unknown error variable. In the superpopulation model context, the function $\mu(x)$ represents the regression of Y on X for the bivariate random variable (X, Y) i.e.

$$\begin{aligned} \mu(x) &= E(Y|X) \\ &= \frac{\int y f(x,y) dy}{f(x)}. \end{aligned}$$

Here $f(x,y)$ is the joint density of the bivariate random variable (X,Y) and $f(x)$ the marginal density of the random variable X. We use this interpretation for developing the Generalized Smoothing Estimator. Särndal (1972, 1976), considered $\mu(.)$ to be known in the form of a linear regression model and used predicted values of y in forming a generalized regression estimator. We feel that the assumption of the knowledge of the

function $\mu(.)$ may be very restrictive and hence wish to predict y in the absence of such a knowledge. In this respect, the theory of nonparametric regression provides a versatile method for exploring a general relationship between variables.

First we provide a review of kernel nonparametric regression along with its properties in section 2.2. In section 2.3 the spline regression model is described, its properties are presently not as well researched as those of the kernel nonparametric regression. The generalized smoothing estimator is described in section 2.4. The estimator is shown to be design-consistent and asymptotically design-unbiased in section 2.5, while its anticipated mean squared error is found in section 2.6. The optimal inclusion probabilities for the estimator will be found in section 2.7 by using the anticipated mean squared error of section 2.5. In section 2.8 the bias and variance of the estimator will be developed, then in section 2.9 a central limit theorem will demonstrate that the estimate has asymptotically a normal distribution. Finally a nonparametric lack of fit test for the simple regression model is discussed in section 2.10.

2.2 Kernel Nonparametric Regression

Smoothing a point scatter (x_k, y_k), involves the approximation of the mean response curve $\mu(x)$, in the regression relationship

$$y_k = \mu(x_k) + \varepsilon_k, \text{ where } k = 1, ..., n.$$

The goal of kernel nonparametric regression is to estimate $\mu(x)$ by a '*local averaging*'. The average will be constructed in such a way that it is defined only in small neighborhoods around x_k. A procedure that has received much attention in the literature is the kernel smoother due to Nadaraya (1964) and Watson (1964) given by

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i y_i,$$

where the function $w_i(x_k)$ is defined by the following:

$$w_i = \frac{K\left(\frac{x\_x_i}{b}\right)}{\sum_{i=1}^{n} K\left(\frac{x\_x_i}{b}\right)}. \tag{2.2}$$

The function $K(.)$ is called the *kernel function*, and has the following properties:

$$K(u) \geq 0 \text{ and continous for all } u, \tag{2.3}$$

$$\int_{-\infty}^{\infty} K(u)du = 1, \tag{2.4}$$

$$K(-u) = K(u) \text{ for all } u \ (K \text{ is a symmetric function about the origin}). \tag{2.5}$$

The parameter b is called the *bandwidth* also known as the *window-width parameter* or the *smoothing parameter*. The value of b determines the level of smoothness. Small values of b reproduce the data while large values give us the sample average of the response variable Y_k. Geometrically this means that small values of b produce curves that are wiggly while large values generate smooth curves (Hastie and Tibshirani 1995). How should b be chosen? Theoretical and data - driven algorithms for optimizing b are described in this section.

The following are the most widely used and studied kernels:

Table 2.1

Kernels used in Smoothing

Kernel	K(u)
Uniform	$\frac{1}{2} I\left(\lvert u\rvert \leq 1\right)$
Triangle	$\left(1-\lvert u\rvert\right) I\left(\lvert u\rvert \leq 1\right)$
Epanechnikov	$\frac{3}{4}(1-u^2) I\left(\lvert u\rvert \leq 1\right)$
Quartic	$\frac{15}{16}\left(1-u^2\right))^2 I\left(\lvert u\rvert \leq 1\right)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$
Triweight	$\frac{35}{32}(1-u^2)^3 I\left(\lvert u\rvert \leq 1\right)$
Cosinus	$\frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right) I\left(\lvert u\rvert \leq 1\right)$

Nonparametric regression on a single predictor generalizes in a straight forward way to multiple predictors. For a multidimensional predictor variable $\mathbf{X}_i = (X_{i1,}, ..., X_{id})$ one uses a multidimensional product kernel function

$$K^*(\mathbf{u}_1, ..., \mathbf{u}_d) = \prod_{h=1}^{d} K_h(\mathbf{u}_h),$$

where $K_h(\mathbf{u}_h)$ is a kernel for predictor $\mathbf{X}_h$, $h = 1, ..., d$. The kernel weights are now defined as

$$w_i(\underline{\mathbf{x}}) = \frac{\prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{b_j}\right)}{\sum_{i=1}^{n}\prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{b_j}\right)}.$$

Then a multivariate version of the local mean regression estimator is

$$\hat{\mu}(\underline{\mathbf{x}}) = \sum_{k=1}^{n} w_i(\underline{\mathbf{x}}) y_k, \tag{2.6}$$

which is the multivariate fitted regression surface.

We now demonstrate that the kernel smoother is consistent and also find its pointwise

bias and mean squared error. The kernel smoother is the ratio of two random variables. In order to find the expectation and variance of the kernel smoother each term in the ratio must be analyzed separately. These are discussed here in infinite population context.

Let

$$\begin{aligned} p(x) &= \int y f(x,y) dy \\ &= \mu(x) f(x) \end{aligned}$$

where $f(x,y)$ is the joint density of the bivariate continuous random vector (X,Y) and $f(x)$ the marginal density of X. Moreover let

$$\begin{aligned} \hat{p}(x) &= \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x-x_i}{b}\right) y_i \\ &= \frac{1}{n} \sum_{i=1}^{n} K_b\left(x-x_i\right) y_i, \end{aligned}$$

where

$$K_b(u) = \frac{1}{b} K\left(\frac{u}{b}\right).$$

Therefore the Nadaraya-Watson estimator

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i y_i,$$

can be written as:

$$\hat{\mu}(x) = \frac{\hat{p}(x)}{\hat{f}(x)}, \tag{2.7}$$

where

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x-x_i}{b}\right).$$

Theorem 2.2.1 *(Nadaraya-Watson, 1964) The estimator $\hat{\mu}(x)$ is a consistent estimate of the regression curve $\mu(x)$ at every point of continuity of $\mu(x)$, as $n \to \infty, b \to 0$ and $nb \to \infty$ under the condition of the existence and continuity of $p''(x)$ and $f''(x)$ at x.*

Proof We first find the expected value of $\hat{p}(x)$.

$$
\begin{aligned}
E\left(\hat{p}(x)\right) &= E\left(\frac{1}{n}\sum_{k=1}^{n} K_b\left(x-x_i\right) y_i\right) \\
&= E\left(K_b\left(x-x_i\right) y_i\right) \\
&= \iint y K_b\left(x-t\right) f\left(y|t\right) f\left(t\right) dydt \\
&= \int K_b\left(x-t\right) f\left(t\right) \left[\int y f\left(y|t\right) dy\right] dt \\
&= \int K_b\left(x-t\right) f\left(t\right) \mu\left(t\right) dt \\
&= \int K_b\left(x-t\right) p\left(t\right) dt \\
&= \frac{1}{b}\int K\left(\frac{x-t}{b}\right) p\left(t\right) dt.
\end{aligned}
$$

Let

$$
-v = \frac{x-t}{b},
$$

therefore

$$
\begin{aligned}
E\left(\hat{p}(x)\right) &= \int K\left(-v\right) p\left(x+bv\right) dv \\
&= \int K\left(v\right) p\left(x+bv\right) dv,
\end{aligned}
$$

since $K(.)$ is a symmetric function about the origin. Now expand $p\left(x+bv\right)$ in a Taylor series about x,(assuming the continuity and existence of $p^{''}\left(x\right)$) we have,

$$
p\left(x+bv\right) = \sum_{j=0}^{t} \frac{\left(vb\right)^j}{j!} p\left(x\right) + o\left(b^t\right) = p\left(x\right) + vbp^{'}\left(x\right) + \frac{v^2b^2}{2} p^{''}\left(x\right) + o\left(b^2\right).
$$

Using the properties 2.3-2.5 of the Kernel function and the Taylor series expansion of $p\left(x+bv\right)$ the expected value of $\hat{p}(x)$ is given by:

$$
\begin{aligned}
E\left(\hat{p}(x)\right) &\approx \int K\left(v\right)\left(p\left(x\right) + vbp^{'}\left(x\right) + \frac{v^2b^2}{2} p^{''}\left(x\right) + o\left(b^2\right)\right) dv \\
&= p\left(x\right)\int K\left(v\right) dv + bp^{'}\left(x\right)\int vK\left(v\right) dv + \\
&\quad \frac{b^2}{2} p^{''}\left(x\right)\int v^2 K\left(v\right) dv + o\left(b^2\right) \\
&= p\left(x\right) + \frac{b^2}{2} p^{''}\left(x\right)\sigma^2 + o\left(b^2\right),
\end{aligned}
\tag{2.8}
$$

where

$$\sigma^2 = \int v^2 K(v)\, dv.$$

Analogous manipulations for the variance show that

$$Var(\hat{p}(x)) = \frac{r^2(x) f(x)}{nb} \alpha_K + o\left((nb)^{-1}\right), \tag{2.9}$$

where

$$\alpha_K = \int K^2(v)\, dv,$$

and

$$r^2(x) = E\left(Y^2 | X = x\right).$$

Now the mean squared error of $\hat{p}(x)$ is given by:

$$MSE(\hat{p}(x)) = \frac{r^2(x) f(x)}{nb} \alpha_K + \frac{b^4}{4}\left(p''(x)\sigma^2\right)^2 + o\left((nb)^{-1}\right) + o\left(b^4\right). \tag{2.10}$$

Therefore, if we let $n \to \infty, b \to 0$ and $nb \to \infty$ then

$$MSE(\hat{p}(x)) \to 0,$$

or $\hat{p}(x)$ is a consistent estimate of $p(x)$.

Using the similiar manipulations one can show that $\hat{f}(x)$ is a consistent estimate of $f(x)$. Using Slutzky's theorem (Roussas, 1997) we have that

$$\hat{\mu}(x) = \frac{\hat{p}(x)}{\hat{f}(x)} \xrightarrow{p} \frac{p(x)}{f(x)} = \mu(x)$$

as $n \to \infty, b \to 0$ and $nb \to \infty$ or $\hat{\mu}(x)$ is a consistent estimate of $\mu(x)$ at every x where $f(x) \neq 0$. ■

We have shown that $\hat{\mu}(x)$ is a *weakly consistent* estimate of $\mu(x)$. *Strong consistency* has also been derived under various conditions Nadaraya (1970), Devroye and Wagner

(1980). The Nadaraya (1970), theorem on strong consistency is now stated and will be used subequently.

Theorem 2.2.2 *(Nadaraya,1970) Let A be a fixed subset of $\Re$ and assume that*

1. *X has a continuous density f,*
2. $\inf_A f(x) > 0$,
3. $\mu(x)$ *is continuous on* $\Re$,
4. $|Y| \leq c < \infty$ *a.s.*,
5. $K(.)$ *is a bounded density on* $\Re$ *satisfying* $|x|\, K(x) \to 0$ *as* $|x| \to \infty$,
6. $K(.)$ *is of bounded variation*
7. $\sum_{n=1}^{\infty} \exp(-\alpha n b^2) < \infty$ *for all* $\alpha > 0$.

Then

$$\operatorname*{ess\,sup}_A \; |\hat{\mu}(x) - \mu(x)| \to 0 \text{ a.s.}$$

or $\hat{\mu}(x)$ *is a strongly consistent estimate of* $\mu(x)$, *if* $n \to \infty, b \to 0$ *and* $nb \to \infty$.

In order to find the bias and mean squared error of $\hat{\mu}(x)$, we linearize the estimate as

$$\begin{aligned} \hat{\mu}(x) - \mu(x) &= \left(\frac{\hat{p}(x)}{\hat{f}(x)} - \mu(x)\right)\left(\frac{\hat{f}(x)}{f(x)} + \left(1 - \frac{\hat{f}(x)}{f(x)}\right)\right) \\ &= \left(\frac{\hat{p}(x) - \mu(x)\,\hat{f}(x)}{f(x)}\right) + (\hat{\mu}(x) - \mu(x))\left(\frac{f(x) - \hat{f}(x)}{f(x)}\right). \end{aligned} \tag{2.11}$$

To find the leading term of the distribution of $\hat{\mu}(x) - \mu(x)$ we use the following concepts

Definition 2.2.1 *Let A_n and B_n be two real-valued random sequences then*

$$A_n = o_p(B_n)$$

if for all $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{A_n}{B_n}\right| > \epsilon\right) = 0.$$

Definition 2.2.2 *Let A_n and B_n be two real-valued random sequences then*

$$A_n = O_p(B_n)$$

if for all $\epsilon > 0$ there exist λ and M such that

$$P\left(\left|\frac{A_n}{B_n}\right| > \lambda\right) < \epsilon$$

for all $n > M$.

Using these definitions we can state that

$$\begin{aligned} \hat{p}(x) - \mu(x)\hat{f}(x) &= (\hat{p}(x) - p(x)) - \mu(x)\left(\hat{f}(x) - f(x)\right) \\ &= O_p(b^2) - \mu(x)O_p(b^2) \\ &= O_p(b^2), \end{aligned} \tag{2.12}$$

and

$$\begin{aligned} (\hat{\mu}(x) - \mu(x))\left(f(x) - \hat{f}(x)\right) &= o_p(1)O_p(b^2) \\ &= o_p(b^2). \end{aligned} \tag{2.13}$$

Therefore the leading term in the distribution of $\hat{\mu}(x) - \mu(x)$ is $\frac{\hat{p}(x) - \mu(x)\hat{f}(x)}{f(x)}$, because the second term is of smaller order in probability as $b \to 0$. These results allow us to state the following theorems:

Theorem 2.2.3 *The bias of the Nadaraya-Watson estimator $\hat{\mu}(x)$ is given by:*

$$E(\hat{\mu}(x) - \mu(x)) = \frac{b^2}{2}\left(\frac{\mu''(x)f(x) + 2\mu'(x)f'(x)}{f(x)}\right)\sigma^2 + o(b^2). \tag{2.14}$$

Proof It was previously shown that

$$E(\hat{p}(x)) = p(x) + \frac{b^2}{2}p''(x)\sigma^2 + o(b^2),$$

and

$$E\left(\hat{f}(x)\right) \approx f(x) + \frac{b^2}{2} f''(x)\sigma^2 + o(b^2).$$

Also since

$$\begin{aligned} p(x) &= \mu(x)f(x), \\ p''(x) &= \mu''(x)f(x) + 2\mu'(x)f'(x) + \mu(x)f''(x). \end{aligned}$$

Then

$$\begin{aligned} E\left(\hat{\mu}(x) - \mu(x)\right) &\approx \frac{E\left(\hat{p}(x) - \mu(x)\,\hat{f}(x)\right)}{f(x)} \\ &= \frac{E\left(\hat{p}(x)\right) - \mu(x)\,E\left(\hat{f}(x)\right)}{f(x)} \end{aligned}$$

and the result follows by making the appropriate substitutions. ■

Remark 2.2.1 *If the regression is linear the above bias reduces to*

$$E\left(\hat{\mu}(x) - \mu(x)\right) = b^2 \frac{\mu'(x)\,f'(x)}{f(x)} \sigma_v^2 + o(b^2).$$

Theorem 2.2.4 *The mean squared error of the Nadaraya- Watson estimator $\hat{\mu}(x)$ is given by:*

$$MSE\left(\hat{\mu}(x)\right) = \frac{1}{nb}\frac{\sigma^2(x)}{f(x)}\alpha_K + \frac{b^4}{4}\left(\frac{\mu''(x)\,f(x) + 2\mu'(x)\,f'(x)}{f(x)}\sigma^2\right)^2 + o(b^4) + o\left((nb)^{-1}\right). \tag{2.15}$$

The proof of this theorem is obtained through similar manipulations as that of Theorem 2.2.5.

Remark 2.2.2 *If the regression is linear the above mean squared error reduces to*

$$MSE\left(\hat{\mu}(x)\right) = \frac{1}{nb}\frac{\sigma^2(x)}{f(x)}\alpha_K + b^4\left(\frac{\mu'(x)\,f'(x)}{f(x)}\sigma^2\right)^2 + o(b^4) + o\left((nb)^{-1}\right).$$

The choice of bandwidth will have an effect on the precision of the estimate of $\mu(x)$. As b tends to get small the bias of $\hat{\mu}(x)$ will get smaller but the variance of $\hat{\mu}(x)$ will get larger. Similarly as b tends to get large the variance of $\hat{\mu}(x)$ will get smaller but the bias will get larger.

The trade-off can be minimized by choosing a b that minimizes the mean squared error. Using standard methods the optimal value of b is given by the following:

$$b_{opt} = \left(\frac{k_1}{nk_2}\right)^{\frac{1}{5}}, \tag{2.16}$$

where

$$\begin{aligned} k_1 &= \frac{\sigma^2(x)}{f(x)}\alpha_K, \\ k_2 &= \left(\frac{\mu^{''}(x)\, f(x) + 2\mu^{'}(x)\, f^{'}(x)}{f(x)}\sigma^2\right)^2. \end{aligned}$$

We should theoretically choose $b_{opt} \sim n^{\frac{1}{5}}$ (Härdle 1989, 1990) but the solution is not very helpful in practice because b_{opt} is a function of unknown parameters. Many methods have been proposed to solve this problem. Silverman (1986) showed that an approximate solution to the optimal bandwidth is given by

$$b_{opt} \approx 1.059An^{-\frac{1}{5}}, \tag{2.17}$$

such that

$$A = \min(S_x, IRQ/1.34),$$

where S_x is the sample standard deviation of the x's and IRQ is the sample interquartile range of the x's.

2.3 Spline Regression

The goal of spline regression is to minimize the penalized residual sum of squares for all units k in the sample s,

$$\sum_{k\in s}(y_k-\mu(x_k))^2+\lambda\int\left(\mu''(t)\right)^2dt, \tag{2.18}$$

over all functions $\mu(.)$ with continuous first and integrable second derivatives. The parameter λ represents the rate of exchange between the residual error and the roughness of the curve $\mu(.)$ and therefore is a smoothing parameter which has the same function as the *bandwidth* and is called the *span*.

The minimizer was found by Schoenberg (1964). The unique solution $\hat{\mu}(x_k)$ is a cubic spline and has the following properties:

a. A cubic polynomial fits the data between two successive sampled x_k values.

b. At the sampled values x_k, $\hat{\mu}(x_k)$ and its two first derivatives are continuous.

c. At the boundary points $x_{(1)}$ and $x_{(n)}$ the second derivatives of $\hat{\mu}(x_k)$ is zero.

Remark 2.3.1 *The following points about the smoothing parameter λ will be useful:*

a. Decreasing λ leads to a less smooth estimate of $\mu(x_k)$.

b. As $\lambda\to 0$, the spline smoother interpolates the sampled y_k values.

c. As $\lambda\to\infty$, $\int\left(\mu''(t)\right)^2dt$ has to be very small with respect to $\sum_{k\in s}(y_k-\mu(x_k))^2$.

Therefore the spline smoother approaches a linear function at the sampled x_k values.

The estimated spline smoother $\hat{\mu}(x_k)$ is a function of the spanning parameter λ in the penalized residual sum of squares

$$S_\lambda(\mu(.))=\sum_{k\in s}(y_k-\mu(x_k))^2+\lambda\int\left(\mu''(t)\right)^2dt.$$

Wahba and Wold (1975) recommend to cross-validating the sum of squares to find the

optimal value of λ. The procedure is to minimize:

$$CV(\lambda) = \frac{1}{n}\sum_{k\in s}\left(y_k - \hat{\mu}_{\lambda}^{-k}(x_k)\right)^2, \tag{2.19}$$

where $\hat{\mu}_{\lambda}^{-k}(x_k)$ denotes the fit at x_k by leaving out that data point. The optimal λ, is found by first computing $CV(\lambda)$ over a suitable range of λ values and then choosing the λ that minimizes $CV(\lambda)$. Let λ_{opt} be the value of λ that minimizes $CV(\lambda)$. We then use λ_{opt} to find $\hat{\mu}(x_k)$ which we shall now call $\hat{\mu}_{\lambda_{opt}}(x_k)$. The predicted y_k and residuals will now be found with the λ_{opt} i.e.

$$\hat{y}_{k\lambda_{opt}} = \hat{\mu}_{\lambda_{opt}}(x_k) \text{ for all } x_k \in U,$$

and

$$e_{k\lambda_{opt}} = y_k - \hat{\mu}_{\lambda_{opt}}(x_k) \text{ for all } x_k \in s.$$

where U is the population and s the sample chosen from U. Silverman (1986) has shown that the spline smoother is related to the equivalent kernel function

$$K(u) = \frac{1}{2}\exp\left(-\frac{|u|}{\sqrt{2}}\right)\sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right).$$

The precision of the spline was shown by Hastie and Tibshirani (1995) to be

$$s_{sp}^2 = \frac{1}{n - trace(\mathbf{S}_{\lambda_{opt}}\mathbf{S}'_{\lambda_{opt}})}\sum_{k\in s}\left(y_k - \hat{\mu}_{\lambda_{opt}}(x_k)\right)^2, \tag{2.20}$$

where $S_{\lambda_{opt}}$ is the symmetric projection matrix.

2.4 The Generalized Smoothing Estimator

Here we consider the case of only one explanatory variable. The material in this section generalizes in a straightforward way to multiple predictors.

Probability sampling is used to select from the finite population $U = \{1, 2, ..., N\}$ a sample s of fixed size n such that s is selected with probability $p(s) \geq 0$. Now $\sum_{s \in \mathcal{L}} p(s) = 1$ where $\mathcal{L}$ is the set of all samples of size n. The sampling design $p(s)$ is characterized by the first order inclusion probabilities, assumed to be positive for $k = 1, 2, ..., N$, $\pi_k = \sum_{s \ni k} p(s)$ where the sum is over all s having k as a member. We will use $E_p(.)$ and $V_p(.)$ to denote expected value and variance with respect to the design $p(s)$.

We now make assumptions about the shape of the finite population point scatter

$$\{(x_k, y_k) : k = 1, 2, ..., N\}.$$

The value of the explanatory variable is denoted by x_k, $k = 1, 2, ...N$ and is assumed to be known for all units in the population. This assumption is realistic on a macro level. The shape of the point scatter is assumed to be generated by a model such as in equation (2.1) called ξ. The assumption usually made is that the scatter of the N points looks as if it had been generated by a linear model ξ, with y as the response variable and x as the explanatory variable.

We generalize this assumption by assuming a model ξ having the following properties:

a. $y_1, y_2, ..., y_N$ are assumed to be realized values of independent random variables $Y_1, Y_2, ..., Y_N$,

b. $\mathcal{E}_\xi(Y_k) = \mu(x_k)$ for $k = 1, 2, ..., N$,

where $\mathcal{E}_\xi(.)$ denotes the expected value with respect to the model ξ, and $\mu(x_k)$ is an unknown functional form.

The goal is to estimate the unknown population mean

$$\bar{Y} = \frac{1}{N} \sum\nolimits_{k \in U} y_k, \tag{2.21}$$

when (x_k, y_k) has been observed for a sample s such that $k \in s$. Särndal (1972, 1976) demonstrated in a series of papers the efficiency and properties of the generalized regres-

sion estimator

$$\hat{m}_{yr} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{N}\sum_{k\in s}\frac{e_k}{\pi_k}, \tag{2.22}$$

where $e_k = y_k - \hat{y}_k$, $\hat{y}_k$ is the predicted value of y_k under the model ξ. A model ξ considered frequently in the literature is

$$\begin{aligned} \mathcal{E}_\xi(Y_k) &= \beta x_k, \\ \mathcal{V}_\xi(Y_k) &= \sigma_k^2. \end{aligned}$$

The generalized regression estimator under this model is

$$\hat{m}_{yr} = \frac{1}{N}\hat{t}_\pi + \frac{1}{N}\hat{B}\left(t_x - \hat{t}_{x\pi}\right), \tag{2.23}$$

where

$$\hat{t}_\pi = \sum_{k\in s}\frac{y_k}{\pi_k},$$

is the π estimator of $t_y = \sum_{k\in U} y_k$, the population total of the y values,

$$\hat{t}_{x\pi} = \sum_{k\in s}\frac{x_{jk}}{\pi_k},$$

is an estimate of the known x total, $t_x = \sum_{k\in U} x_k$ and

$$\hat{B} = \left(\sum_{k\in s}\frac{x_k^2}{\pi_k\sigma_k^2}\right)^{-1}\left(\sum_{k\in s}\frac{x_k y_k}{\pi_k\sigma_k^2}\right).$$

Särndal, C. E., Swensson, B. Wretman J. (1992), use this model and comment 'The role of the model is to describe the finite population point scatter. We hope that the model fits the population reasonably well. We think that the finite population looks as if it might have been generated in accordance with the model. However the assumption is never made that the population was really generated by the model. Our conclusion about the finite population parameters are therefore independent of model assumptions'.

This statement is a little questionable. First a model is used to estimate the finite population parameter B, which is then used to estimate the population total t_y. Are the

conclusions model-independent or do we '*look*' and '*hope that the population was really generated by the model*'? Researchers in survey sampling use a parametric approach for its simplicity in computation, for its compatibility with model assumptions and for mathematical tractability.

We ascertain that a better solution to this problem is to make no assumption regarding the point scatter. The motivation for the nonparametric regression approach as described by Härdle (1989, 1990) are as follows:

i. to provide a versatile method for exploring a general relationship between two variables,

ii. to give predictions for observations yet to be made without reference to a fixed parametric model,

iii. to furnish a tool for finding spurious observations by studying the influence of isolated points,

iv. to have a flexible method of substituting for missing values or interpolating between adjacent X values.

We adapt and modify equation (2.23). We replace $\hat{B}\left(t_x - \hat{t}_{x\pi}\right)$ by a nonparametric regression estimate. We call this estimator the **generalized smoothing estimate**, denote it by $\hat{m}_{sm}$ and define it as

$$\hat{m}_{sm} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{N}\sum_{k\in s}\frac{e_k}{\pi_k}, \tag{2.24}$$

where $\hat{y}_k = \hat{\mu}_{\pi}(x_k)$ and $e_k = y_k - \hat{y}_k$. Now $\hat{\mu}_{\pi}(x_k)$ is the sample kernel estimator and is defined as

$$\hat{\mu}_{\pi}(x_k) = \frac{\sum_{j\in s}\dfrac{K\left(\dfrac{x_k - x_j}{b}\right)y_j}{\pi_j}}{\sum_{j\in U}K\left(\dfrac{x_k - x_j}{b}\right)} = \frac{\sum_{j\in s}\dfrac{W_j(x_k)\,y_j}{\pi_j}}{\sum_{j\in U}W_j(x_k)}$$

where

$$W_j(x_k) = K\left(\frac{x_k - x_j}{b}\right).$$

The estimate $\hat{m}_{sm}$ is the mean of fitted values $\frac{1}{N}\sum_{k\in U}\hat{y}_k$ and an adjustment term $\frac{1}{N}\sum_{k\in s}\frac{e_k}{\pi_k}$.

2.5 Asymptotic Unbiasedness and Consistency of the Generalized Smoothing Estimator

Here we will demonstrate that $\hat{m}_{sm}$ is an asymptotically design-unbiased and design-consistent estimator of $\bar{Y}$. If an estimator $\hat{\theta}$ is asymptotically design-unbiased then it can be considered unbiased when the sample size is large. Also if $\hat{\theta}$ is design-consistent this implies that the sampling error for $\hat{\theta}-\theta$ is likely to be small for large n.

For the asymptotics considered here we use the mathematical formulation given by Issaki and Fuller (1981) and Robinson and Särndal (1983). Consider a sequence of populations $U_1, U_2, ...$where U_t consists of the first N_t units from the infinite sequence of populations i.e. $U_t = \{k : 1, 2, ...N_t\}$. It is assumed that $U_1 \subset U_2 \subset U_3...$ which implies that $N_1 < N_2 < N_3 < ....$ Also consider a probability sampling design $p_t(.)$ for each of the populations U_t. Now $p_t(.)$ gives every element of U_t a probability of being included in the sample s_t. Let π_{kt} and π_{klt} denote the inclusion probabilities of the k'th unit and joint inclusion of the (k,l) unit, $\{k, l = 1, 2, ...N_t\}$ associated with the design $p_t(.)$. Moreover assume that the design is a fixed effective design, i.e. the sample size n_t is fixed such that $n_1 < n_2 < n_3 < ....$ Now when $t \to \infty$, both $n_t \to \infty$ and $N_t \to \infty$.

Let

$$I_{kt} = \begin{cases} 1 \text{ for all } k \in s_t \\ 0 \text{ otherwise} \end{cases}$$

therefore

$$P(I_{kt} = 1) = \pi_{kt} \text{ and } P(I_{kt} = 0) = 1 - \pi_{kt},$$

and

$$\sum_{k\in U_t} I_{kt} = \sum_{k\in U_t} \pi_{kt} = n_t.$$

Now rewrite (2.24) as

$$\hat{m}_{sm} = \frac{1}{N}\sum_{k\in s_t}\frac{y_k}{\pi_{kt}} + \frac{1}{N}\sum_{k\in U_t}\hat{\mu}_{\pi}(x_k) - \frac{1}{N}\sum_{k\in s_t}\frac{\hat{\mu}_{\pi}(x_k)}{\pi_{kt}}, \tag{2.25}$$

and furthermore as,

$$M_t = \frac{1}{N_t}\sum_{k\in U_t}\frac{Y_k I_{kt}}{\pi_{kt}} - \frac{1}{N_t}\sum_{k\in U_t}\hat{\mu}_{\pi}(x_k)\left(\frac{I_{kt}}{\pi_{kt}} - 1\right).$$

Consider

$$\hat{m}^{o}_{sm} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_{kt}} + \frac{1}{N}\sum_{k\in U_t}\hat{\mu}_t(x_k) - \frac{1}{N}\sum_{k\in s_t}\frac{\hat{\mu}_t(x_k)}{\pi_{kt}}, \tag{2.26}$$

where $\hat{\mu}_t(x_k)$ is the finite population estimator for population U_t and is defined as

$$\hat{\mu}_t(x) = \frac{\sum_{j\in U_t} W_j(x) y_j}{\sum_{j\in U_t} W_j(x)} = \sum_{i\in U_t} w_i(x)\, Y_i,$$

where

$$w_i(x) = \frac{W_i(x)}{\sum_{j\in U_t} W_j(x)}.$$

Now

$$\hat{m}_{sm} - \hat{m}^{o}_{sm} = \frac{1}{N}\sum_{k\in U_t}\left(\hat{\mu}_{\pi}(x_k) - \hat{\mu}_t(x_k)\right) - \frac{1}{N}\sum_{k\in s_t}\frac{\hat{\mu}_{\pi}(x_k) - \hat{\mu}_t(x_k)}{\pi_{kt}}. \tag{2.27}$$

A definition of asymptotic design unbiasedness and design-consistency is now presented. Allow U_t to get larger by making $N_t \to \infty$, $n_t \to \infty$ as $t \to \infty$. The sample size $n_t \to \infty$ but not at the same rate as N_t. Let ξ denote the probability distribution of the infinite random variable $(Y_1, Y_2, ...)$ and $\bar{Y}_t$ denote the mean of the t'th population. In the following definitions '*probability one*' refers to the probability under the distribution

ξ.

Definition 2.5.1 *A predictor M_t is said to be asymptotically design-unbiased if*

$$\lim_{t\to\infty}\left(E_p\left(M_t|\mathbf{Y}_t\right)-\bar{Y}_t\right)=0$$

with probability one.

Definition 2.5.2 *A predictor M_t is said to be design-consistent if for all $\varepsilon>0$,*

$$\lim_{t\to\infty}P_p\left(\left|M_t-\bar{Y}_t\right|>\varepsilon|\mathbf{Y}_t\right)=0$$

with probability one.

The following lemmas will enable us to show that M_t^o and consequently M_t to be asymptotically design-unbiased and design-consistent.

Lemma 2.5.1 $\lim_{t\to\infty}\left(E_p\left(\hat{\mu}_*(x_k)-\mu(x_k)|\mathbf{Y}_t\right)\right)=0$, *with probability one.*

Proof Now

$$\hat{\mu}_*(x_k)=\frac{\sum_{j\in s_t}\frac{W_j(x_k)y_j}{\pi_j}}{\sum_{j\in U_t}W_j(x_k)},$$

and

$$\hat{\mu}_t(x_k)=\frac{\sum_{j\in U_t}W_j(x_k)y_j}{\sum_{j\in U_t}W_j(x_k)}.$$

But

$$E_p\left(\sum_{j\in s_t}\frac{W_j(x_k)y_j}{\pi_j}-\sum_{j\in U_t}W_j(x)y_j|\mathbf{Y}_t\right)=0 \text{ for all } t,$$

hence

$$E_p\left(\frac{\sum_{j\in s_t}\frac{W_j(x_k)y_j}{\pi_j}}{\sum_{j\in U_t}W_j(x_k)}-\frac{\sum_{j\in U_t}W_j(x_k)y_j}{\sum_{j\in U_t}W_j(x_k)}|\mathbf{Y}_t\right)=0.$$

Therefore

$$\lim_{t\to\infty}\left(E_p\left(\hat{\mu}_*(x_k)-\hat{\mu}_t(x_k)|\mathbf{Y}_t\right)\right)=0 \text{ with probability one.}$$

Now

$$\begin{aligned}\lim_{t\to\infty}\left(E_p\left(\hat{\mu}_*(x_k)-\mu(x_k)|\mathbf{Y}_t\right)\right) &= \lim_{t\to\infty}\left(E_p\left(\hat{\mu}_*(x_k)-\hat{\mu}_t(x_k)|\mathbf{Y}_t\right)\right)\\ &\quad+\lim_{t\to\infty}\left(E_p\left(\hat{\mu}_t(x_k)-\mu(x_k)|\mathbf{Y}_t\right)\right).\end{aligned}$$

Using Theorem 2.2.2 we have that

$$\lim_{t\to\infty}\left(E_p\left(\hat{\mu}_t(x_k)-\mu(x_k)|\mathbf{Y}_t\right)\right)=\lim_{t\to\infty}\left(\hat{\mu}_t(x_k)-\mu(x_k)\right)\ =0,$$

almost surely. Using Slutzky's theorem (Roussas, 1997)we have that

$$\lim_{t\to\infty}\left(E_p\left(\hat{\mu}_*(x_k)-\mu(x_k)|\mathbf{Y}_t\right)\right)=0, \text{with probability one.}\ \blacksquare$$

Lemma 2.5.2 $\lim_{t\to\infty} E_p\left(\hat{m}_{sm}-\hat{m}^o_{sm}\right)|\mathbf{Y}_t)=0$ *with probability one.*

Proof Apply Lemma 2.5.1. ■

Now rewrite

$$\hat{m}^o_{sm}=\frac{1}{N}\sum\nolimits_{k\in s}\frac{y_k}{\pi_{kt}}+\frac{1}{N}\sum\nolimits_{k\in U_t}\hat{\mu}_t(x_k)-\frac{1}{N}\sum\nolimits_{k\in s_t}\frac{\hat{\mu}_t(x_k)}{\pi_{kt}},$$

as

$$\begin{aligned}M^o_t &= \frac{1}{N_t}\sum\nolimits_{k\in U_t}\frac{Y_kI_{kt}}{\pi_{kt}}-\frac{1}{N_t}\sum\nolimits_{k\in U_t}\hat{\mu}_t(x_k)\left(\frac{I_{kt}}{\pi_{kt}}-1\right)\\ &= \frac{1}{N_t}\sum\nolimits_{k\in U_t}\frac{Y_kI_{kt}}{\pi_{kt}}-\frac{1}{N_t}\sum\nolimits_{k\in U_t}\left(\sum\nolimits_{i\in s_t}w_i\left(x_k\right)Y_i\right)\left(\frac{I_{kt}}{\pi_{kt}}-1\right).\end{aligned}\tag{2.28}$$

Lemma 2.5.3 *If* $\mu(x_k)$ *is bounded for all* $x_k\in U_t$, *then* $\lim_{t\to\infty}\frac{1}{N_t}\sum_{k\in U_t}\left(\sum_{i\in s_t}w_i\left(x_k\right)Y_i\right)^2$ *is bounded with probability one.*

Proof Now $\hat{\mu}_t(x_k)=\sum_{i\in U_t}w_i\left(x_k\right)Y_i$ is the finite population estimator for population U_t such that $\hat{\mu}_t(x_k)\xrightarrow{d}\mu(x_k)$ with probability one when $t\to\infty$, $b\to 0$ and $n_tb\to\infty$.

Now if $Z_t \xrightarrow{p} \theta$ then $g(Z_t) \xrightarrow{p} g(\theta)$ if $g(.)$ is continuous. Therefore $\hat{\mu}_t^2(x_k) \xrightarrow{p} \mu^2(x_k)$ when $t \to \infty, b \to 0$ and $n_t b \to \infty$.

Since $\mu(x_k)$ is bounded for all $x_k \in U_t$, then $\lim_{t\to\infty} \frac{1}{N_t} \sum_{k\in U_t} \mu^2(x_k)$ is bounded for all $x_k \in U_t$ and $\lim_{t\to\infty} \frac{1}{N_t} \sum_{k\in U_t} \left(\sum_{i\in s_t} w_i(x_k) Y_i\right)^2$ is bounded with probability one. ■

The following inequality is needed to show the asymptotic results:

Lemma 2.5.4 *If A and B are random variables then*

$$E_p|A-B| \leq \left(E_p(A^2)\right)^{\frac{1}{2}} + \left(E_p(B^2)\right)^{\frac{1}{2}}. \tag{2.29}$$

Proof We have that (almost surely)

$$|A-B| \leq |A| + |B|.$$

Therefore applying the expectation operator on both sides of the above inequality gives us the following:

$$E_p|A-B| \leq E_p|A| + E_p|B|.$$

Using Jensen's inequality for concave functions

$$E_p(g(X)) \leq g(E_p(X)),$$

we obtain

$$E_p|A| \leq \left(E_p(A^2)\right)^{\frac{1}{2}}$$

and the result follows. ■

As in Issaki and Fuller (1981) and Robinson and Särndal (1983), the following assumptions are made because in classical sampling theory the Y_k's and π_k are fixed constants:

A.1 $\lim_{t\to\infty} \frac{1}{N_t} \sum_{k\in U_t} Y_{kt}^2 < \infty$ with probability one,

A.2 $\lim_{t\to\infty} N_t \min_{1\leq k\leq N_t} \pi_{kt} = \infty$,

A.3 $\lim_{t\to\infty} \max_{1\le k\neq l\le N_t} \left| \frac{\pi_{klt}}{\pi_{kt}\pi_{lt}} - 1 \right| = 0.$

The following theorem and proof is analogous to that of Robinson and Särndal (1983).

Theorem 2.5.5 *Under A.1 - A.3, M_t^o is asymptotically design unbiased and design-consistent.*

Proof The use of Markov's inequality

$$P_p\left(\left|M_t^o - \bar{Y}_t\right| > \varepsilon | \mathbf{Y}_t\right) \le \frac{E_p\left(\left|M_t^o - \bar{Y}_t\right| | \mathbf{Y}_t\right)}{\varepsilon^{\frac{1}{2}}},$$

will establish the design-consistency and the asymptotically design unbiasedness of M_t as per *Definitions* 2.5.1 and 2.5.2. It will be sufficient to show that

$$\lim_{t\to\infty} E_p\left(\left|M_t^o - \bar{Y}_t\right| | \mathbf{Y}_t\right) = 0, \text{ with probability one.}$$

Let

$$\begin{aligned} M_t^o - \bar{Y}_t &= a_t - b_t, \\ a_t &= \frac{1}{N_t}\sum\nolimits_{k\in U_t} Y_k \left(\frac{I_{kt}}{\pi_{kt}} - 1\right), \\ b_t &= \frac{1}{N_t}\sum\nolimits_{k\in U_t} \left(\sum\nolimits_{i\in s_t} w_i(x_k) Y_i\right)\left(\frac{I_{kt}}{\pi_{kt}} - 1\right). \end{aligned}$$

Therefore

$$E_p\left(\left|M_t^o - \bar{Y}_t\right| | \mathbf{Y}_t\right) = E_p\left(|a_t - b_t| | \mathbf{Y}_t\right)$$

and using Lemma 2.5.4 we have that

$$E_p\left(|a_t - b_t| | \mathbf{Y}_t\right) \le \left(E_p(a_t^2 | \mathbf{Y}_t)\right)^{\frac{1}{2}} + \left(E_p(b_t^2 | \mathbf{Y}_t)\right)^{\frac{1}{2}}.$$

Now

$$\begin{aligned} E_p(a_t^2 | \mathbf{Y}_t) &= \frac{1}{N_t^2}\sum\nolimits_{k\in U_t} Y_k^2 E_p\left(\frac{I_{kt}}{\pi_{kt}} - 1\right)^2 \\ &+ \frac{1}{N_t^2}\sum\nolimits_{k\in U_t} \sum_{\substack{k,l\in U_t \\ k\neq l}} Y_k Y_l E_p\left(\frac{I_{kt}}{\pi_{kt}} - 1\right)\left(\frac{I_{lt}}{\pi_{lt}} - 1\right). \end{aligned}$$

Also

$$\begin{aligned} E_p\left(\frac{I_{kt}}{\pi_{kt}}-1\right)^2 &= E_p\left(\frac{I_{kt}^2}{\pi_{kt}^2}-\frac{2I_{kt}}{\pi_{kt}}+1\right) \\ &= \left(\frac{\pi_{kt}}{\pi_{kt}^2}-\frac{2\pi_{kt}}{\pi_{kt}}+1\right) \\ &= \left(\frac{1}{\pi_{kt}}-1\right). \end{aligned}$$

But

$$\begin{aligned} E_p\left(\frac{I_{kt}}{\pi_{kt}}-1\right)\left(\frac{I_{lt}}{\pi_{lt}}-1\right) &= E_p\left(\frac{I_{kt}I_{lt}}{\pi_{kt}\pi_{lt}}-\frac{I_{kt}}{\pi_{kt}}-\frac{I_{lt}}{\pi_{lt}}+1\right) \\ &= \left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-\frac{\pi_{kt}}{\pi_{kt}}-\frac{\pi_{lt}}{\pi_{lt}}+1\right) \\ &= \left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right) \end{aligned}$$

Therefore,

$$\begin{aligned} E_p(a_t^2\,|\mathbf{Y}_t) &= \frac{1}{N_t^2}\sum\nolimits_{k\in U_t} Y_k^2\left(\frac{1}{\pi_{kt}}-1\right) \\ &\quad + \frac{1}{N_t^2}\sum\nolimits_{k\in U_t}\sum_{\substack{k,l\in U_t\\ k\neq l}} Y_kY_l\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right). \end{aligned}$$

Now

$$\begin{aligned} E_p(a_t^2\,|\mathbf{Y}_t) &\leq \frac{1}{N_t^2}\sum\nolimits_{k\in U_t} Y_k^2\left(\frac{1}{\pi_{kt}}-1\right) \\ &\quad + \left|\frac{1}{N_t^2}\sum\nolimits_{k\in U_t}\sum_{\substack{k,l\in U_t\\ k\neq l}} Y_kY_l\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right|. \end{aligned}$$

Since

$$\frac{1}{N_t^2}\sum\nolimits_{k\in U_t} Y_k^2\left(\frac{1}{\pi_{kt}}-1\right) \leq \left(\lim_{t\to\infty} N_t \min_{1\leq k\leq N_t}\pi_{kt}\right)^{-1}\frac{1}{N_t}\sum\nolimits_{k\in U_t} Y_k^2,$$

and because of A.1 and A.2

$$\left(\lim_{t\to\infty} N_t \min_{1\leq k\leq N_t}\pi_{kt}\right)^{-1}\frac{1}{N_t}\sum\nolimits_{k\in U_t} Y_k^2 \to 0 \text{ as } t\to\infty \text{ with probability one,}$$

therefore

$$\frac{1}{N_t^2}\sum\nolimits_{k\in U_t} Y_k^2\left(\frac{1}{\pi_{kt}}-1\right) \to 0 \text{ as } t\to\infty \text{ with probability one.}$$

Also

$$\left|\frac{1}{N_t^2}\sum_{k\in U_t}\sum_{\substack{k,l\in U_t\\k\neq l}} Y_kY_l\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right| \leq \max_{1\leq k\neq l\leq N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\left|\frac{1}{N_t^2}\sum_{k\in U_t}\sum_{\substack{k,l\in U_t\\k\neq l}} Y_kY_l\right|.$$

But

$$\max_{1\leq k\neq l\leq N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\left|\frac{1}{N_t^2}\sum_{k\in U_t}\sum_{\substack{k,l\in U_t\\k\neq l}} Y_kY_l\right| \leq \frac{1}{N_t}\max_{1\leq k\neq l\leq N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\frac{1}{N_t}\sum_{k\in U_t} Y_k^2.$$

Therefore because of A.1 and then A.3,

$$\left|\frac{1}{N_t^2}\sum_{k\in U_t}\sum_{\substack{k,l\in U_t\\k\neq l}} Y_kY_l\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right| \leq \frac{1}{N_t}\max_{1\leq k\neq l\leq N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\frac{1}{N_t}\sum_{k\in U_t} Y_k^2 \to 0,$$

when $t \to \infty$ with probability one. Therefore

$$E_p(a_t^2\,|\mathbf{Y}_t) \to 0$$

as $t \to \infty$ with probability one.

Now

$$\begin{aligned} E_p(b_t^2\,|\mathbf{Y}_t) &= \frac{1}{N_t^2}\sum_{k\in U_t}\left(\sum_{i\in s_t} w_i(x_k)\,Y_i\right)^2\left(\frac{1}{\pi_{kt}}-1\right)+ \\ &\quad \frac{1}{N_t^2}\sum_{k\in U_t}\sum_{\substack{k,l\in U_t\\k\neq l}}\left(\sum_{i\in s_t} w_i(x_k)\,Y_i\right)\left(\sum_{i\in s_t}\frac{w_{li}Y_i}{\pi_{lt}}\right)\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right) \end{aligned}$$

but

$$E_p(b_t^2 \,|\mathbf{Y}_t) \;\leq\; \frac{1}{N_t^2}\sum\nolimits_{k\in U_t}\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)^2\left(\frac{1}{\pi_{kt}}-1\right)+$$
$$\left|\frac{1}{N_t^2}\sum\nolimits_{k\in U_t}\sum_{\substack{k,l\in U_t\\ k\neq l}}\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right|.$$

We now study the first term of the above expression,

$$\frac{1}{N_t^2}\sum\nolimits_{k\in U_t}\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)^2\left(\frac{1}{\pi_{kt}}-1\right)\leq\left(\lim_{t\to\infty} N_t \min_{1\leq k\leq N_t}\pi_{kt}\right)^{-1}\frac{1}{N_t}\sum\nolimits_{k\in U_t}\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)$$

but using A.2 and Lemma 2.5.3, we obtain:

$$\left(\lim_{t\to\infty} N_t \min_{1\leq k\leq N_t}\pi_{kt}\right)^{-1}\frac{1}{N_t}\sum\nolimits_{k\in U_t}\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)^2 \xrightarrow{p} 0 \text{ as } t\to\infty \text{ with probability one.}$$

Therefore as $t\to\infty$

$$\frac{1}{N_t^2}\sum\nolimits_{k\in U_t}\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)^2\left(\frac{1}{\pi_{kt}}-1\right)\to 0,$$

with probability one. For the second term we have that,

$$\left|\frac{1}{N_t^2}\sum\nolimits_{k\in U_t t}\sum_{\substack{k,l\in U_t\\ k\neq l}}\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right|$$

$$\leq \max_{1\leq k\neq l\leq N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\frac{1}{N_t^2}\left|\sum\nolimits_{k\in U_t}\sum_{\substack{k,l\in U_t\\ k\neq l}}\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)\left(\sum\nolimits_{i\in s_t} w_i(x_k)\,Y_i\right)\right|,$$

but because of A.3 and Lemma 2.5.4,

$$\max_{1\le k\ne l\le N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\frac{1}{N_t^2}\left|\sum\nolimits_{k\in U_t}\sum_{\substack{k,l\in U_t\\ k\ne l}}\left(\sum\nolimits_{i\in s_t}w_i(x_k)Y_i\right)\left(\sum\nolimits_{i\in s_t}w_i(x_k)Y_i\right)\right|$$

$$\le\frac{1}{N_t}\max_{1\le k\ne l\le N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\frac{1}{N_t}\left(\sum\nolimits_{k\in U_t}\left(\sum\nolimits_{i\in s_t}w_i(x_k)Y_i\right)\right)^2\to 0,$$

as $t\to\infty$ with probability one. Therefore,

$$\left|\frac{1}{N_t^2}\sum\nolimits_{k\in U_t t}\sum_{\substack{k,l\in U_t\\ k\ne l}}\left(\sum\nolimits_{i\in s_t}w_i(x_k)Y_i\right)\left(\sum\nolimits_{i\in s_t}w_i(x_k)Y_i\right)\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right|\to 0,$$

as $t\to\infty$ with probability one. Hence,

$$E_p(b_t^2\,|\mathbf{Y}_t)\to 0 \text{ as } t\to\infty \text{ with probability one.}$$

Therefore the above facts demonstrate that M_t^o is asymptotically design unbiased and design-consistent. ■

Corollary 2.5.6 *$M_t=\hat{m}_{sm}$ is asymptotically design unbiased and design-consistent.*

Remark 2.5.1 *Robinson and Särndal (1983), proved this theorem for the case of the generalized linear regression model. Our theorem generalizes their result, because we make no linearity assumption between the response and explanatory variable.*

2.6 Anticipated Mean Square Error

A general expression is derived for the expected model mean square error $\mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t)$. This measure was appropriately named by Fuller and Issaki (1982) as the anticipated mean square error of M_t. Furthermore, it is shown that as $t\to\infty$, the anticipated mean square error is 0.

The following assumptions will be used in the sequel:

A.4 $\lim_{t\to\infty} \frac{1}{N_t} \sum_{k\in U_t} \sigma_{kt}^2 = 0,$

A.5 $\lim_{t\to\infty} \frac{N_t}{n_t} \min_{1\le k\le N_t} \pi_{kt} \to \infty.$

Theorem 2.6.1 *The anticipated mean square error of M_t is given by*

$$\mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t) = A_t + \mathcal{E}_\xi(g_t^2) + 2\mathcal{E}_\xi E_p(h_t g_t), \tag{2.30}$$

where

$$\begin{aligned} A_t &= \frac{n_t}{N_t^2} \sum_{k\in U_t} \sigma_k^2 \left(\frac{1}{\pi_{kt}} - 1\right) \\ \mathcal{E}_\xi(g_t^2) &= \frac{n_t}{N_t^2} \sum_{k\in U_t} \mathcal{E}_\xi \left(\varepsilon_k^2 \left(\frac{1}{\pi_{kt}} - 1\right)\right) + \\ &\quad + \frac{n_t}{N_t^2} \sum_{k\in U_t} \sum_{\substack{l\in U_t \\ k\neq l}} \mathcal{E}_\xi \left(\varepsilon_k \varepsilon_l \left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}} - 1\right)\right), \end{aligned}$$

and

$$\varepsilon_k = \mu(x_k) - \hat{\mu}_t(x_k).$$

Proof Let

$$\begin{aligned} \sqrt{n_t}\left(M_t - \bar{Y}_t\right) &= h_t + g_t \\ h_t &= \frac{\sqrt{n_t}}{N_t} \sum_{k\in U_t} (Y_k - \mu(x_k)) \left(\frac{I_{kt}}{\pi_{kt}} - 1\right) \\ g_t &= \frac{\sqrt{n_t}}{N_t} \sum_{k\in U_t} (\mu(x_k) - \hat{\mu}_t(x_k)) \left(\frac{I_{kt}}{\pi_{kt}} - 1\right). \end{aligned}$$

Now

$$\begin{aligned} n_t \mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t) &= n_t \mathcal{E}_\xi E_p \left(M_t - \bar{Y}_t\right)^2 \\ &= \mathcal{E}_\xi E_p\left(h_t^2\right) + \mathcal{E}_\xi E_p\left(g_t^2\right) + 2\mathcal{E}_\xi E_p(h_t g_t). \end{aligned}$$

But

$$\begin{aligned} \mathcal{E}_\xi E_p\left(h_t^2\right) &= \mathcal{E}_\xi E_p \left(\frac{\sqrt{n_t}}{N_t} \sum_{k\in U_t} (Y_k - \mu(x_k)) \left(\frac{I_{kt}}{\pi_{kt}} - 1\right)\right)^2 \\ &= \mathcal{E}_\xi \left(\frac{n_t}{N_t^2} \sum_{k\in U_t} (Y_k - \mu(x_k))^2\right) \left(\frac{1}{\pi_{kt}} - 1\right) \\ &\quad + \mathcal{E}_\xi \left(\frac{n_t}{N_t^2} \sum_{k\in U_t} \sum_{\substack{l\in U_t \\ k\neq l}} (Y_k - \mu(x_k))(Y_l - \mu(x_l)) \left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}} - 1\right)\right), \end{aligned}$$

since we assumed that the Y_k are uncorrelated the above expression reduces to

$$\mathcal{E}_\xi E_p\left(h_t^2\right) = \frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\sigma_k^2\left(\frac{1}{\pi_{kt}}-1\right).$$

Now let us write

$$\varepsilon_k = \mu(x_k) - \hat{\mu}_t(x_k),$$

$$\begin{aligned}
\mathcal{E}_\xi E_p\left(g_t^2\right) &= \mathcal{E}_\xi E_p\left(\frac{\sqrt{n_t}}{N_t}\sum\nolimits_{k\in U_t}\left(\mu(x_k)-\hat{\mu}_t(x_k)\right)\left(\frac{I_{kt}}{\pi_{kt}}-1\right)\right)^2 \\
&= \mathcal{E}_\xi E_p\left(\frac{\sqrt{n_t}}{N_t}\sum\nolimits_{k\in U_t}\varepsilon_k\left(\frac{I_{kt}}{\pi_{kt}}-1\right)\right)^2 \\
&= \mathcal{E}_\xi\left(\frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\varepsilon_k^2\left(\frac{1}{\pi_{kt}}-1\right)\right)+ \\
&\quad \mathcal{E}_\xi\left(\frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\sum_{\substack{k,l\in U_t\\ k\neq l}}\varepsilon_k\varepsilon_l\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right),
\end{aligned}$$

the result follows by making the appropriate substitutions. ■

Assume that the bias of $\hat{\mu}_t(x_k)$ is of negligible size compared with the variance i.e. $\varepsilon_k = \mu(x_k) - \hat{\mu}_t(x_k) \approx 0$. Therefore the terms $\mathcal{E}_\xi(g_t^2)$ and $2\mathcal{E}_\xi E_p\left(h_t g_t\right)$ vanish in the above theorem and

$$\mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t) = A_t.$$

Example 2.6.1 *If the sampling plan is simple random sampling* $\pi_{kt} = \frac{n_t}{N_t}$, *then the anticipated mean squared error is*

$$\mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t) = (1-f_t)\sum\nolimits_{k\in U_t}\sigma_k^2.$$

It is now shown that the anticipated asymptotic mean square error approaches 0.

Theorem 2.6.2 $\mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t) \to 0$, *as* $n_t \to \infty, b \to 0$, $n_t b \to \infty$, *and* $t \to \infty$.

Proof Using A.4 and A.5 we obtain that

$$\frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\sigma_k^2\left(\frac{1}{\pi_{kt}}-1\right) \leq \left(\lim_{t\to\infty}\frac{N_t}{n_t}\min_{1\leq k\leq N_t}\pi_{kt}\right)^{-1}\frac{1}{N_t}\sum\nolimits_{k\in U_t}\sigma_k^2 \to 0,$$

as $t \to \infty$ or

$$\mathcal{E}_\xi E_p\left(h_t^2\right) \to 0 \text{ as } b \to 0, n_t b \to \infty \text{ and } t \to \infty. \tag{2.31}$$

We previously found that

$$\begin{aligned}\mathcal{E}_\xi(g_t^2) &= \frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\mathcal{E}_\xi\left(\varepsilon_k^2\left(\frac{1}{\pi_{kt}}-1\right)\right) \\ &+ \frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\sum_{\substack{k,l\in U_t\\ k\neq l}}\mathcal{E}_\xi\left(\varepsilon_k\varepsilon_l\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right).\end{aligned}$$

Now using A.5,

$$\frac{n_t}{N_t^2}\mathcal{E}_\xi\left(\sum\nolimits_{k\in U_t}\varepsilon_k^2\left(\frac{1}{\pi_{kt}}-1\right)\right) \leq \left(\left(\lim_{t\to\infty}\frac{N_t}{n_t}\min_{1\leq k\leq N_t}\pi_{kt}\right)^{-1}\frac{1}{N_t}\sum\nolimits_{k\in U_t}\mathcal{E}_\xi\left(\varepsilon_k^2\right)\right) \to 0,$$

as $b \to 0, n_t b \to \infty$ and $t \to \infty$. Therefore

$$\mathcal{E}_\xi\left(\frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\varepsilon_k^2\left(\frac{1}{\pi_{kt}}-1\right)\right) \to 0 \text{ as } b \to 0, n_t b \to \infty \text{ and } t \to \infty.$$

Also

$$\mathcal{E}_\xi\left(\frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\sum\nolimits_{\substack{l\in U_t\\ k\neq l}}\varepsilon_k\varepsilon_l\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right) \leq$$

$$\frac{n_t}{N_t^2}\max_{1\leq k\neq l\leq N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\sum\nolimits_{k\in U_t}\sum\nolimits_{\substack{l\in U_t\\ k\neq l}}\mathcal{E}_\xi\left(\varepsilon_k\varepsilon_l\right)$$

and using A.3 we have that

$$\frac{n_t}{N_t}\max_{1\leq k\neq l\leq N_t}\left|\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right|\frac{1}{N_t}\sum\nolimits_{k\in U_t}\sum\nolimits_{\substack{l\in U_t\\ k\neq l}}\mathcal{E}_\xi\left(\varepsilon_k\varepsilon_l\right) \to 0$$

as $b \to 0, n_t b \to \infty$ and $t \to \infty$. Therefore

$$\mathcal{E}_\xi\left(\frac{n_t}{N_t^2}\sum\nolimits_{k\in U_t}\sum\nolimits_{\substack{l\in U_t\\ k\neq l}}\varepsilon_k\varepsilon_l\left(\frac{\pi_{klt}}{\pi_{kt}\pi_{lt}}-1\right)\right) \to 0$$

as $b \to 0, n_t b \to \infty$ and $t \to \infty$. Hence

$$\mathcal{E}_\xi E_p\left(g_t^2\right) \to 0 \text{ as } b \to 0, n_t b \to \infty \text{ and } t \to \infty. \tag{2.32}$$

Now

$$\mathcal{E}_\xi E_p(h_t g_t) \leq \left(\mathcal{E}_\xi E_p\left(h_t^2\right) \mathcal{E}_\xi E_p\left(g_t^2\right)\right)^{\frac{1}{2}}.$$

Therefore

$$\mathcal{E}_\xi E_p(h_t g_t) \to 0 \text{ as } b \to 0, n_t b \to \infty \text{ and } t \to \infty,$$

using (2.31) and (2.32) and the theorem follows. ■

2.7 Near Optimal Inclusion Probabilities

We now find the values of π_{kt} that minimize the asymptotic anticipated mean squared error, $\mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t)$. Moreover these values of π_{kt} induce $\mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t)$ to attain the Godambe and Joshi (1965) lower bound.

Theorem 2.7.1 *Let $p(.)$ be any probability sampling design such that the expected sample size satisfies*

$$E_p(n_{s_t}) = n_t,$$

for some given n_t. For the general smoothing estimator the values of π_{kt} that minimize the asymptotic $\mathcal{E}_\xi MSE_p(M_t|\mathbf{Y}_t) = A_t$ when $\varepsilon_k \approx 0$ are given by

$$\pi_{kt} = \frac{n_t \sigma_k}{\sum_{k \in U_t} \sigma_k}. \tag{2.33}$$

The value of the minimized anticipated MSE is

$$\mathcal{E}_\xi MSE_{p_0}(M_t|\mathbf{Y}_t) = \frac{1}{N_t^2}\left(\sum_{k \in U_t} \sigma_k\right)^2 - \frac{n_t}{N_t^2} \sum_{k \in U_t} \sigma_k^2. \tag{2.34}$$

Proof The anticipated mean squared error under any probability design is

$$V = \frac{n_t}{N_t^2} \sum_{k \in U_t} \frac{\sigma_k^2}{\pi_{kt}} - \frac{n_t}{N_t^2} \sum_{k \in U_t} \sigma_k^2,$$

if $\varepsilon_k \approx 0$. Also let

$$\begin{aligned} V_1 &= V + \frac{n_t}{N_t^2} \sum_{k \in U_t} \sigma_k^2 \\ &= \frac{n_t}{N_t^2} \sum_{k \in U_t} \frac{\sigma_k^2}{\pi_{kt}}, \end{aligned}$$

which is dependent on π_{kt}. The constraint $E_p(n_{s_t}) = n_t$ maybe rewritten as

$$C = \sum_{k \in U_t} \pi_{kt} = n_t.$$

Now

$$\frac{N_t^2}{n_t} V_1 C = \sum_{k \in U_t} \frac{\sigma_k^2}{\pi_{kt}} \sum_{k \in U_t} \pi_{kt}.$$

We now use the Cauchy-Schwartz inequality to find the optimal value of π_{kt}. The Cauchy-Schwartz inequality states that

$$\left(\sum_{\substack{k,l \in U_t \\ k \neq l}} a_k^2 \right) \left(\sum_{\substack{k,l \in U_t \\ k \neq l}} b_k^2 \right) \geqslant \left(\sum_{\substack{k,l \in U_t \\ k \neq l}} a_k b_k \right)^2,$$

and equality holds if and only if $\frac{b_k}{a_k}$ is a constant for every k.Take

$$a_k = \frac{\sigma_k}{\sqrt{\pi_{kt}}}$$

and

$$b_k = \sqrt{\pi_{kt}}.$$

We have equality when

$$\left(\frac{\frac{\sqrt{\pi_{kt}}}{\sigma_k}}{\sqrt{\pi_{kt}}}\right)^{\frac{1}{2}} = \left(\frac{\pi_{kt}}{\sigma_k}\right)^{\frac{1}{2}} = \text{constant.}$$

Let

$$\left(\frac{\pi_{kt}}{\sigma_k}\right)^{\frac{1}{2}} = \kappa^{\frac{1}{2}},$$

then

$$\pi_{kt} = \kappa\sigma_k,$$

but $\sum_{k\in U_t} \pi_{kt} = n_t$ this implies that

$$\kappa = \frac{n_t}{\sum_{k\in U_t} \sigma_k},$$

therefore

$$\pi_{kt} = \frac{n_t\sigma_k}{\sum_{k\in U_t} \sigma_k}.$$

Placing this value of π_{kt} in the asymptotic anticipated mean squared error the desired result is found. ■

We now state the Godambe and Joshi (1965) inequality:

Theorem 2.7.2 *Godambe and Joshi (1965) Let $\hat{t}$ be any estimator of t satisfying*

$$\mathcal{E}_\xi E_p\left(\hat{t} - t\right) = 0,$$

then

$$\mathcal{E}_\xi E_p\left(\hat{t} - t\right)^2 \geqslant \sum_{k\in U}\left(\frac{1}{\pi_k} - 1\right)\sigma_k^2. \tag{2.35}$$

Applying this theorem with the optimal π_{kt}:

$$\frac{n_t}{N_t^2}\sum_{k\in U_t}\left(\frac{1}{\pi_k}-1\right)\sigma_k^2=\frac{1}{N_t^2}\left(\sum_{k\in U_t}\sigma_k\right)^2-\frac{n_t}{N_t^2}\sum_{k\in U_t}\sigma_k^2,$$

which is the same as the minimum anticipated asymptotic mean squared error of the generalized smoothing estimator as given in (2.34) when the bias is negligible. Thus we conclude that the generalized smoothing estimator with optimum inclusion probability attains asymptotically the minimum anticipated mean squared error of any design unbiased estimator.

2.8 The Bias and Variance of the Generalized Smoothing Estimator

Consider a complete enumeration of the population where we observe (x_k, y_k) for all $k \in U$. In this setup one can find the population kernel predictor of y_k i.e.

$$\begin{aligned} y_k^o &= \hat{\mu}(x_k) \\ &= \sum_{j\in U} w_j(x_k)\, y_j, \end{aligned}$$

where

$$w_j(x_k)=\frac{K\left(\frac{x_k-x_j}{b}\right)}{\sum\limits_{j\in U}K\left(\frac{x_k-x_j}{b}\right)},$$

and $K(.)$ is any of the kernel functions described in Section 2.2. Also the population fitted residuals are now defined as

$$E_k = y_k - \hat{\mu}(x_k),\ k \in U.$$

In order to find the bias and variance of $\hat{m}_{sm}$, the following theorem will be used.

Theorem 2.8.1 *(Särndal, Swensson and Wretman 1992)The regression estimator* $\hat{t}_{yr}$

$$\hat{t}_{yr} = \sum_{k\in U} \hat{y}_k + \sum_{k\in s} \frac{e_k}{\pi_k}$$

is approximately design unbiased for $t = \sum_{k\in U} y_k$ *with variance*

$$V_p\left(\hat{t}_{yr}\right) = \sum_{\substack{k,l\in U_t \\ k\neq l}} \sum_{k\in U} \left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}. \tag{2.36}$$

Provided $\pi_{kl} > 0$ *for all* $k, l \in U$, *an unbiased estimator of* $V_p\left(\hat{t}_{\pi}\right)$ *is given by*

$$\hat{V}_p\left(\hat{t}_{yr}\right) = \sum_{\substack{k,l\in U_t \\ k\neq l}} \sum_{k\in U} \frac{1}{\pi_{kl}} \left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}. \tag{2.37}$$

The main result of this section is contained in the following theorem.

Theorem 2.8.2 *The finite general smoothing estimator* $\hat{m}_{sm}$ *is approximately design unbiased for* $\bar{Y}$ *and has an approximate design variance*

$$V_p\left(\hat{m}_{sm}\right) \doteq \frac{1}{N^2} \sum_{\substack{k,l\in U_t \\ k\neq l}} \sum_{k\in U} \left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}, \tag{2.38}$$

which can be estimated by

$$\hat{V}_p\left(\hat{m}_{sm}\right) \doteq \frac{1}{N^2} \sum_{\substack{k,l\in U_t \\ k\neq l}} \sum_{k\in s} \frac{1}{\pi_{kl}} \left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}, \tag{2.39}$$

where $e_k = y_k - \hat{y}_k$, e_k *being the sample counterpart of* E_k.

Proof We first express

$$\hat{m}_{sm} = \frac{1}{N} \sum_{k\in U} \hat{y}_k + \frac{1}{N} \sum_{k\in s} \frac{e_k}{\pi_k},$$

as

$$\begin{aligned}\hat{m}_{sm} &= \frac{1}{N}\sum_{k\in U}\hat{\mu}_{\pi}(x_k) + \frac{1}{N}\sum_{k\in s}\frac{y_k - \hat{\mu}_{\pi}(x_k)}{\pi_k} \\ &= \frac{1}{N}\sum_{k\in U}\hat{\mu}(x_k) + \frac{1}{N}\sum_{k\in s}\frac{y_k - \hat{\mu}(x_k)}{\pi_k} \\ &\quad + \frac{1}{N}\sum_{k\in U}\left(\hat{\mu}_{\pi}(x_k) - \hat{\mu}(x_k)\right) - \frac{1}{N}\sum_{k\in s}\frac{\hat{\mu}_{\pi}(x_k) - \hat{\mu}(x_k)}{\pi_k}.\end{aligned}$$

Hence

$$\begin{aligned}E_p(\hat{m}_{sm}) &= E_p\left(\frac{1}{N}\sum_{k\in U}\hat{\mu}(x_k) + \frac{1}{N}\sum_{k\in s}\frac{E_k}{\pi_k}\right) \\ &\quad + E_p\left(\frac{1}{N}\sum_{k\in U}\left(\hat{\mu}_{\pi}(x_k) - \hat{\mu}(x_k)\right) - \frac{1}{N}\sum_{k\in s}\frac{\hat{\mu}_{\pi}(x_k) - \hat{\mu}(x_k)}{\pi_k}\right).\end{aligned}$$

Using Lemmas 2.5.1 and 2.5.2 we have that

$$\begin{aligned}E_p(\hat{m}_{sm}) &\doteq \frac{1}{N}\sum_{k\in U}\mu_o(x_k) + \frac{1}{N}\sum_{k\in U}E_k, \\ &\doteq \bar{Y},\end{aligned}$$

which demonstrates that $\hat{m}_{sm}$ is approximately design unbiased and the other results follow using Theorem 2.8.1 ■

Example 2.8.1 *Suppose the sampling design is simple random sampling then* $\pi_k = \frac{n}{N}$ *and* $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$. *Now with this sampling design the G.S.E.* $\hat{m}_{sm}$ *is easily shown to be*

$$\hat{m}_{sm} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{n}\sum_{k\in s}e_k,$$

with approximate design variance

$$V_p(\hat{m}_{sm}) = \frac{1-f}{n}S_E^2,$$

where

$$S_E^2 = \frac{1}{N-1}\sum_{k\in U}(y_k - y_k^{\circ})^2.$$

Now we estimate $V_p(\hat{m}_{sm})$ *by*

$$\hat{V}_p(\hat{m}_{sm}) = \frac{1-f}{n} s_e^2.$$

The above theorem shows that $\hat{m}_{sm}$ is approximately unbiased. Consequently we can use a normal approximation to find confidence intervals for $\bar{Y}$ using $\hat{m}_{sm}$ and $\hat{V}_p(\hat{m}_{sm})$ because $\hat{m}_{sm}$ is asymptotically unbiasedness and consistent (Hájek, 1960).

2.9 The Asymptotic Distribution of the Generalized Smoothing Estimator under a Simple Random Sample Design

In this section we specify conditions under which the finite population central limit theorem holds for the generalized smoothing estimator under simple random sampling. Then the population variance is replaced by an estimator, which is also consistent. Therefore the central limit property holds for the standardized form with the variance replaced by its estimator. The population finite population central limit theorem is only valid for the simple random sample design.

Our results follow from the central limit theorem for finite populations, Hájek (1960):

Theorem 2.9.1 *Suppose* $N_t \to \infty$ *and* $n_t \to \infty$ *as* $t \to \infty$. *Then under simple random sampling,*

$$\frac{\sqrt{n_t}\,(\bar{y}_t - \bar{Y}_t)}{\sqrt{1-f_t}\,S_t} \xrightarrow{\mathcal{L}} N(0,1) \text{ as } t \to \infty,$$

if and only if $\{Y_{tj}\}$ *satisfy the Lindeberg-Hájek condition*

$$\lim_{t\to\infty} \sum\nolimits_{k\in R(\delta)} \frac{(Y_{kt} - \bar{Y}_t)^2}{(N_t - 1)\,S_t^2} = 0 \text{ for any } \delta > 0,$$

where $R(\delta)$ is the set of units in U_t, for which

$$\frac{|Y_{kt} - \bar{Y}_t|}{\sqrt{1 - f_t} S_t} > \delta \sqrt{n_t}.$$

Under simple random sampling $\pi_{kt} = \frac{n_t}{N_t}$ and our general smoothing estimator has the following form:

$$M_t = \bar{y}_t + \frac{1}{N_t} \sum\nolimits_{k \in U_t} \hat{\mu}(x_{kt}) - \frac{1}{n_t} \sum\nolimits_{k \in S_t} \hat{\mu}(x_{kt}).$$

We now define the population residual about the regression curve as follows

$$E_{kt} = (Y_{kt} - \bar{Y}_t) - \left(\mu(x_{kt}) - \frac{1}{N_t} \sum\nolimits_{k \in U_t} \mu(x_{kt}) \right), \tag{2.40}$$

it can easily be shown that

$$\sum_{k=1}^{N_t} E_{kt} = 0.$$

The population variance of E_{kt} is

$$S_{E_t}^2 = \frac{\sum_{k=1}^{N_t} E_{kt}^2}{N_t - 1}.$$

The sample mean of the E_{kt} is defined as

$$\bar{e}_t = (\bar{y}_t - \bar{Y}_t) - \left(\frac{1}{n_t} \sum\nolimits_{k \in S_t} \mu(x_{kt}) - \frac{1}{N_t} \sum\nolimits_{k \in U_t} \mu(x_{kt}) \right), \tag{2.41}$$

therefore

$$\begin{aligned} M_t - \bar{Y}_t &= \bar{e}_t + \frac{1}{N_t} \sum\nolimits_{k \in U_t} (\hat{\mu}(x_{kt}) - \mu(x_{kt})) - \frac{1}{n_t} \sum\nolimits_{k \in S_t} (\hat{\mu}(x_{kt}) - \mu(x_{kt})) \\ &= \bar{e}_t + \bar{D}_t - \bar{d}_t, \end{aligned} \tag{2.42}$$

where

$$\begin{aligned}\bar{D}_t &= \frac{1}{N_t}\sum_{k\in U_t}(\hat{\mu}(x_{kt})-\mu(x_{kt}))\\ \bar{d}_t &= \frac{1}{n_t}\sum_{k\in S_t}(\hat{\mu}(x_{kt})-\mu(x_{kt}))\,.\end{aligned}$$

Theorem 2.9.2 *Under simple random sampling*

$$\frac{\sqrt{n_t}\,(M_t-\bar{Y}_t)}{\sqrt{1-f_t}S_{E_t}} \xrightarrow{\mathcal{L}} N(0,1) \text{ as } t\to\infty, b\to 0 \text{ and } n_t b\to\infty,$$

provided $\{E_{kt}\}$ *satisfy the Lindeberg-Hájek condition.*

Proof We express

$$\frac{\sqrt{n_t}\,(M_t-\bar{Y}_t)}{\sqrt{1-f_t}S_{E_t}} = \frac{\sqrt{n_t}\bar{e}_t}{\sqrt{1-f_t}S_{E_t}} + \frac{\sqrt{n_t}\bar{D}_t}{\sqrt{1-f_t}S_{E_t}} - \frac{\sqrt{n_t}\bar{d}_t}{\sqrt{1-f_t}S_{E_t}}.$$

Using Theorem 2.5.5 we have that

$$\frac{\sqrt{n_t}\bar{D}_t}{\sqrt{1-f_t}S_{E_t}} \xrightarrow{p} 0, \text{ as } b\to 0, n_t b\to\infty, \text{ as } t\to\infty.$$

Similarly

$$\frac{\sqrt{n_t}\bar{d}_t}{\sqrt{1-f_t}S_{E_t}} \xrightarrow{p} 0, \text{ as } b\to 0, n_t b\to\infty, \text{ as } t\to\infty.$$

Now we verify the Lindeberg-Hájek condition

$$\begin{aligned}\lim_{t\to\infty}\sum_{R(\delta)}\frac{E_{kt}^2}{(N_t-1)\,S_{E_t}^2} &= \lim_{t\to\infty}\left(\sum_{k=1}^{N_t}E_{kt}^2\right)^{-1}\sum_{k\in R(\delta)}E_{kt}^2\\ &= 0.\end{aligned}$$

Therefore

$$\frac{\sqrt{n_t}\bar{e}_t}{\sqrt{1-f_t}S_{E_t}} \xrightarrow{\mathcal{L}} N(0,1) \text{ as } t\to\infty.$$

Now using Slutsky's theorem (Roussas, 1997) we have that

$$\frac{\sqrt{n_t}\bar{e}_t}{\sqrt{1-f_t}S_{E_t}} + \frac{\sqrt{n_t}\bar{D}_t}{\sqrt{1-f_t}S_{E_t}} - \frac{\sqrt{n_t}\bar{d}_t}{\sqrt{1-f_t}S_{E_t}} \xrightarrow{\mathcal{L}} N(0,1) \text{ as } t\to\infty. \blacksquare$$

We now state a law of large numbers for finite populations due to Hájek (1960):

Theorem 2.9.3 *Suppose* $\{Y_{tj}\}$ *satisfies the condition*

$$\frac{(1-f_t)\,S_t^2}{n_t} \xrightarrow{p} 0 \text{ as } t \to \infty \ .$$

Then under simple random sampling,

$$(i)\ E\left|\bar{y}_t - \bar{Y}_t\right| \xrightarrow{p} 0 \text{ as } t \to \infty,$$

and

$$(ii)\ \bar{y}_t - \bar{Y}_t \xrightarrow{p} 0 \text{ as } t \to \infty.$$

The variance for M_t, $S_{E_t}^2$ is not usually known but must be estimated. We consider as its estimate

$$s_{E_t}^2 = \frac{\sum_{k \in S_t} (y_{kt} - \hat{\mu}(x_{kt}))}{n_t - 1}.$$

Our aim is to show that $\frac{s_{E_t}^2}{S_{E_t}^2} \xrightarrow{p} 1$ in probability as $t \to \infty$. We then can estimate $\frac{\sqrt{n_t}\,(M_t - \bar{Y}_t)}{\sqrt{1-f}\,S_{E_t}}$ by $\frac{\sqrt{n_t}\,(M_t - \bar{Y}_t)}{\sqrt{1-f}\,s_{E_t}}$ which will converge in distribution to a $N(0,1)$ as $t \to \infty$. We now adapt Theorem 2 of Scott and Wu (1981) for our problem.

Theorem 2.9.4 *Under simple random sampling* $\frac{s_{E_t}^2}{S_{E_t}^2} \xrightarrow{p} 1$ *as* $t \to \infty$ *provided the random variables* $\left\{\frac{(Y_{kt} - \mu(x_{kt}))^2}{S_{E_t}^2}\right\}$ *satisfy the conditions of Theorem 2.9.3.*

Proof We now express

$$\frac{s_{E_t}^2}{S_{E_t}^2} = A_1 + A_2 + A_3,$$

where

$$A_1 = \frac{\sum_{k\in S_t}(Y_{kt}-\mu(x_{kt}))^2}{(n_t-1)\,S_{E_t}^2}$$

$$A_2 = \frac{\sum_{k\in S_t}(\mu(x_{kt})-\hat{\mu}(x_{kt}))^2}{(n_t-1)\,S_{E_t}^2}$$

$$A_3 = -2\frac{\sum_{k\in S_t}(Y_{kt}-\mu(x_{kt}))\,(\mu(x_{kt})-\hat{\mu}(x_{kt}))}{(n_t-1)\,S_{E_t}^2},$$

it follows from our assumption and Theorem 2.9.3 that $A_1 \xrightarrow{p} 1$ as $t \to \infty$. Now using Theorem 2.9.2, $A_2 \xrightarrow{p} 0$ as $b \to 0, n_t \to \infty, n_t b \to \infty$ and $t \to \infty$. Lastly using Theorem 2.9.2, $A_3 \xrightarrow{p} 0$ as $b \to 0, n_t \to \infty, n_t b \to \infty$ and $t \to \infty$. ■

Corollary 2.9.5 *Under the conditions of Theorems 2.9.2 and 2.9.3,*

$$\frac{\sqrt{n_t}\,(M_t-\bar{Y}_t)}{\sqrt{1-f_t}\,s_{E_t}} \xrightarrow{\mathcal{L}} N(0,1) \text{ as } t\to\infty, b\to 0 \text{ and } n_t b\to\infty.$$

Example 2.9.1 *Therefore under simple random sampling we can use*

$$\hat{m}_{sm} \pm z_\alpha \sqrt{\hat{V}_p\,(\hat{m}_{sm})},$$

as a $100(1-\alpha)\%$ *confidence interval for* $\bar{Y}$.

2.10 Lack of Fit Test

The purpose of this section is to assess the lack of fit of a simple regression model. Azzalini, Bowman and Härdle (1989) proposed a pseudo-likelihood ratio test. The formal structure of the problem is as follows:

$$H_o : E(e) = 0$$

$$H_1 : E(e) = \text{smooth function of } x.$$

The hypothesis can be tested with

$$F_{NP} = \frac{RSS_0 - RSS_1}{RSS_0}, \tag{2.43}$$

where RSS_0 represents the residual sum of squares under the simple linear regression model

$$RSS_0 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{2.44}$$

and RSS_1 is the residual sum of squares of the fitted values using a nonparametric regression

$$RSS_1 = \sum_{i=1}^{n} (y_i - \hat{\mu}(x_i))^2. \tag{2.45}$$

.

The quadratic forms are respectively

$$\begin{aligned} RSS_0 &= \underline{e}^{\mathbf{T}}\underline{e} \\ &= \underline{e}^{\mathbf{T}}\mathbf{I}\underline{e}, \end{aligned}$$

and

$$\begin{aligned} RSS_1 &= \underline{e}^{\mathbf{T}}(\mathbf{I}-\mathbf{W})^{\mathbf{T}}(\mathbf{I}-\mathbf{W})\underline{e} \\ &= \underline{e}^{\mathbf{T}}\mathbf{A}\underline{e}, \end{aligned}$$

such that $\mathbf{W}$ is the smoothing matrix associated with nonparametric regression. Now we can rewrite F_{NP} as

$$F_{NP} = \frac{\underline{e}^{\mathbf{T}}\mathbf{B}\underline{e}}{\underline{e}^{\mathbf{T}}\mathbf{A}\underline{e}},$$

where $\mathbf{B} = \mathbf{I} - \mathbf{A}$. The p-value associated with F_{NP} can be written

$$\begin{aligned} p &= \Pr(F_{NP} > F_{obv}) \\ &= \Pr(\underline{e}^{\mathbf{T}}(\mathbf{I} - (1 + F_{obv})\mathbf{A})\underline{e} > 0) \\ &= \Pr(\underline{e}^{\mathbf{T}}\mathbf{C}\underline{e} > 0) \\ &= \Pr(\mathbf{Q} > 0), \end{aligned} \tag{2.46}$$

where F_{obv} is the value from the observed data.

The statistic F_{NP} does not have the standard properties of the usual F statistic. The quadratic form $\mathbf{Q}$ associated with F_{NP} is not positive definite. Procedures have been formulated by Bowman and Azzalini (1997) to find the distribution of F_{NP}. Basically the authors match the first three moments of $\mathbf{Q}$ with that of a shifted and scaled χ^2 distribution. The cummulants of $\mathbf{Q}$ are found as follows:

$$\kappa_j = 2^{j-1}(j-1)!tr\{(\mathbf{VC})^j\},$$

where tr is the trace operator and V is var-cov matrix of $\mathbf{e}$,

$$Var(\underline{e}) = \mathbf{I} - \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T} = \mathbf{V}.$$

The matching of moments of Q with a shifted and scaled χ^2 i.e. $a\chi^2_b + c$ distribution defines the following parameters

$$a = \frac{|\kappa_3|}{4\kappa_4},\ b = \frac{8\kappa_2^3}{\kappa_3^2},\ c = \kappa_1 - ab. \tag{2.47}$$

We therefore approximate the distribution of $\mathbf{Q}$ as a χ^2 with b degrees of freedom. Hence we find the p-value as follows:

$$p = 1 - \Pr(\mathbf{Q} > c/a). \tag{2.48}$$

Raz (1990) proposed a permutation test (bootstrap test) to find the distribution of F_{NP}. If H_o is true then the coupling of the x and y in the observed sample is random. The distribution of the statistic F_{NP} can then be generated by simulation, using parings of the observed $x's$ and $y's$ and constructing the corresponding F_{NP} statistic which we call F_{obv}. The empirical $p-$value of the test is then the proportion of F_{obv} which are larger

than the F_{NP} observed in the original data set or

$$\begin{aligned} \hat{p} &= \Pr(F_{obv} > F_{NP}) \\ &= \frac{\#(F_{obv} > F_{NP})}{k}, \end{aligned} \tag{2.49}$$

where k is the number of times F_{obv} was simulated.

Chapter 3

Empirical Comparisons of Generalized Regression and the Generalized Smoothing Estimators under a Simple Random Sampling Design

3.1 Introduction

This chapter will compare estimates of the population mean by contrasting the generalized regression and the generalized smoothing estimator developed in Chapter 2. For the generalized smoothing estimator the normal kernel regression and the spline regression will be considered. The sampling design that will be used throughout this chapter is *simple random sampling without replacement.* In order to understand the behavior of these different estimates the populations under investigation will have known point scatters.

In Section 3.2 we present a method to generate populations having a known point scatter. Subsequently in the section we simulate five populations having a known point

scatter and also present their finite population characteristics. The ratio estimator, the kernel estimator, and the spline estimator are developed under a simple random sample design in Section 3.3. In Section 3.4 criteria are developed to judge the efficiciency of the estimators. The results of the simulation procedure are presented in Section 3.5 while in Section 3.6 the conclusions are presented.

3.2 The Populations

Five artificial populations of size $N = 1000$ were created. The x_k values were generated by a gamma distribution $\Gamma(\theta_1, \theta_2)$ with density given by

$$f(x) = \frac{x^{\theta_1 - 1} \exp(-\frac{x}{\theta_2})}{\Gamma(\theta_1)\theta_2^{\theta_1}} \text{ such that } \theta_1 > 0,\ \theta_2 > 0, \text{ and } x > 0.$$

Conditional on x_k, the y_k values were also generated by a gamma distribution $\Gamma(\phi_1, \phi_2)$. Now x_k and y_k are related as follows $\phi_1 = \frac{\beta^2 x_k^r}{\sigma^2}$ and $\phi_2 = \frac{\sigma^2}{\beta x_k^m}$. The conditional mean and variance of y_k are as follows:

$$\begin{aligned} E_\epsilon(Y_k|X_k) &= \phi_1\phi_2 \\ &= \beta x_k^{r-m}, \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} V_\epsilon(Y_k|X_k) &= \phi_1\phi_2^2 \\ &= \sigma^2 x_k^{r-2m}. \end{aligned} \tag{3.2}$$

Equations (3.1) and (3.2) are used to create populations having different point scatters. If the point scatter has a mean $E_\epsilon(Y_k|X_k) = \beta x_k^l$ and a variance $V_\epsilon(Y_k|X_k) = \sigma^2 x_k^g$ we must then choose our parameters r and m as follows:

$$\begin{aligned} r &= 2l - g, \\ m &= l - g. \end{aligned}$$

The five populations were created using the following procedure:

1. 1000 x_k *values were simulated with a* $\Gamma(\theta_1 = 2, \theta_2 = 10)$,

2. *For each* x_k *value, and pairs* (β, σ) , (l, g) *a* y_k *was simulated as a* $\Gamma(\phi_1, \phi_2)$.

The parameter $\beta = 0.4, 2, 4$ and the parameter σ was fixed at 0.4. The parameters (l, g) were assigned the following values: $(1,1)$, $(2,1)$, $(0.5,1)$, $(1,.5)$, $(0.5,0.5)$, and $(0.5,1)$. The following two tables summarizes the major characteristics of the five populations.

Table 3.1

Population

Summary Statistics

	Population 1		Population 2		Population 3	
	$l = 1$		$l = 2$		$l = 0.5$	
	$g = 1$		$g = 1$		$g = 2$	
	$\beta = \sigma = 0.4$		$\beta = 10$, $\sigma = 0.4$		$\beta = 2$, $\sigma = 0.4$	
	X	Y	X	Y	X	Y
Mean	20.51	8.18	20.48	109.69	20.11	8.81
Standard Deviation	14.2	6.01	14.36	65.77	6.17	2.30
Skewness	1.27	1.39	1.36	2.57	0.55	0.62
Kurtosis	2.20	2.91	2.28	10.92	0.54	0.61
Coefficient of Variation	0.69	0.74	0.70	0.60	0.31	0.26
Correlation	0.957		0.569		0.61	

Table 3.2

Population

Summary Statistics

	Population 4		Population 5	
	$l = 1$, $g = 0.5$		$l = 0.5$, $g = 0.5$	
	$\beta = 2.0$, $\sigma = 0.4$		$\beta = 4.0$, $\sigma = 0.4$	
	X	Y	X	Y
Mean	19.55	8.18	20.01	25.53
Standard Deviation	13.67	12.52	11.42	1.39
Skewness	1.40	4.50	0.15	.036
Kurtosis	2.57	32.74	2.48	2.60
Coefficient of Variation	0.70	1.53	0.14	0.05
Correlation	0.521		0.579	

3.3 The Estimators

In this section the generalized regression estimate, and the generalized smoothing estimate are defined under a simple random sampling design. Their associated measures of precision are also presented.

3.3.1 Generalized Regression Estimator

Consider a model in which the point scatter $\left(x_k, \frac{y_k}{x_k}\right)$ is constant such that

$$E_\xi(y_k) = \beta x_k.$$

Moreover assume that the variance structure is proportional to the x_k around the regression line

$$V_\epsilon(y_k) = \sigma^2 x_k.$$

The above model is called a common ratio model . Särndal, Swenson and Wretman (1992) have shown that the generalized regression estimator for the population mean under a simple random sampling design has the following form:

$$\hat{m}_{yr} = \bar{X}_U \frac{\bar{y}}{\bar{x}}, \tag{3.3}$$

which is called the *ratio estimator* in the survey sampling literature. Now $\bar{X}_U$ is the population mean of the known x_k's while $\bar{y}$ and $\bar{x}$ are the sample means.

Using a Taylor series expansion, it can be shown that the estimator is approximately unbiased with variance estimator (see Cochran, 1977, Chapter 5)

$$\hat{V}_{srs}(\hat{m}_{yr}) = \left(\frac{\bar{X}_U}{\bar{x}}\right)^2 \hat{V}_0, \tag{3.4}$$

such that

$$\hat{V}_0 = \frac{1-f}{n} \frac{\sum_{k\in s}\left(y_k - \hat{\beta}x_k\right)^2}{n-1}, \tag{3.5}$$

with $\hat{\beta} = \dfrac{\sum_{k\in s} y_k}{\sum_{k\in s} x_k}$.

3.3.2 Kernel Regression Estimators

In the previous chapter we described and analyzed thegeneralized smoothing estimator

$$\hat{m}_{sm} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{N}\sum_{k\in s}\frac{e_k}{\pi_k},$$

where

$$\hat{y}_k = \frac{\sum_{j \in U} K\left(\frac{x_k - x_j}{b}\right) y_j}{\sum_{j \in U} K\left(\frac{x_k - x_j}{b}\right)},$$

and

$$e_k = y_k - \frac{\sum_{i \in s} K\left(\frac{x_k - x_i}{b}\right) y_i}{\sum_{i \in s} K\left(\frac{x_k - x_i}{b}\right)}$$

is the sample residuals where $K(.)$ is any kernel. Under a simple random sampling design $\pi_k = \frac{n}{N}$, the finite generalized smoothing estimate then has the following form:

$$\hat{m}_{sm} = \frac{1}{N} \sum_{k \in U} \hat{y}_k + \frac{1}{n} \sum_{k \in s} e_k. \tag{3.6}$$

In the simulation study we only considered the normal kernel which is defined as

$$K_N(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right), -\infty < u < \infty,$$

and this weighting function was used to estimate the finite population mean $\bar{Y}$. As previously stated all kernel regression estimators involve the bandwidth b, and optimal properties of the kernel estimates are functions of this parameter. The choice of b used throughout the simulation is due to Silverman (1986) who showed that an approximate value of the optimal b, b_{opt} is given by:

$$b_{opt} \approx 1.059 A n^{-\frac{1}{5}}, \tag{3.7}$$

such that

$$A = \min(S_x, IRQ/1.34),$$

where S_x is the sample standard deviation of the x's and IRQ is the sample interquartile range of the x's. Therefore the finite generalized smoothing estimate will have as

predicted values

$$\hat{y}_k = \frac{\sum_{j \in s} K\left(\frac{x_k - x_j}{b_{opt}}\right) y_j}{\sum_{j \in s} K\left(\frac{x_k - x_j}{b_{opt}}\right)}, \; k = 1, 2, ...N, \tag{3.8}$$

and residuals

$$e_k = y_k - \frac{\sum_{i \in s} K\left(\frac{x_k - x_i}{b_{opt}}\right) y_i}{\sum_{i \in s} K\left(\frac{x_k - x_i}{b_{opt}}\right)}, \; k = 1.2...n. \tag{3.9}$$

The precision of the fitted nonparametric regression curve was studied by Rice (1984), Gasser, Stroka, Steinmetz (1986), Hall, Marron (1988) and Hall, Kay, Titterington (1990). These papers propose to estimate the sampling variance of the kernel regression curve as:

$$s_e^2 = \left(n - 2trace(\mathbf{W}_s) + trace(\mathbf{W}_s^2)\right)^{-1} \sum_{k \in s} (y_k - \hat{y}_k)^2, \tag{3.10}$$

such that the elements of $\mathbf{W}_s$, are

$$w_{ij} = \frac{K\left(\frac{x_i - x_j}{b_{opt}}\right)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{b_{opt}}\right)} \text{ for } i = 1...n \text{ and } j = 1...n.$$

Under a simple random sampling design it was shown in the previous chapter that the estimated variance of $\hat{m}_{sm}$ was

$$\hat{V}_{srs}(\hat{m}_{sm}) = \frac{1 - f}{n} s_e^2.$$

Hence for kernel smoothing the estimated sampling variance is

$$\hat{V}_{srs}(\hat{m}_{sm}) = \frac{1 - f}{n} \left(\left(n - 2trace(\mathbf{W}_s) + trace(\mathbf{W}_s^2)\right)^{-1} \sum_{k \in s} (y_k - \hat{y}_k)^2\right). \tag{3.11}$$

3.3.3 Spline Estimators

Thegeneralized smoothing estimator

$$\hat{m}_{gm} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{N}\sum_{k\in s}\frac{e_k}{\pi_k},$$

is now adjusted for the spline estimator. Under a simple random sampling design $\pi_k = \frac{n}{N}$, and thegeneralized smoothing estimator using a spline method has the following form:

$$\hat{m}_{sp} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{n}\sum_{k\in s}e_k, \tag{3.12}$$

where $\hat{y}_k$ is a predicted spline value of y_k.

A cubic spline $\hat{\mu}(x_k)$ will be fitted for the sample values as follows:

a. A cubic polynomial fits the data between two successive sampled x_k values.

b. At the sampled values x_k, $\hat{\mu}(x_k)$ and its two first derivatives are continuous.

c. At the boundary points $x_{(1)}$ and $x_{(n)}$, the second derivative of $\hat{\mu}(x_k)$ is zero.

Therefore

$$\hat{y}_k = \hat{\mu}(x_k) \text{ for all } x_k \in U,$$

and

$$e_k = y_k - \hat{\mu}(x_k) \text{ for all } x_k \in s.$$

The estimated spline smoother $\hat{\mu}(x_k)$ is a function of the spanning parameter λ in the penalized residual sum of squares

$$S_\lambda(\mu(.)) = \sum_{k\in s}(y_k - \mu(x_k))^2 + \lambda\int\left(\mu''(t)\right)^2 dt.$$

The optimal value of λ was found by cross-validating the sum of squares

$$CV(\lambda) = \frac{1}{n}\sum_{k\in s}\left(y_k - \hat{\mu}_\lambda^{-k}(x_k)\right)^2,$$

where $\hat{\mu}_{\lambda}^{-k}(x_k)$ denotes the fit at x_k by leaving out that data point. Let λ_{opt} be the value of λ that minimizes $CV(\lambda)$. We then use λ_{opt} to find $\hat{\mu}(x_k)$ which we shall now call $\hat{\mu}_{\lambda_{opt}}(x_k)$. The predicted y_k and residuals will now be found with the λ_{opt} i.e.

$$\hat{y}_{k\lambda_{opt}} = \hat{\mu}_{\lambda_{opt}}(x_k) \text{ for all } x_k \in U,$$

and

$$e_{k\lambda_{opt}} = y_k - \hat{\mu}_{\lambda_{opt}}(x_k) \text{ for all } x_k \in s.$$

The optimal spline regression estimator will now be

$$\hat{m}_{sp} = \frac{1}{N}\sum_{k\in U}\hat{y}_{k\lambda_{opt}} + \frac{1}{n}\sum_{k\in s} e_{k\lambda_{opt}}. \tag{3.13}$$

The precision of the spline was shown by Hastie and Tibshirani (1995) to be

$$s^2_{sp} = \frac{1}{n - trace(\mathbf{S}_{\lambda_{opt}}\mathbf{S}'_{\lambda_{opt}})}\sum_{k\in s}\left(y_k - \hat{\mu}_{\lambda_{opt}}(x_k)\right)^2, \tag{3.14}$$

where $S_{\lambda_{opt}}$ is the symmetric projection matrix which is the same as the smoothing matrix $\mathbf{W}_s$. Under a simple random sampling design it was demonstrated in the previous chapter that the estimated variance was

$$\hat{V}_{srs}(\hat{m}_{sm}) = \frac{1-f}{n}s^2_e,$$

hence for spline smoothing the estimated sampling variance is

$$\hat{V}_{srs}(\hat{m}_{sp}) = \frac{1-f}{n}\left(\frac{1}{n - trace(\mathbf{S}_{\lambda_{opt}}\mathbf{S}'_{\lambda_{opt}})}\sum_{k\in s}\left(y_k - \hat{\mu}_{\lambda_{opt}}(x_k)\right)^2\right). \tag{3.15}$$

3.4 The Simulation Procedure

The simulations were carried out in S-Plus Version 3.3 as follows: First a simple random sample of size $n = 10, 25$ was chosen in each of the five populations. Secondly for each sample the following estimates of the finite population mean were calculated: $\hat{m}_{GRE}$, $\hat{m}_{sm}$, $\hat{m}_{sp}$. Thirdly for each sample the following measures were calculated,

1. the sample bias $\hat{B}_{srs}(\hat{m}_{(.)}) = \bar{Y} - \hat{m}_{(.)}$,
2. the sample variance $\hat{V}_{srs}(\hat{m}_{(.)})$.

The following relative measures of performance were also found:

1. The absolute bias ratio

$$BR(\hat{m}_{(.)}) = \frac{|B(\hat{m}_{(.)})|}{\sqrt{\hat{V}_{srs}(\hat{m}_{(.)})}} \times 100. \tag{3.16}$$

2. The absolute relative bias

$$RB(\hat{m}_{(.)}) = \left| \frac{B(\hat{m}_{(.)})}{\bar{Y}} \right| \times 100. \tag{3.17}$$

The procedure is repeated for a total of $K = 1000$ times for each of the sample sizes $n = 10, 25$. Let $\hat{m}_{j(.)}$ denote the estimate from the j th sample, we then calculate

$$\bar{m}_{(.)} = \frac{1}{1000} \sum_{j=1}^{1000} \hat{m}_{j(.)}, \tag{3.18}$$

which is an estimate of $E\left(\hat{m}_{(.)}\right)$. The same calculations were done for $\bar{B}(\hat{m}_{(.)})$, $\bar{V}_{srs}(\hat{m}_{(.)})$, $\overline{BR}(\hat{m}_{(.)})$ and $\overline{RB}(\hat{m}_{(.)})$. We also calculated the confidence interval at the approximate 95% level,

$$\hat{m}_{(.)} \pm 1.96 \left[\hat{V}_{srs}(\hat{m}_{(.)})\right]^{\frac{1}{2}}$$

and then counted the number of intervals C out of K that contain the true value of $\bar{Y}$.

We call this the coverage ratio and define it as

$$C.R.\left(\hat{m}_{(.)}\right) = \frac{C}{K} \times 100. \tag{3.19}$$

Lastly we calculated the simulation mean squared error and the simulation variance with the following:

$$\overline{mse}\left(\hat{m}_{(.)}\right) = \frac{1}{1000}\sum_{j=1}^{1000}\left(\hat{m}_{j(.)} - \bar{Y}\right)^2, \tag{3.20}$$

$$\overline{V}_{sim}\left(\hat{m}_{(.)}\right) = \overline{mse}\left(\hat{m}_{(.)}\right) - \bar{B}^2(\hat{m}_{(.)}). \tag{3.21}$$

We compared the efficiency of our proposed estimators with that of the generalized regression estimator with

$$Eff\left(\hat{m}_{(.)}\right) = \frac{\overline{mse}_{srs}\left(\hat{m}_{GRE}\right)}{\overline{mse}_{srs}\left(\hat{m}_{(.)}\right)}. \tag{3.22}$$

Values of $Eff\left(\hat{m}_{(.)}\right) > 1$ imply that the proposed method of estimation is superior to that of the generalized regression estimator. If $Eff\left(\hat{m}_{(.)}\right) < 1$ this implies that the proposed method of estimation is inferior to the generalized regression estimator.

In the analysis the relative accuracy of $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$was compared to $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$ for the three methods of estimation. To be more specific the relative accuracy was defined as:

$$R.A. = \left(1 - \frac{\overline{V}_{sim}\left(\hat{m}_{(.)}\right) - \bar{V}_{srs}\left(\hat{m}_{(.)}\right)}{\overline{V}_{sim}\left(\hat{m}_{(.)}\right)}\right). \tag{3.23}$$

Values of $R.A. > 1$ imply that $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$ over-estimates $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$, while $R.A. < 1$, implies that $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$ under-estimates $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$.

3.5 Results of the Simulation Study

The following tables summarizes the Monte Carlo characteristics of the five populations considered for sample sizes $n = 10$ and 25. Table 3.3 contains the value of the population

parameter $\bar{Y}$ and the average estimates of this parameter found by using $\bar{m}_{GRE}$, $\bar{m}_{sm}$ and $\bar{m}_{sp}$. Table 3.4 contains the average bias of the estimators, while Table 3.5 presents the average absolute relative bias of the estimators.

Table 3.3

Average Estimate of Population Mean

		Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\bar{Y}$	$\bar{m}_{GRE}$	$\bar{m}_{sm}$	$\bar{m}_{sp}$	$\bar{m}_{GRE}$	$\bar{m}_{sm}$	$\bar{m}_{sp}$
Population 1	8.18	8.17	7.86	8.18	8.19	8.02	8.20
Population 2	109.69	116.66	110.87	105.73	111.67	111.45	108.28
Population 3	8.81	8.85	8.78	8.85	8.85	8.82	8.84
Population 4	8.18	8.23	7.83	8.27	8.10	7.95	8.13
Population 5	25.53	25.59	25.51	25.52	25.57	25.53	25.54

Table 3.4

Average Bias of the Estimators

$\bar{B}(\hat{m}_{(.)})$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	−0.02	−0.32	0.00	0.01	−0.15	0.02
Population 2	6.96	1.18	−3.97	1.98	1.76	−1.42
Population 3	0.03	−0.03	0.04	0.03	0.01	0.03
Population 4	0.04	−0.35	0.08	−0.08	−0.24	−0.06
Population 5	0.06	−0.02	−0.02	0.04	−0.00	0.01

Table 3.5

Average Absolute Relative Bias of the Estimators

$\overline{RB}(\hat{m}_{(.)})$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	0.2%	3.8%	0.0%	0.2%	1.9%	0.2%
Population 2	6.3%	1.1%	3.6%	1.8%	1.6%	1.3%
Population 3	0.4%	0.4%	0.4%	0.4%	0.1%	0.4%
Population 4	0.5%	4.3%	0.9%	1.0%	3.0%	0.7%
Population 5	0.2%	0.1%	0.1%	0.2%	0.0%	0.0%

The following patterns emerge form these tables:

1. For all estimators as the sample size increases the bias $\bar{B}(\hat{m}_{(.)})$ and relative bias $\overline{RB}(\hat{m}_{(.)})$ decrease for the five populations considered. In Chapter 2 it was shown that the generalized smoothing estimator is asymptotically unbiased. Tables 3.4 and 3.5 confirm this fact albeit the sample sizes were small.

2. In population 1 the true regression model is linear and the bias and relative bias of the $\hat{m}_{GRE}$ is negligible. The result is not surprising because the generalized regression estimate $\hat{m}_{GRE}$ is the optimal estimate for the superpopulation model when $l = 1$ and $g = 1$.

3. In populations 3 and 5 for both sample sizes the bias $\bar{B}(\hat{m}_{(.)})$ and the absolute relative bias $\overline{RB}(\hat{m}_{(.)})$ of the three estimators are essentially the same.

4. In population 2 the bias $\bar{B}(\hat{m}_{GRE})$ and the absolute relative bias $\overline{RB}(\hat{m}_{GRE})$ of the generalized regression estimate $\hat{m}_{GRE}$ was much larger than the other two estimators considered. The result is not surprising because the generalized regression estimate $\hat{m}_{GRE}$ is not the optimal estimate for this superpopulation model.

5. In population 4 at the sample sizes considered, the bias $\bar{B}(\hat{m}_{sm})$ and the absolute relative bias $\overline{RB}(\hat{m}_{sm})$ of the finite generalized smoother $\hat{m}_{sm}$ is much larger than the other two estimates considered.

6. The bias that is always present in non-parametric regression near the edges of the region over which the data have been collected, has been reduced. The reduction is due to the fact that we are measuring the center of the regression curve and the bias that occurs at both extremes of the regression curve cancel out. Moreover the results confirm the approximate design unbiasedness property of Theorem 2.5.2.

Table 3.6, 3.7 and 3.8 contain the average sample variance $\bar{V}_{srs}(\hat{m}_{(.)})$,coverage ratio $C.R.(\hat{m}_{(.)})$ and finally the average bias ratio $\overline{BR}\left(\hat{m}_{(.)}\right)$. The bias ratio is considered because it has an effect on the coverage ratio. When $\left|\overline{BR}\left(\hat{m}_{(.)}\right)\right| \leq 10\%$ the bias effect may be ignored because the coverage ratio is approximately equal to the nominal coverage probability.

Table 3.6

Average Sample Variance

$\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	0.302	1.469	0.558	0.121	1.000	0.152
Population 2	2063.7	391.7	464.1	686.6	208.8	108.4
Population 3	0.481	0.452	0.561	0.203	0.257	0.165
Population 4	8.742	11.244	32.831	3.876	7.157	5.913
Population 5	0.935	0.131	0.208	0.364	0.068	0.059

Table 3.7

Average Absolute Bias Ratio

$\overline{BR}(\hat{m}_{(.)})$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	11.5%	9.6%	0.0%	2.87%	1.5%	5.1%
Population 2	15.3%	15.6%	18.4%	7.5%	12.2%	13.6%
Population 3	4.3%	4.5%	5.3%	4.2%	1.9%	3.5%
Population 4	13.4%	12.5%	12.4%	12.2%	11.0%	12.4%
Population 5	6.2%	6.5%	4.4%	6.1%	0.0%	4.1%

Table 3.8

Coverage Ratio

$C.R.(\hat{m}_{(.)})$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	90.2%	92.6%	90.1%	92.5%	99.2%	91.4%
Population 2	88.9%	84.3%	87.3%	92.6%	83.8%	89.9%
Population 3	92.2%	93.2%	91.9%	93.9%	97.8%	93.6%
Population 4	82.3%	84.8%	80.0%	87.3%	85.4%	83.9%
Population 5	90.7%	87.6%	93.2%	93.8%	94.8%	93.7%

The following conclusions can be inferred from these tables:

8. For all estimators as the sample size increases the sample variance $\bar{V}_{srs}(\hat{m}_{(.)})$and the average bias ratio decreases for the five populations considered.

9. In population 1 the true regression model is linear and the sample variance of $\hat{m}_{GRE}$ is much smaller than the other two estimates considered. As previously stated the generalized regression estimate $\hat{m}_{GRE}$ is optimal for this population.

10. In population 3 for both sample sizes the average sample variance of the three estimators are essentially the same magnitude. While in population 2 and 5 the average sample variance of $\hat{m}_{GRE}$ is much larger than the other two estimates. But in population 4 the average sample variance of $\hat{m}_{GRE}$ is smaller than $\hat{m}_{sm}$ and $\hat{m}_{sp}$.

11. The sample size has an effect on the coverage rate $C.R.(\hat{m}_{(.)})$. As the sample size increases the coverage rate $C.R.(\hat{m}_{(.)})$ approaches the nominal 95%, which confirms the asymptotic theory developed in section 2.9 of Chapter 2.

12. Table 3.7 demonstrates the effect that the bias can have on the coverage ratio Table 3.8. As the bias ratio decreases the coverage ratio approaches the 95% nominal level. Therefore the coverage rate $C.R.(\hat{m}_{(.)})$ is affected by the sample size and the bias ratio.

Tables 3.9 and 3.10 contain the mean squared error of simulation $\overline{mse}_{sim}(\hat{m}_{(.)})$and the associated sample variance $\overline{V}_{sim}(\hat{m}_{(.)})$. These estimators are consistent estimates of the design mean squared error $mse_{srs}(\hat{m}_{(.)})$ and of the design variance $V_{srs}(\hat{m}_{(.)})$.

Table 3.9

Mean Squared Error of Simulation

$\overline{mse}_{sim}(\hat{m}_{(.)})$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	0.315	0.790	0.707	0.121	0.215	0.249
Population 2	1880.8	926.1	366.8	657.5	461.8	108.7
Population 3	0.483	0.376	0.584	0.203	0.143	0.165
Population 4	11.224	10.702	22.601	4.105	4.232	11.442
Population 5	0.988	0.141	0.243	0.366	0.054	0.227

Table 3.10

Variance of Simulation

$\overline{V}_{sim}(\hat{m}_{(.)})$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	0.315	0.691	0.707	0.121	0.191	0.249
Population 2	1832.3	924.7	351.1	653.7	458.7	106.7
Population 3	0.484	0.373	0.585	0.202	0.143	0.165
Population 4	11.223	10.648	22.675	4.132	4.245	11.476
Population 5	0.985	0.140	0.242	0.365	0.054	0.227

From Tables 3.9 and 3.10 the following can be inferred:

13. For all estimators as the sample size increases the mean squared error of simulation $\overline{mse}_{sim}(\hat{m}_{(.)})$and the sample variance $\overline{V}_{sim}(\hat{m}_{(.)})$ decreases.

14. For population 1 the $\overline{mse}_{sim}(\hat{m}_{GRE})$ and $\overline{V}_{sim}(\hat{m}_{GRE})$ is smaller than that of the other two estimators because $\hat{m}_{GRE}$ is optimal.

15. In populations 2, 3 and 5 and for both sample sizes the $\overline{mse}_{sim}(\hat{m}_{GRE})$ and $\overline{V}_{sim}(\hat{m}_{GRE})$ is much larger than the other two estimators. While in population 4 the $\overline{mse}_{sim}(\hat{m}_{GRE})$ and $\overline{V}_{sim}(\hat{m}_{GRE})$ is of the same magnitude as that of $\overline{mse}_{sim}(\hat{m}_{sm})$ and $\overline{V}_{sim}(\hat{m}_{sm})$.

Table 3.11 demonstrates the efficiency of each estimator considered in the simulation for each of the populations studied.

Table 3.11

Efficiency

$Eff(\hat{m}_{(.)})$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	1.00	0.40	0.45	1.00	0.56	0.49
Population 2	1.00	2.03	5.13	1.00	1.42	6.05
Population 3	1.00	1.32	0.84	1.00	1.47	1.27
Population 4	1.00	1.04	0.49	1.00	0.97	0.36
Population 5	1.00	7.02	4.07	1.00	6.73	1.62

The following can be inferred from Table 3.11:

16. The sample size has an effect on the efficiency of the estimator. As the sample size increases the efficiency increases.

17. The estimator $\hat{m}_{GRE}$ is only efficient in population 1 because in this population the generalized regression estimator has optimal properties.

18. The estimators $\hat{m}_{GRE}$ and $\hat{m}_{sm}$ are essentially of equal efficiency for population 4 and for both sample sizes considered.

19. The estimators $\hat{m}_{sm}$ and $\hat{m}_{sp}$ are more efficient than $\hat{m}_{GRE}$ for populations 2, 3, and 5.

The p-values for the lack of fit test presented in section 2.10 using a shifted and scaled χ^2 distribution is presented in Table 3.12 while the p-values for the permutation (bootstrap) test is presented in Table 3.13. In both tables we considered $\hat{m}_{GRE}$ for the null hypothesis and $\hat{m}_{em}$ for the research hypothesis.

Table 3.12

Lack of Fit Test

p-values for χ^2

	$n = 10$	$n = 25$
Population 1	0.50	0.95
Population 2	0.00	0.00
Population 3	0.00	0.00
Population 4	0.00	0.00
Population 5	0.00	0.00

Table 3.13

Lack of Fit Test

p-values for Permutation

	$n = 10$	$n = 25$
Population 1	1.00	1.00
Population 2	0.01	0.00
Population 3	0.02	0.00
Population 4	0.01	0.01
Population 5	0.01	0.01

The shifted and scaled χ^2 test and the permutation (bootstrap) lack of fit test permits us to make the following conclusions:

20. For population 1 the linear model adequately describes the fit of the point scatter. Consequently, the generalized regression estimator $\hat{m}_{GRE}$ would the appropriate estimator of the finite population mean. This result confirms the efficiency analysis found in Table 3.11.

21. We can reject the hypothesis of a linear fit at a significance level of 5% for populations 2, 3, 4 and 5. For these finite populations a kernel or a spline regression would fit the point scatter adequately and ultimately provide an efficient estimator of the finite population mean. Again these results confirm those found in Table 3.11.

Finally Table 3.14 shows the relative accuracy of the variance estimator $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$ with respect to the simulation variance $V_{sim}\left(\hat{m}_{(.)}\right)$.

Table 3.14

Relative Accuracy

$RA(\hat{m}_{(.)})$

	Sample Size = 10(1%)			*Sample Size* = 25(2.5%)		
	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$	$\hat{m}_{GRE}$	$\hat{m}_{sm}$	$\hat{m}_{sp}$
Population 1	0.96	2.13	0.79	1.00	5.22	0.61
Population 2	1.13	0.42	1.32	1.05	0.46	1.02
Population 3	1.00	1.23	0.96	1.00	1.78	1.00
Population 4	0.78	1.06	1.45	0.94	1.70	0.52
Population 5	0.95	0.93	0.86	1.00	1.26	0.26

The above table demonstrates that:

22. As the sample size increases the relative accuracy of the sampled variance estimator $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$ approaches the sample variance $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$.

23. The relative accuracy of $\bar{V}_{srs}\left(\hat{m}_{GRE}\right)$ was very good for $\hat{m}_{GRE}$.

24. The relative accuracy of $\bar{V}_{srs}\left(\hat{m}_{sm}\right)$ and $\bar{V}_{srs}\left(\hat{m}_{sp}\right)$ were not as favorable as the relative accuracy of $\bar{V}_{srs}\left(\hat{m}_{GRE}\right)$. The methods used to find $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$ were studied

by Rice (1984), Gasser, Stroka, Steinmetz (1986), Hall, Marron (1988) and Hall, Kay, Titterington (1990). Research is still being pursued in this area because of the bias present in $\hat{m}_{(.)}$ which will have the effect of inflating the size of the residual sum of squares.

3.6 Conclusions

This chapter has shown that it is possible to estimate a finite population mean with a nonparametric model that makes no assumption about the underlying point scatter. We are using the advice of Eubank (1988) '*let the data speak for itself*' or as stated by Hastie and Tibshirani (1990) '*let the data show us the appropriate functional form*' without making parametric model assumptions.

In keeping with the spirit of Härdle (1989, 1990) the motivation for the nonparametric regression approach is as follows:

i. to provide a versatile method for exploring a general relationship between two variables,

ii. to give predictions for observations yet to be made without reference to a fixed parametric model,

iii. to furnish a tool for finding spurious observations by studying the influence of isolated points,

iv. to have a flexible method of substituting for missing values or interpolating between adjacent X values.

Also Hansen, Madow and Tepping (1981) asked researchers the following '*The proper use of models has much to contribute to survey design. We urge continuing strong efforts, taking the fullest feasible advantage of models, but ordinarily within the framework of probability sampling, i. e., using designs and estimators that are not model-dependent*'.

We urge survey statisticians to use nonparametric estimators because they are model-independent and to use a probability sampling so as to protect the survey statistician

'*against failures of* ***assumed models and provide robustness for all estimators***' (**Hansen** et al. 1981).

Chapter 4

Empirical Comparisons of Generalized Regression and the Generalized Smoothing Estimators under a Stratified Sample Design

4.1 Introduction

In this chapter we compare estimates of the population mean using parametric and nonparametric models. The sampling design that will be used throughout this chapter is *stratified sampling without replacement.* Considerable gains in efficiency can be obtained with this design if the strata are well defined and if the allocation of sampled units in each strata is done with the Neyman (1934) criterion. In order to understand the behavior of these different estimates the populations under investigation will have known point scatters.

In section 4.2 we present a method to generate populations having a known point scatter in each strata. Subsequently in the section we simulate the three populations and also present their finite population characteristics. The ratio estimator, the kernel

estimator, and the spline estimator are developed under a stratified sample design in section 4.3. In section 4.4 criteria are developed to judge the efficacy of the estimators under a stratified random sample design. The results of the simulation procedure are presented in section 4.5 while in section 4.6 the conclusions are presented.

4.2 The Populations

Three artificial populations of size $N = 1000$ were created. The populations had two strata of size $N_h = 500$ where $h = 1, 2$. For each strata x_{hk} values were generated by a gamma distribution $\Gamma(\theta_{h1}, \theta_{h2})$, Conditional on x_{hk}, y_{hk} values for strata h were also generated by a gamma distribution $\Gamma(\phi_{h1}, \phi_{h2})$, where $\phi_{h1} = \frac{\beta_h^2 x_{hk}^r}{\sigma_h^2}$ and $\phi_{h2} = \frac{\sigma_h^2}{\beta_h x_{hk}^m}$. The conditional mean and variance of y_{hk} in stratum h is as follows:

$$\begin{aligned} E_\epsilon(Y_{hk}|X_{hk}) &= \phi_{h1}\phi_{h2} \\ &= \beta_h x_{hk}^{r-m}, \end{aligned} \tag{4.1}$$

and

$$\begin{aligned} V_\epsilon(Y_{hk}|X_{hk}) &= \phi_{h1}\phi_{h2}^2 \\ &= \sigma_h^2 x_{hk}^{r-2m}. \end{aligned} \tag{4.2}$$

Equations (2.1) and (2.2) are used to create strata having different point scatters. If the point scatter has a mean $E_\epsilon(Y_{hk}|X_{hk}) = \beta_h x_{hk}^l$ and a variance $V_\epsilon(Y_{hk}|X_{hk}) = \sigma_h^2 x_{hk}^g$ we must then choose our parameters r and m as follows:

$$\begin{aligned} r &= 2l - g, \\ m &= l - g. \end{aligned} \tag{4.3}$$

The three populations were created as follows:

Strata 1

1. *Simulated* 500 x_{1k} *with a* $\Gamma(\theta_{11} = 2, \theta_{12} = 10)$,
2. *For each* x_{1k} *value, and pairs* (β, σ) and (l, g), *simulated a* y_{1k} *value with a*

$\Gamma(\phi_{11}, \phi_{12})$.

Strata 2

1. Simulated 500 x_{2k} *with a* $\Gamma(\theta_{21} = 2.5, \theta_{22} = 20)$,

2. For each x_{2k} *value, and pairs* (β, σ) and (l, g) *simulated a* y_{2k} *value with a* $\Gamma(\phi_{21}, \phi_{22})$.

The following values of (l, g) were considered $(0.5, 1)$, $(1, 1)$, $(1.5, 1)$. The pair (β, σ) were given values of $(0.4, 0.4)$, $(4.0, 0.4)$ and $(-10.0, 0.4)$ The following three tables summarizes the major characteristics of the populations created.

Table 4.1

Population 6

Summary Statistics

	Strata 1		Strata 2		Population	
	$l = 1$ and $g = 1$		$l = 1$ and $g = 1$			
	$\beta = \sigma = 0.4$		$\beta = \sigma = 0.4$			
	X	Y	X	Y	X	Y
Mean	19.68	7.94	50.87	20.25	35.27	14.09
Standard Deviation	6.25	3.13	32.82	13.13	28.30	11.35
Skewness	0.69	0.84	1.11	1.11	1.94	1.89
Kurtosis	0.88	1.68	1.15	1.26	4.02	3.99
Coefficient of Variation	0.32	0.39	0.65	0.65	0.80	0.81
Correlation	0.829		0.980		0.981	

Table 4.2

Population7

Summary Statistics

	Strata 1		Strata 2		Population	
	$l = 1$ and $g = 1$		$l = 1.5$ and $g = 1$			
	$\beta = \sigma = 0.4$		$\beta = \sigma = 0.4$			
	X	Y	X	Y	X	Y
Mean	19.92	7.89	49.38	12.67	34.65	10.28
Standard Deviation	13.71	5.77	31.92	2.02	28.64	4.94
Skewness	1.30	1.19	1.50	1.05	1.89	0.12
Kurtosis	2.21	1.63	3.95	0.97	5.72	0.73
Coefficient of Variation	0.69	0.73	0.65	0.16	0.83	0.48
Correlation	0.955		0.407		0.609	

Table 4.3

Population 8

Summary Statistics

	Strata 1		Strata 2		Population	
	$l = 0.5$ and $g = 1$		$l = 0.5$ and $g = 1$			
	$\beta = -10$ and $\sigma = 0.4$		$\beta = 4$ and $\sigma = 0.4$			
	X	Y	X	Y	X	Y
Mean	19.53	108.72	9.95	25.42	14.74	67.07
Standard Deviation	13.04	55.42	1.36	1.42	10.43	57.20
Skewness	1.16	1.45	0.03	0.110	2.20	1.68
Kurtosis	1.96	2.64	3.94	4.09	6.18	3.18
Coefficient of Variation	0.67	0.51	0.14	0.06	0.71	0.85
Correlation	−0.625		0.545		−0.043	

4.3 The Estimators

In this section we adjust the estimators analyzed in Chapter 3. We take into consideration the fact that our population is stratified into two strata.

4.3.1 Generalized Regression Estimator

We first consider a model in which the point scatter $\left(x_{hk}, \frac{y_{hk}}{x_{hk}}\right)$ $(h = 1, 2, ...H)$ is constant in each strata such that

$$E_\epsilon(y_{hk}) = \beta_h x_{hk}.$$

Moreover assume that the variance structure in each strata is proportional to the x_{hk} around the regression line

$$V_\epsilon(y_{hk}) = \sigma_h^2 x_{hk}.$$

The above model is called the group ratio model.

The generalized regression estimator for the h'th stratum mean is given by:

$$\hat{m}_{yrh} = \bar{X}_{U_h} \frac{\bar{y}_h}{\bar{x}_h}, \ h = 1, ..., H,$$

where $\bar{X}_{U_h}$ is the population mean in strata h of the known x_{hk}'s while $\bar{y}_h$ and $\bar{x}_h$ are the sample means of the sample chosen in strata h. Pooling information from each stratum an estimate of the population mean is given by:

$$\hat{m}_{styr} = \sum_{h=1}^{H} W_h \hat{m}_{yrh}, \tag{4.4}$$

where $W_h = \frac{N_h}{N}$ and $N = \sum_{h=1}^{H} N_h$, this estimate is called the *separate ratio estimator* of the population mean.

Using a Taylor series expansion it can be shown that the estimator is approximately

unbiased with an estimate of its variance (see Särndal et al Chapter 11),

$$\hat{V}_{strs}(\hat{m}_{styr}) = \sum_{h=1}^{H} W_h^2 \left(\frac{\hat{X}_{U_h}}{\bar{x}_h}\right)^2 \hat{V}_h, \tag{4.5}$$

such that

$$\hat{V}_h = \frac{1-f_h}{n_h} \frac{\sum_{k\in s}\left(y_{hk} - \hat{\beta}_h x_{hk}\right)^2}{n_h - 1},$$

where $\hat{\beta}_h = \dfrac{\sum_{k\in s} y_{hk}}{\sum_{k\in s} x_{hk}}$.

4.3.2 Kernel Regression Estimators

In the previous chapter it was shown that the estimate of the population mean using kernel regression is

$$\hat{m}_{sm} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{n}\sum_{k\in s} e_k, \tag{4.6}$$

under a simple random sampling design The fitted values are defined as

$$\hat{y}_k = \frac{\sum_{j\in U} K\left(\frac{x_k - x_j}{b_{opt}}\right) y_j}{\sum_{j\in U} K\left(\frac{x_k - x_j}{b_{opt}}\right)},$$

and the residuals by

$$e_k = y_k - \frac{\sum_{i\in s} K\left(\frac{x_k - x_i}{b_{opt}}\right) y_i}{\sum_{i\in s} K\left(\frac{x_k - x_i}{b_{opt}}\right)}.$$

The finite generalized smoothing estimate for the h'th stratum mean is given by:

$$\hat{m}_{smh} = \frac{1}{N_h}\sum_{k\in U_h}\hat{y}_{hk} + \frac{1}{n_h}\sum_{k\in s_h} e_{hk}.$$

Pooling information from each stratum an estimate of the population mean is given by:

$$\hat{m}_{stsm} = \sum_{h=1}^{H} W_h \hat{m}_{smh}, \tag{4.7}$$

where $W_h = \frac{N_h}{N}$ and $N = \sum_{h=1}^{H} N_h$. The fitted values are defined as

$$\hat{y}_{hk} = \frac{\sum_{j \in U_h} K\left(\frac{x_{hk} - x_{hj}}{b_{hopt}}\right) y_{hj}}{\sum_{j \in U_h} K\left(\frac{x_{hk} - x_{hj}}{b_{hopt}}\right)},$$

while the residuals have the following form

$$e_{hk} = y_{hk} - \frac{\sum_{i \in s} K\left(\frac{x_{hk} - x_{hi}}{b_{hopt}}\right) y_{hi}}{\sum_{i \in s} K\left(\frac{x_{hk} - x_{hi}}{b_{hopt}}\right)}.$$

The optimal bandwidth b_{hopt} is found for each strata as follows

$$b_{hopt} \approx 1.0059 A n^{-\frac{1}{5}}, \tag{4.8}$$

such that

$$A_h = \min(S_{x_h}, IRQ_h/1.34)$$

and S_{x_h} is the sample standard deviation of the x's in strata h and IRQ_h the sample interquartile range of the x again in strata h.

As in Chapter 3 an estimate for the variance in strata h is given by

$$s^2_{e_h} = \left(n_h - 2trace(\mathbf{W}_{s_h}) + trace(\mathbf{W}^2_{s_h})\right)^{-1} \sum_{k \in s_h} (y_{hk} - \hat{y}_{hk})^2, \tag{4.9}$$

such that the elements of $\mathbf{W}_{s_h}$, are

$$w_{ij_h} = \frac{K\left(\frac{x_{hi}-x_{hj}}{b_{hopt}}\right)}{\sum_{j\in s} K\left(\frac{x_{hi}-x_{hj}}{b_{hopt}}\right)} \text{ for } i = 1...n. \text{ and } h = 1...H.$$

The estimated variance $\hat{V}_{strs}(\hat{m}_{stsm})$ is found by pooling the estimate for the variance $s^2_{e_h}$ in strata h

$$\hat{V}_{strs}(\hat{m}_{stsm}) = \sum_{h=1}^{H} W_h^2 \frac{1-f_h}{n_h} s^2_{e_h}. \tag{4.10}$$

4.3.3 Spline Estimators

The spline regression estimator

$$\hat{m}_{sp} = \frac{1}{N}\sum\nolimits_{k\in U} \hat{y}_k + \frac{1}{n}\sum\nolimits_{k\in s} e_k,$$

is adjusted for a stratified sampling design. A cubic spline $\hat{\mu}_h(x_{hk})$ is fitted for the sample values in each strata using the same procedures as described in Chapter 3. Therefore

$$\hat{y}_{hk} = \hat{\mu}_h(x_{hk}) \text{ for all } x_{hk} \in U_h,$$

and

$$e_{hk} = y_{hk} - \hat{\mu}(x_{hk}) \text{ for all } x_{hk} \in s_h.$$

The optimal value of the spanning parameter λ_h is found by the cross-validating the sum of squares in each strata

$$CV(\lambda_h) = \frac{1}{n_h}\sum\nolimits_{k\in s_h} \left(y_{hk} - \hat{\mu}^{-k}_{h_\lambda}(x_{hk})\right)^2,$$

where $\hat{\mu}^{-k}_{h_\lambda}(x_{hk})$ denotes the fit at x_{hk} by leaving out that data point. Let λ_{hopt} be the value of λ_h that minimizes $CV(\lambda_h)$. We then use λ_{hopt} to find $\hat{\mu}_h(x_k)$ which we shall now

call $\hat{\mu}_{h\lambda_{opt}}(x_k)$. The predicted y_k and residuals will now be found with the λ_{hopt} i.e.

$$\hat{y}_{hk_{\lambda_{hopt}}} = \hat{\mu}_{h_{\lambda_{hopt}}}(x_{hk}) \text{ for all } x_{hk} \in U_h,$$

and

$$e_{hk_{\lambda_{hopt}}} = y_{hk} - \hat{\mu}_{h\lambda_{hopt}}(x_k) \text{ for all } x_{hk} \in s_h.$$

The spline estimate in each strata is

$$\hat{m}_{sph} = \frac{1}{N_h}\sum\nolimits_{k\in U_h} \hat{y}_{hk} + \frac{1}{n_h}\sum\nolimits_{k\in s} e_{hk},$$

where $\hat{y}_{hk}$ is a predict y_{hk} value for all $x_{hk} \in U_h$. Pooling information from each stratum an estimate of the population mean is given by:

$$\hat{m}_{stsp} = \sum_{h=1}^{H} W_h \hat{m}_{sph}, \tag{4.11}$$

where $W_h = \frac{N_h}{N}$ and $N = \sum_{h=1}^{H} N_h$.

The precision of the spline is given by the following

$$s_h^2 = \left(\sum\nolimits_{k\in s_h} \frac{\left(y_{hk} - \hat{\mu}_{h_{\lambda h_{opt}}}(x_{hk})\right)^2}{n_h - tr(S_{h_{\lambda_{hopt}}} S^T_{h_{\lambda_{hopt}}})} \right),$$

where $S_{h_{\lambda_{hopt}}}$ is the symmetric projection matrix for strata h. An estimate of the variance of $\hat{m}_{stsp}$ is is found by pooling the estimate for the variance $s^2_{e_h}$ in strata h

$$\hat{V}_{strs}(\hat{m}_{stsp}) = \sum_{h=1}^{H} W_h^2 \frac{1-f_h}{n_h} s_h^2. \tag{4.12}$$

4.4 The Simulation Procedure

The simulations were carried out in S Plus Version 3.3. Now since all the x_{hk} are known for each strata considerable gains in efficiency can be made if the total sample size is allocated correctly. Since the population standard deviation S_{xU_h} are known, the Neyman criterion allocates the total sample n to the strata as follows:

$$n_h = n\frac{N_h S_{xU_h}}{\sum_{h=1}^{H} N_h S_{xU_h}}$$

The simulation study for each artificial population was carried out as follows: First a stratified random sample of size $n = 20, 40$ was chosen in each of the three populations. Secondly for each sample the following estimates of the finite population mean were calculated: $\hat{m}_{styr}$, $\hat{m}_{stsm}$ and $\hat{m}_{stsp}$. Finally the measures of precision, relativity, efficiency and accuracy defined in Chapter 3 were calculated so as to judge the behavior of $\hat{m}_{styr}$, $\hat{m}_{stsm}$ and $\hat{m}_{stsp}$.

4.5 The Simulations Results

The following tables summarizes the Monte Carlo characteristics of the five populations considered for sample sizes $n = 20$ and 40. Table 4.4 contains the value of the population parameter $\bar{Y}$ and the average estimates of this parameter found by using : $\hat{m}_{styr}$, $\hat{m}_{stsm}$ and $\hat{m}_{stsp}$. Table 4.5 contains the average bias of the estimators, while Table 4.6 presents the average absolute relative bias of the estimators.

Table 4.4

Average Estimate of Population Mean

		Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\bar{Y}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	14.09	14.09	13.70	14.12	14.11	13.87	14.10
Population 7	10.28	10.49	10.18	10.39	10.39	10.22	10.31
Population 8	67.07	71.65	68.18	65.48	68.61	67.60	66.84

Table 4.5

Average Bias of the Estimators

$\bar{B}(\hat{m}_{(.)})$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	0.00	−0.39	0.03	0.02	−0.22	0.01
Population 7	0.21	−0.11	0.11	0.11	−0.06	0.03
Population 8	4.58	1.11	−1.59	1.54	0.52	−0.24

Table 4.6

Average Absolute Relative Bias of the Estimators

$\overline{RB}(\hat{m}_{(.)})$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	0.0%	2.8%	0.2%	0.1%	1.6%	0.1%
Population 7	2.0%	1.0%	1.0%	1.1%	0.6%	0.3%
Population 8	6.8%	1.7%	2.4%	2.3%	0.8%	0.4%

From tables 4.4, 4.5 and 4.6 the following can be inferred:

1. For all estimators as the sample size increases the bias $\bar{B}(\hat{m}_{(.)})$ and relative bias $\overline{RB}(\hat{m}_{(.)})$ decreases for the three populations considered. This confirms the theory developed in Chapter 2 with respect to the asymptotic unbiasedness.

2. In population 6 the true regression model is linear and the bias and relative bias of the $\hat{m}_{styr}$ and $\hat{m}_{stsp}$ are negligible. The result is not surprising for the generalized regression estimate, $\hat{m}_{styr}$ is the optimal estimate for the superpopulation model when $l = 1$ and $g = 1$.

3. In population 7 for both sample sizes the bias $\bar{B}(\hat{m}_{(.)})$ and the absolute relative bias $\overline{RB}(\hat{m}_{(.)})$ of $\hat{m}_{stsm}$ and $\hat{m}_{stsp}$ are essentially the same.

4. In population 8 for both sample sizes the bias $\bar{B}(\hat{m}_{(.)})$ and the absolute relative bias $\overline{RB}(\hat{m}_{(.)})$ of $\hat{m}_{styr}$ is much larger than the other two estimators considered.

5. Again the bias that is always present in non-parametric regression has been reduced. The reduction is due to the fact that we are measuring the center of the regression curve and the bias that occurs at both extremes of the regression curve cancel out. Moreover the results confirm the approximate design unbiasedness property of Theorem 2.5.2.

Table 4.7, 4.8 and 4.9 contain the average sample variance $\bar{V}_{strs}(\hat{m}_{(.)})$,coverage ratio $C.R.(\hat{m}_{(.)})$and finally the average bias ratio $\overline{BR}(\hat{m}_{(.)})$. The bias ratio is considered because it has an effect on the coverage ratio. When $|\overline{BR}(\hat{m}_{(.)})| \leq 10\%$ the bias effect may be ignored because the coverage ratio is approximately equal to the nominal coverage probability.

Table 4.7

Average Sample Variance

$\bar{V}_{strs}(\hat{m}_{(.)})$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	0.23	0.86	0.37	0.11	0.32	0.14
Population 7	2.12	0.22	0.50	1.02	0.16	0.19
Population 8	483.3	155.2	78.4	186.9	94.1	25.06

Table 4.8

Average Absolute Bias Ratio

$\overline{BR}(\hat{m}_{(.)})$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	0.8%	42.2%	4.3%	2.4%	5.7%	2.6%
Population 7	20.3%	22.4%	21.2%	15.7%	15.8%	8.6%
Population 8	20.8%	17.8%	17.9%	11.3%	5.4%	4.7%

Table 4.9

Coverage Ratio

$C.R.(\hat{m}_{(.)})$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	90.5%	94.5%	92.5%	91.8%	99.0%	92.7%
Population 7	90.9%	91.0%	91.0%	93.1%	95.6%	91.3%
Population 8	90.2%	75.9%	89.3%	92.4%	84.2%	92.7%

The following patterns emerge from the tables:

6. For all estimators as the sample size increases the sample variance $\bar{V}_{strs}\left(\hat{m}_{(.)}\right)$ and the average bias ratio decreases for the three populations considered.

7. In population 6 the true regression model is linear and the sample variance of $\hat{m}_{stsm}$ is much smaller than the other two estimates considered. As previously stated the generalized regression estimate $\hat{m}_{stsm}$ is optimal for this population.

8. In populations 7 and 8 the sample variance of $\hat{m}_{stsm}$ is much larger than the sample variance of $\hat{m}_{stsm}$ and $\hat{m}_{stsp}$.

9. The sample size has an effect on the coverage rate $C.R.(\hat{m}_{(.)})$. As the sample size increases the coverage rate $C.R.(\hat{m}_{(.)})$ approaches the nominal 95%, which confirms the asymptotic theory developed in section 2.9 of Chapter 2. This result is similar to result 10 of Chapter 3 which was for simple random sampling.

10. Table 4.8 demonstrates the effect that the bias can have on the coverage ratio. As the bias ratio decreases the coverage ratio approaches the 95% nominal level.

Tables 4.10 and 4.11 contain the mean squared error of simulation $\overline{mse}_{sim}\left(\hat{m}_{(.)}\right)$and the associated sample variance $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$. These estimators are consistent estimates of the design mean squared error $mse_{strs}\left(\hat{m}_{(.)}\right)$ and of the design variance $V_{strs}\left(\hat{m}_{(.)}\right)$.

Table 4.10

Mean Squared Error of Simulation

$\overline{mse}_{sim}\left(\hat{m}_{(.)}\right)$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	0.26	0.80	0.45	0.13	0.31	0.16
Population 7	1.89	0.26	0.89	0.76	0.12	0.12
Population 8	508.4	217.0	56.7	182.8	99.8	24.9

Table 4.11

Variance of Simulation

$\overline{V}_{sim}(\hat{m}_{(.)})$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	0.26	0.65	0.45	0.13	0.26	0.15
Population 7	1.84	0.25	0.87	0.74	0.12	0.12
Population 8	487.5	215.8	54.1	180.4	99.5	24.8

11. For all estimators as the sample size increases the mean squared error of simulation $\overline{mse}_{sim}(\hat{m}_{(.)})$ and the sample variance $\overline{V}_{sim}(\hat{m}_{(.)})$ decreases.

12. For population 6 the $\overline{mse}_{sim}(\hat{m}_{styr})$ and $\overline{V}_{sim}(\hat{m}_{styr})$ is smaller than that of the other two estimators because $\hat{m}_{styr}$ is optimal.

13. In populations 7, 8 and for both sample sizes the $\overline{mse}_{sim}(\hat{m}_{GRE})$ and $\overline{V}_{sim}(\hat{m}_{GRE})$ is much larger than the other two estimators.

Table 4.12 demonstrates the efficiency of each estimator considered in the simulation for each of the populations studied.

Table 4.12

Efficiency

$Eff(\hat{m}_{(.)})$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	1.00	0.32	0.58	1.00	0.42	0.85
Population 7	1.00	7.19	2.13	1.00	6.08	6.44
Population 8	1.00	2.34	8.95	1.00	1.83	7.34

The following results are apparent:

14. The sample size has an effect on the efficiency of the estimator. As the sample size increases the efficiency increases.

15. The estimator $\hat{m}_{styr}$ is the only efficient estimator in population 6 because in this population the generalized regression estimator is optimal.

16. The estimators $\hat{m}_{stsm}$ and $\hat{m}_{stsp}$ are more efficient than $\hat{m}_{styr}$ for populations 7, and 8.

Finally Table 4.13 shows the relative accuracy of the variance estimator $\bar{V}_{strs}\left(\hat{m}_{(.)}\right)$ with respect to the design variance $V_{sim}\left(\hat{m}_{(.)}\right)$.

Table 4.13

Relative Accuracy

$RA(\hat{m}_{(.)})$

	Sample Size = 20(2%)			*Sample Size* = 40(4%)		
	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$	$\hat{m}_{styr}$	$\hat{m}_{stsm}$	$\hat{m}_{stsp}$
Population 6	0.89	1.31	0.83	0.84	1.22	0.89
Population 7	1.15	0.87	0.57	1.38	1.31	1.64
Population 8	0.99	0.72	1.45	1.04	0.95	1.01

The above table demonstrates that:

17. As the sample size increases the relative accuracy of the sampled variance estimator $\bar{V}_{strs}\left(\hat{m}_{(.)}\right)$ approaches the sample variance $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$ in population 8. For the other two populations the results are mixed.

4.6 Conclusions

This chapter has demonstrated that a finite population mean can be estimated with a nonparametric model under a stratified sampling design. The methodology makes no as-

sumption about the underlying point scatter. In practice, any regression model is likely to have some error, by misspecification of the mathematical form of the model or by omitting important explanatory variables. The nonparametric regression methods presented in this theses are in a sense robust, every survey statistician should have this tool in her/his toolbox, because inferences made with nonparametric methodologies protect the statistician against model misspecification.

Chapter 5

The Generalized Smoothing Estimator and Nonparametric Binary Regression for Nonresponse

5.1 Introduction

In many sample surveys some of the units contacted do not respond to all the items on a questionnaire. Such non- response, is common in practice whenever the population consists of units such as individual people, households, or businesses.

Suppose a population consists of N units and y is the characteristic we are interested to measure. For simplicity assume the total population is surveyed and denote by μ_R and μ_{R^c} the population means of the responding population and the nonresponding population. Also let N_R and N_{R^c} ($N = N_R + N_{R^c}$) represent the sizes of the responding and nonresponding populations. The population mean is a weighted average of μ_R and μ_{R^c}

$$\mu = \frac{N_R\mu_R + N_{R^c}\mu_{R^c}}{N}. \tag{5.1}$$

Let us now discard the nonresponding population and use μ_R the population mean of the

respondents as an estimate of μ. If the population mean of the respondents μ_R is different than the population mean of the nonrespondents μ_{R^c} we induce a bias in our estimate of the population mean μ. The following Lemma quantifies this bias as a product of the difference of the population means of both subpopulations and the nonresponse rate.

Lemma 5.1.1 *The bias B incurred by using μ_R to estimate μ is given by*

$$B = R^c(\mu_{R^c} - \mu_R), \tag{5.2}$$

where $R^c = \dfrac{N_{R^c}}{N}$ is the nonresponse rate.

Proof *Since*

$$\mu = \frac{N_R\mu_R + N_{R^c}\mu_{R^c}}{N},$$

and

$$N = N_R + N_{R^c},$$

we have

$$\mu = \frac{(N - N_{R^c})\,\mu_R + N_{R^c}\mu_{R^c}}{N},$$

or

$$\begin{aligned} \mu &= \mu_R + \frac{N_{R^c}}{N}(\mu_{R^c} - \mu_R) \\ &= \mu_R + R^c(\mu_{R^c} - \mu_R). \end{aligned}$$

But the bias is defined as

$$\begin{aligned} B &= \mu - \mu_R \\ &= R^c(\mu_{R^c} - \mu_R). \end{aligned}$$

■

Reducing R^c so as to reduce the bias B is a problem of data collection. Many methods have been developed to reduce the nonresponse rate. The fundamental work in this area are attributed to Hansen and Hurwitz (1946) and to Politz and Simmons (1949, 1950). This thesis does not address this issue as it is not a statistical problem.

Reducing $\mu_{R^c} - \mu_R$, so as to reduce the bias B is on the other hand a problem of statistical methodology. Consider a population $U = \{1, 2, ..., k, ...N\}$, and let U_R represent those units in the population that respond and U_{R^c} those units that do not respond ($U_R \cup U_{R^c} = U$). Suppose the missing values Y_k ($k \in U_{R^c}$) are estimated by $\hat{y}_k$. We then can estimate μ by

$$\bar{y} = \frac{N_R \mu_R + N_{R^c} \hat{\mu}_{R^c}}{N}, \tag{5.3}$$

where

$$\hat{\mu}_{R^c} = \frac{1}{N_{R^c}} \sum_{k \in U_{R^c}} \hat{y}_k.$$

and N_R and N_{R^c} are the sizes of the populations U_R and U_{R^c} respectively. Note that μ_R is considered known because it represents the mean of the responding units.

Lemma 5.1.2 *The bias B incurred by using $\bar{y}$ to estimate μ is given by*

$$B = R^c(\mu_{R^c} - \hat{\mu}_{R^c}), \tag{5.4}$$

where $R^c = \frac{N_{R^c}}{N}$ is the nonresponse rate.

Remark 5.1.1 *The above lemma demonstrates that the nonresponse bias can be reduced for a fixed nonresponse rate if $\hat{\mu}_{R^c}$ is a good estimate of the population mean of the nonrespondents μ_{R^c}.*

The main thrust of this chapter is to develope methods that reduce the bias incurred because of nonresponse. In section 5.2 we review a nonresponse model due to Nargundkar and Joshi (1975), and introduce the concept of response probability. Section 5.3 reviews the superpopulation model literature for the nonresponse problem. In this framework two models are considered, namely a regression model for the point scatter and a response model for the respondents. The theory developed in Chapter 2 is now adapted in section 5.4 and 5.5 for the nonresponse problem. Finally in section 5.6 the response probabilities are estimated by binary regression and these estimates are then used to estimate the population mean.

5.2 A Model for the Response Mechanism

Nargundkar and Joshi (1975) adjusted the Horvitz - Thompson (1952) estimator

$$\hat{m}_{HT} = \sum_{k \in s} \frac{y_k}{N\pi_k},$$

so as to take into account the nonresponse behavior. Consider a population $U = \{1, 2, ..., k, ...N\}$, of size N. Let s be a sample of fixed size n drawn from U according to a known sampling design $p(s)$ such that

$$p(s) \geq 0 \text{ for all } s \in \mathcal{L},$$

and

$$\sum_{s \in \mathcal{L}} p(s) = 1,$$

where $\mathcal{L}$ is the set of all s of fixed size n. The inclusion probability of unit k is defined as

$$\pi_k = \sum_{s \in \mathcal{L}_k} p(s) \text{ such that } \pi_k > 0 \text{ for all } k \in U,$$

where $\mathcal{L}_k = \{s : k \in s\}$. We also define the joint inclusion probability of units k and l as

$$\pi_{kl} = \sum_{s \in \mathcal{L}_{kl}} p(s) \text{ such that } \pi_{kl} > 0 \text{ for all } k, l \in U,$$

where $\mathcal{L}_{kl} = \{s : k, l \in s\}$.

The response mechanism has the following distribution $q(r|s)$ such that for every fixed s, r denoting the members of s which is responding,

$$q(r|s) \geq 0 \text{ for all } r \in \mathcal{R}_s,$$

and

$$\sum_{r \in \mathcal{R}_s} q(r|s) = 1,$$

where $\mathcal{R}_s = \{r : r \subseteq s\}$. We define the response probability of unit k given s as

$$\phi_k = \sum_{r \in \mathcal{R}_{ks}} q(r|s) \text{ such that } \phi_k > 0 \text{ for all } k \in U,$$

where $\mathcal{R}_{ks} = \{r : k \in R \subseteq s\}$. Also the joint response probabilities of units k and l are defined as

$$\phi_{kl} = \sum_{r \in \mathcal{R}_{kls}} q(r|s) \text{ such that } \phi_{kl} > 0 \text{ for all } k, l \in U$$

where $\mathcal{R}_{kls} = \{r : k, l \in r \subseteq s\}$. Assume that the units respond independently of each other and of s :

$$q(r|s) = \prod_{k \in R} \phi_k \prod_{k \in s-R} (1 - \phi_k).$$

The ϕ_k may be functions of unknown parameters and auxiliary variables x_k known for all population units $\{k = 1, ...N\}$ such that $\phi_k = f_k(x_k; \theta)$.

Using these response probabilities Nargundkar and Joshi (1975) modified the Hovitz-Thompson estimator as

$$\hat{m}_{NJ} = \sum_{k \in R} \frac{y_k}{N \pi_k \phi_k}. \tag{5.5}$$

An estimator M is said to be design unbiased for the population mean $\bar{Y}$ if

$$E_p(M) = \bar{Y},$$

where the expectation is taken with respect to the sampling design p. This is also known as 'p unbiased'. Since we have introduced a nonresponse distribution into the estimation procedure unbiasedness must now be defined with respect to the design p and the response

mechanism q. This is called 'pq unbiased' and is defined as

$$\begin{aligned} E_{pq}(M) &= E_p(E_q(M|s)) \\ &= \bar{Y}. \end{aligned}$$

Lemma 5.2.1 *The modified Hovitz- Thompson estimator*

$$\hat{m}_{NJ} = \sum_{k \in R} \frac{y_k}{N \pi_k \phi_k},$$

is pq unbiased.

Proof We have for all $\phi_k > 0$ the following relationship

$$\begin{aligned} E_{pq}(\hat{m}_{NJ}) &= E_p(E_q(\hat{m}_{NJ}|s)) \\ &= E_p\left(E_q\left(\sum_{k \in R} \frac{y_k}{N \pi_k \phi_k}\right)\right) \\ &= E_p\left(\sum_{k \in s} \frac{y_k}{N \pi_k}\right). \end{aligned}$$

Now for all $\pi_k > 0$ we have

$$E_p\left(\sum_{k \in s} \frac{y_k}{N \pi_k}\right) = \bar{Y}. \blacksquare$$

Remark 5.2.1 *The modified Hovitz- Thompson estimator is pq unbiased if the true distribution of the respondents is $q(r|s)$. In practice this distribution is unknown. The analyst must make assumptions about this distribution.*

Nargundkar and Joshi (1975) also showed that an unbiased estimate of the variance is given by:

$$\hat{V}_{pq}(\hat{m}_{NJ}) = \frac{1}{N^2}\left(\sum_{k\in R}\phi_k^2 y_k^2\left(\pi_k^{-1}-1\right)+2\sum_{\substack{k,l\in U_t\\ k\neq l}}\sum_{\substack{k,l\in U_t\\ k\neq l}}\phi_k\phi_l y_k y_l\left(\pi_{kl}\pi_k^{-1}\pi_l^{-1}-1\right)\right)$$
$$\substack{k,l\in R\\ k<l}$$
$$+\frac{1}{N^2}\left(\sum_{k\in R}\phi_k\left(1-\phi_k\right)y_k^2\pi_k^{-1}+2\sum_{\substack{k,l\in U_t\\ k\neq l}}\sum_{\substack{k,l\in U_t\\ k\neq l}}\left(\phi_{kl}-\phi_k\phi_l\right)y_k y_l\pi_{kl}\pi_k^{-1}\pi_l^{-1}\right) \tag{5.6}$$
$$\substack{k,l\in R\\ k<l}$$

with the first two terms due to the sampling design while the last two terms attributable to the response mechanism which is unknown.

Remark 5.2.2 *The above estimate of variance has the same inherent problems as pq unbiasedness. The $q(r|s)$ distribution is unknown and therefore $\hat{V}_{pq}(\hat{m}_{NJ})$ is ineffective if the ϕ_k are unknown.*

The above procedure can be useful for surveys that are repetitious. A statistician can then formulate and test different response distributions $q(r|s)$ and hence find her/his ϕ_k. In subsequent sections of this chapter we are going to review and make proposals to estimate ϕ_k.

5.3 Superpopulation Regression Models and the Non-response Problem

Hansen, Madow and Tepping (1983) has become a corner stone in the survey sampling literature. The thesis of this paper was that only inferences based on probability sampling protected the survey statistician '*against failures of assumed models and provide*

robustness for all estimators'. The authors also encouraged research in what they coined model dependent designs. They stated '*The proper use of models has much to contribute to survey design. We urge continuing strong efforts, taking the fullest feasible advantage of models, but ordinarily within the framework of probability sampling, i. e., using designs and estimators that are not model-dependent*'. Casel, Särndal and Wretman (1979) (hereafter CSW) undertook this challenge and developed a statistical model for the non-response problem within the framework of probability sampling.

Consider a linear model ξ such that $Y_1, Y_2, ...Y_N$ are independent and

$$\begin{aligned} E_\xi(Y_k) &= \underset{\sim}{\mathbf{x}}_k' \underline{\beta} \\ &= \sum_{j=0}^{p} \beta_j x_{kj} \\ &= \mu_k, \end{aligned}$$

and

$$V_\xi(Y_k) = \sigma^2 v_k,$$

where $\underline{\beta}$ is a vector of $(p+1)$ unknown coefficients, $\underline{\mathbf{x}}_k'$ is a vector of $(p+1)$ known auxiliary variables for all $k = 1, 2, ...N$. Also the σ^2 is unknown and $v_k = v(\underline{\mathbf{x}}_k')$ for a known function $v(.)$.

Assume that the individual response probabilities ϕ_k are dependent on the known vector of auxiliary variables $\underset{\sim}{\mathbf{x}}_k'$

$$\phi_k = f(\underset{\sim}{\mathbf{x}}_k'; \underline{\theta}),$$

where $\underline{\theta}$ is a $(p+1)$ unknown vector of coefficients that can be estimated from the available data.

The method proposed by CSW can be summarized by the following steps:

1. *The unknown $\underline{\theta}$ is estimated by minimizing the likelihood function*

$$L(\underline{\theta}) = \prod_{k \in R} \phi_k \prod_{k \in s-R} (1 - \phi_k).$$

2. *The estimated parameter* $\hat{\underline{\theta}}$ *is used to estimate the individual response probabilities*

$$\hat{\phi}_k = f\left(\underline{\mathbf{x}}_k'; \underline{\theta}\right).$$

3. *The estimated response probabilities are then used to estimate the unknown vector of regression coefficients* $\underline{\beta}$

$$\hat{\underline{\beta}} = \left(\mathbf{x}_r' V_r^{-1} \Pi_r^{-1} \Phi_r^{-1} \mathbf{x}_r\right)^{-1} V_r^{-1} \Pi_r^{-1} \Phi_r^{-1} \mathbf{Y}_r,$$

where V_r, Π_r, *and* Φ_r *are* $(n_r \times n_r)$ *diagonal matrices with diagonal elements respectively of* v_k, π_k *and* $\hat{\phi}_k$ $(k \in r)$ *;now* $\mathbf{x}_r'$ *is a* $((p+1) \times n_r)$*matrix with column vectors* $\mathbf{x}_k, (k \in r)$ *and* $\mathbf{Y}_r$ *is a column vector having* $Y_k, (k \in r)$ *as elements.*

4. *The estimated* $\hat{\underline{\beta}}$ *are then used in forming an estimate* $\hat{m}_r$ *of the population mean*

$$\hat{m}_{rq} = \sum_{k \in R} \frac{Y_k}{N \pi_k \hat{\phi}_k} + \sum_{j=0}^{p} \hat{\beta}_j \left(\bar{x}_{.j} - \sum_{k \in R} \frac{x_k}{N \pi_k \hat{\phi}_k} \right). \tag{5.7}$$

Remark 5.3.1 *The main difficulty with the above procedure is specifying a proper model for* $\phi_k = f\left(\underline{\mathbf{x}}_k'; \underline{\theta}\right)$. *The* $q(r|s)$ *distribution is unknown and therefore* m_r *maybe model biased because of response model mispecification.*

Särndal and Hui (1981) investigated the properties of this method by means of Monte Carlo experiments. They concluded that if the regression model is representative of the population point scatter, then the estimator $\hat{m}_r$ may still be design unbiased even if the response probabilities are wrongly estimated using the model. If the regression model is not representative population point scatter then $\hat{m}_r$ is still design unbiased if the response mechanism is correctly modeled but then the variance of $\hat{m}_r$ increases.

The choice of a response model introduces assumptions regarding the behavior of the response mechanism. As we have seen these assumptions have an effect on the robustness of the estimate $\hat{m}_r$ and $\hat{V}_{pq}(\hat{m}_r)$. When auxiliary information is available Giommi (1985, 1987) showed that the response probabilities can be estimated nonparametrically.

Giommi (1985) proposed to estimate ϕ_k by considering a nearest neighbor estimate:

$$\hat{\phi}_k^D = \frac{\sum_{j \in s} i_k D(x_k - x_j)}{\sum_{j \in s} D(x_k - x_j)}, \tag{5.8}$$

where

$$D(x_k - x_j) = \begin{cases} 1 & \text{if } |x_k - x_j| \leq h \\ 0 & \text{otherwise} \end{cases}$$

such that h is a percentage (10%, 30% and 50%) of the range of the sampled x values and $i_k = 1$ if $k \in R$ and 0 otherwise. In the nonparametric literature h is called the span.

Giommi (1987) considered the following, kernel estimate of ϕ_k is

$$\hat{\phi}_k^{D^*} = \frac{\sum_{j \in r} D^*(x_k - x_j)}{\sum_{j \in s} D^*(x_k - x_j)}, \tag{5.9}$$

such that

$$D^*(x_k - x_j) = (2\pi h^2)^{-\frac{1}{2}} \exp\left(-\frac{(x_k - x_j)^2}{2h^2}\right),$$

where h is again a percentage (10%, 30% and 50%) of the range of the sampled x values.

The $\hat{\phi}_k$ are then used to estimate $\bar{Y}$ with

$$\hat{m}_r^G = \sum_{k \in R} \frac{y_k}{N \pi_k \hat{\phi}_k} + \hat{\beta}\left(\bar{X} - \sum_{k \in R} \frac{x_k}{N \pi_k \hat{\phi}_k}\right). \tag{5.10}$$

Monte Carlo experiments were conducted to study the characteristics of $\hat{\phi}_k^D$ and $\hat{\phi}_k^{D^*}$. The bias and variance of $\hat{m}_r^G$ using $\hat{\phi}_k^D$ are less than $\hat{\phi}_k^{D^*}$ for various values of h. Both procedures produced better estimates in terms of bias and variance than

$$\hat{m}_r = \sum_{k \in R} \frac{y_k}{N \pi_k} + \hat{\beta}\left(\bar{X} - \sum_{k \in R} \frac{x_k}{N \pi_k}\right) \tag{5.11}$$

which is the estimate $\bar{Y}$ with no correction for nonresponse. The problem with both methods of estimating ϕ_k is the specification of an optimal value of h. Niyonsenga (1994) used the same procedure as Giommi (1985, 1987) but the h was based on a function of the ranks of the auxiliary variables. The h value again was not optimal. Latter in this chapter we will present a new procedure to estimate the values of ϕ_k. The bandwidth used in the procedure will be optimal and will not have the same deficiencies as that of Giommi (1985, 1987) and Niyonsenga (1994).

5.4 Nonparametric Regression Estimation for the Nonresponse Problem

It was demonstrated in Theorem 2.5.2 that the nonparametric regression estimator

$$\hat{m}_{st} = \frac{1}{N}\sum_{k\in U}\hat{\mu}(x_k) + \frac{1}{N}\sum_{k\in s}\frac{e_k}{\pi_k},$$

is asymtotically design-unbiased (*p unbiased)* and design-consistent for $\bar{Y}$. The above model is modified as

$$\hat{m}_{smr} = \frac{1}{N}\sum_{k\in U}\hat{\mu}(x_k) + \frac{1}{N}\sum_{k\in R}\frac{e_k}{\pi_k}, \tag{5.12}$$

so as to take into account the nonresponse. The estimates $\hat{\mu}(x_k)$ can be either a kernel or a spline estimates of y_k based on a responding set r instead of a sample set s.

Remark 5.4.1 *If the sampling design is simple random sampling take* $\pi_k = \frac{r}{N}$, *because in this case all units in the population have the same probability of response.*

Now using Theorem 2.8.2 with R replacing s one observes that $\hat{m}_{smr}$ is approximately design unbiased with design variance

$$V_p(\hat{m}_{smr}) = \frac{1}{N^2}\sum_{k}\sum_{l\in R}\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)\frac{E_k}{\pi_k}\frac{E_l}{\pi_l}, \tag{5.13}$$

which can be estimated by

$$\hat{V}_p(\hat{m}_{smr}) = \frac{1}{N^2}\sum_k\sum_{l\in R}\frac{1}{\pi_{kl}}\left(\frac{\pi_{kl}}{\pi_k\pi_l}-1\right)\frac{e_k}{\pi_k}\frac{e_l}{\pi_l}, \tag{5.14}$$

such that $e_k = y_k - \hat{y}_k$. Again with r replacing s in Theorem 2.5.2 one observes that $\hat{m}_{str}$ is asymtotically design unbiased and consistent for $\bar{Y}$.

The estimate $\hat{m}_{smr}$ may still be *q biased* because of the unknown nature of the responding mechanism. CSW (1979), introduced responding probabilities so as to take into account the problem of *q unbiasedness*.. Using the same procedure we adjust $\hat{m}_{smr}$ as

$$\hat{m}_{smrq} = \frac{1}{N}\sum_{k\in U}\hat{\mu}(x_k) + \frac{1}{N}\sum_{k\in R}\frac{e_k}{\phi_k\pi_k}, \tag{5.15}$$

where ϕ_k is the response probability for unit $k \in R$. The bias and variance of $\hat{m}_{smrq}$ are established in the following theorem.

Theorem 5.4.1 *The estimator $\hat{m}_{smrq}$ is approximately p-q unbiased for $\bar{Y}$ and has an approximate p-q variance*

$$\begin{aligned} V_p(\hat{m}_{smrq}) \doteq\ & \frac{1}{N^2}\sum_k\sum_{l\in U}\left(\frac{\pi_{kl}}{\pi_k\pi_l}-1\right)\frac{E_k}{\pi_k}\frac{E_l}{\pi_l} \\ & + \frac{1}{N^2}E_p\left(\sum_k\sum_{l\in s}\left(\frac{\phi_{kl}}{\phi_k\phi_l}-1\right)\frac{E_k^*}{\phi_k}\frac{E_l^*}{\phi_k}\,|s\right), \end{aligned} \tag{5.16}$$

such that $E_k^* = \dfrac{E_k}{\pi_k}$. An unbiased estimator of $V_p(\hat{m}_{smrq})$ is

$$\begin{aligned} \hat{V}_p(\hat{m}_{smrq}) =\ & \frac{1}{N^2}\sum_k\sum_{l\in R}\frac{1}{\pi_{kl}}\left(\frac{\pi_{kl}}{\pi_k\pi_l}-1\right)\frac{e_k}{\pi_k}\frac{e_l}{\pi_l} \\ & + \frac{1}{N^2}\sum_k\sum_{l\in R}\frac{1}{\phi_{kl}}\left(\frac{\phi_{kl}}{\phi_k\phi_l}-1\right)\frac{e_k^*}{\phi_k}\frac{e_l^*}{\phi_l}, \end{aligned} \tag{5.17}$$

such that $e_k^* = \dfrac{e_k}{\pi_k}$ and $e_k = y_k - \hat{y}_k$ is the sample residual.

Proof In Theorem 2.8.2 we showed that

$$\hat{m}_{sm} \doteq \frac{1}{N}\sum_{k\in U} y_k^o + \frac{1}{N}\sum_{k\in s}\frac{E_k}{\pi_k},$$

therefore we can approximate $\hat{m}_{smrq}$ by

$$\hat{m}_{smrq} \doteq \frac{1}{N}\sum_{k\in U} y_k^o + \frac{1}{N}\sum_{k\in R}\frac{E_k}{\phi_k\pi_k}.$$

Now

$$\begin{aligned} E_p(\hat{m}_{smrq}) &= E_q(E_p(\hat{m}_{smrq}|s)) \\ &= \frac{1}{N}\sum_{k\in U} y_k^o + E_q\left(\frac{1}{N}\sum_{k\in s}\frac{E_k}{\phi_k}.\right) \\ &= \frac{1}{N}\sum_{k\in U} y_k^o \\ &= \bar{Y}, \end{aligned}$$

is approximately pq unbiased.

Now the variance of a random variable can be written as the sum of the variance of the conditional expectation and the expected value of conditional variances i.e.

$$V(X) = V_A(E(X|A)) + E_A(V(|A)).$$

Using the above identity we find the variance of $\hat{m}_{smrq}$ as

$$V_p(\hat{m}_{smrq}) = V_p(E_q(\hat{m}_{smrq}|s)) + E_p(V_q(\hat{m}_{smrq}|s)).$$

But

$$\begin{aligned} E_q(\hat{m}_{smrq}|s) &= E_q\left(\frac{1}{N}\sum_{k\in R}\frac{y_k}{\pi_k\phi_k} - \frac{1}{N}\left(\sum_{k\in R}\frac{\hat{\mu}_w(x_k)}{\phi_k} - \sum_{k\in U}\hat{\mu}(x_k)\right)\right) \\ &= \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} - \frac{1}{N}\left(\sum_{k\in R}\frac{\hat{\mu}_w(x_k)}{\phi_k} - \sum_{k\in U}\hat{\mu}(x_k)\right) \\ &= \hat{m}_{sm}. \end{aligned}$$

But in Theorem 2.8.2 we found

$$V_p(\hat{m}_{sm}) = \frac{1}{N^2} \sum_{\substack{k,l \in U_t \\ k \neq l}} \sum_{k,l \in U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) \frac{E_k}{\pi_k} \frac{E_l}{\pi_l},$$

therefore

$$V_p(E_q(\hat{m}_{smrq} | s)) = \frac{1}{N^2} \sum_{\substack{k,l \in U_t \\ k \neq l}} \sum_{k,l \in U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}$$

Now

$$\hat{m}_{smrq} = \frac{1}{N} \sum_{k \in U} y_k^o + \frac{1}{N} \sum_{k \in R} \frac{E_k}{\phi_k \pi_k},$$

maybe written as

$$\hat{m}_{smrq} = \frac{1}{N} \sum_{k \in U} y_k^o + \frac{1}{N} \sum_{k \in R} \frac{E_k^*}{\phi_k},$$

where $E_k^* = \frac{E_k}{\pi_k}$. Therefore using Theorem 2.8.1, we have that

$$V_q(\hat{m}_{smrq} | s) = \frac{1}{N^2} \sum_{\substack{k,l \in U_t \\ k \neq l}} \sum_{k,l \in R} \left(\frac{\phi_{kl}}{\phi_k \phi_l} - 1 \right) \frac{E_k^*}{\phi_k} \frac{E_l^*}{\phi_k}.$$

Therefore

$$\begin{aligned} V_p(\hat{m}_{smrq}) &= \frac{1}{N^2} \sum_{\substack{k,l \in U_t \\ k \neq l}} \sum_{k,l \in U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) \frac{E_k}{\pi_k} \frac{E_l}{\pi_l} \\ &\quad + \frac{1}{N^2} E_p \left(\sum_{\substack{k,l \in U_t \\ k \neq l}} \sum_{k,l \in R} \left(\frac{\phi_{kl}}{\phi_k \phi_l} - 1 \right) \frac{E_k^*}{\phi_k} \frac{E_l^*}{\phi_l} | s \right) \end{aligned}$$

An unbiased estimate is found by using Theorem 2.8.2. ■

Example 5.4.1 *Suppose the sampling design is simple random sampling then* $\pi_k = \frac{n}{N}$ *and* $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$. *Now with this sampling design thegeneralized smoothing estimator*

$\hat{m}_{smrq}$ *is*

$$\hat{m}_{smrq} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{n}\sum_{k\in R}\frac{e_k}{\phi_k},$$

with approximate design variance

$$V_p\left(\hat{m}_{smrq}\right) = \frac{1-f'}{r}S_E^2 + \frac{1}{N^2}E_p\left(\sum_{\substack{k,l\in U_t\\ k\neq l}}\sum_{k,l\in r}\left(\frac{\phi_{kl}}{\phi_k\phi_l}-1\right)\frac{E_k^*}{\phi_k}\frac{E_l^*}{\phi_l}\,|s\right),$$

where

$$S_E^2 = \frac{1}{N-1}\sum_{k\in U}\left(y_k - y_k^o\right)^2,$$

and $E_k^* = \dfrac{E_k}{\pi_k}$. *Now* $V_p\left(\hat{m}_{smrq}\right)$ *is estimated by*

$$\hat{V}_p\left(\hat{m}_{smrq}\right) = \frac{1-f'}{n_r}s_c^2 + \frac{1}{n_r^2}\sum_{\substack{k,l\in U_t\\ k\neq l}}\sum_{k,l\in r}\frac{1}{\phi_{kl}}\left(\frac{\phi_{kl}}{\phi_k\phi_l}-1\right)\frac{e_k}{\phi_k}\frac{e_l}{\phi_l},$$

such that

$$s_c^2 = \frac{1}{n_r-1}\sum_{k\in R}\left(y_k - \hat{\mu}(x_k)\right)^2,$$

$f' = \dfrac{n_r}{N}$ *and* $e_k = y_k - \hat{\mu}(x_k)$.

Remark 5.4.2 *We observe from the theorem and the example that the variance is decomposed into the sampling variance and the response variance. If we have full response then* $\phi_k = 1$ *for all* $k \in s$ *and* $r = n$. *The last term in* $V_p\left(\hat{m}_{smrq}\right)$*then vanishes and* $V_p\left(\hat{m}_{smrq}\right) = V_p\left(\hat{m}_{sm}\right)$ *which was developed in Chapter 2 for the full sample case.*

The theorem demonstrates that $\hat{m}_{smrq}$ is approximately unbiased. Consequently we can use a normal approximation to find confidence intervals for $\bar{Y}$ using $\hat{m}_{smrq}$ and $\hat{V}_p\left(\hat{m}_{smrq}\right)$ only if $\hat{m}_{smrq}$ is asymptotic unbiasedness and consistent.

5.5 Asymptotic Unbiasedness and Consistency of the generalized smoothing estimator having a Non-response Distribution

We now demonstrate that $\hat{m}_{gmrq}$ is an asymptotic design-unbiased and design-consistent estimator of $\bar{Y}$. Consider the same sequence of populations $U_1, U_2, \ldots$ as in Issaki - Fuller (1979) and used in Chapter 2.

Definition 5.5.1 *A predictor m is said to be asymptotically pq unbiased if*

$$\lim_{t\to\infty} \left(E_{pq}\left(m|\mathbf{Y}_t\right) - \bar{Y}_t\right) = 0,$$

with probability.

Definition 5.5.2 *A predictor m is said to be $p-q$ consistent if for all $\varepsilon > 0$,*

$$\lim_{t\to\infty} P_{pq}\left(\left|m - \bar{Y}_t\right| > \varepsilon|\mathbf{Y}_t\right) = 0$$

with probability one.

The assumptions made in Chapter 2 are now adjusted so as to take into account the responding probabilities. The assumptions are

A'.1 $\lim_{t\to\infty} \frac{1}{N_t} \sum_{k\in U_t} Y_{kt}^2 < \infty$ with d probability one,

A'.2 $\lim_{t\to\infty} N_t \min_{1\le k\le N_t} \phi_{kt}\pi_{kt} = \infty$,

A'.3 $\lim_{t\to\infty} \max_{1\le k\ne l\le N_t} \left|\frac{\phi_{klt}\pi_{klt}}{\phi_{kt}\phi_{lt}\pi_{kt}\pi_{lt}} - 1\right| = 0.$

Let

$$I_{kt} = \begin{cases} 1 \forall k, \in s_t \\ 0 \text{ otherwise} \end{cases}$$

and

$$Z_{kt} = \begin{cases} 1 \forall k, \in r_t \\ 0 \text{ otherwise} \end{cases}$$

now

$$\hat{m}_{smrq} = \frac{1}{N}\sum_{k\in R}\frac{y_k}{\pi_k\phi_k} - \frac{1}{N}\left(\sum_{k\in R}\frac{\hat{\mu}_{\pi}(x_k)}{\phi_k} - \sum_{k\in U}\hat{\mu}(x_k)\right), \tag{5.18}$$

can be written as

$$\begin{aligned} M_{tq} &= \frac{1}{N_t}\sum_{k\in U_t}\frac{Y_k I_{kt} Z_{kt}}{\phi_{kt}\pi_{kt}} - \frac{1}{N_t}\sum_{k\in U_t}\hat{\mu}(x_k)\left(\frac{I_{kt}Z_{kt}}{\phi_{kt}\pi_{kt}} - 1\right) \\ &= \frac{1}{N_t}\sum_{k\in U_t}\frac{Y_k I_{kt} Z_{kt}}{\phi_{kt}\pi_{kt}} - \frac{1}{N_t}\sum_{k\in U_t}\left(\sum_{i\in s_t} w_{ki} Y_i\right)\left(\frac{I_{kt}Z_{kt}}{\phi_{kt}\pi_{kt}} - 1\right). \end{aligned} \tag{5.19}$$

Theorem 5.5.1 *Under $A^{'}.1$ - $A^{'}.3$, M_{tq} is asymptotically $p-q$ unbiased and consistent.*

The proof of the above theorem follow the same steps as that of theorem 2.5.2.

The regression models $\hat{m}_{smr}$ and $\hat{m}_{smrq}$ will both be representative of the population point scatter and both are *model and design unbiased.* But the regression model $\hat{m}_{smrq}$ will also be *q- unbiased* with respect to the response mechanism. But again we are in the situation in which the ϕ_k are unknown.. The following section will introduce a nonparametric method to estimate the unknown ϕ_k.

5.6 Nonparametric Regression Estimation of the Response Probabilities

Giommi (1985, 1987) uses nearest neighbors and kernels to estimate the response probabilities. We show explicitly that the ϕ_k of Giommi (1987) can be achieved using nonparametric regression.

Let

$$Z_k = \begin{cases} 1 \text{ if unit } k \text{ responds} \\ 0 \text{ otherwise} \end{cases}$$

be an indicator function for the respondents. Under this model the expected response $E(Z_k) = \phi_k$. The response probabilities are estimated by

$$\hat{\phi}_k = \sum_{k \in s} w_{ki} z_k,$$

where

$$w_{ki} = \frac{K\left(\frac{x_k - x_i}{b}\right)}{\sum_{i \in s} K\left(\frac{x_k - x_i}{b}\right)}.$$

Now we can rewrite $\hat{\phi}_k = \sum_{k \in s} w_{ki} z_k$ as

$$\begin{aligned}
\hat{\phi}_k &= \frac{\sum_{k \in s} K\left(\frac{x_k - x_i}{b}\right) z_k}{\sum_{i \in s} K\left(\frac{x_k - x_i}{b}\right)} \\
&= \frac{\sum_{k \in R} K\left(\frac{x_k - x_i}{b}\right) 1}{\sum_{i \in s} K\left(\frac{x_k - x_i}{b}\right)} + \frac{\sum_{k \in R^c} K\left(\frac{x_k - x_i}{b}\right) 0}{\sum_{i \in s} K\left(\frac{x_k - x_i}{b}\right)} \\
&= \frac{\sum_{k \in R} K\left(\frac{x_k - x_i}{b}\right)}{\sum_{i \in s} K\left(\frac{x_k - x_i}{b}\right)},
\end{aligned}$$

which is precisely Giommi's (1987) estimate.

Remark 5.6.1 *$\hat{\phi}_k$ is the ratio of two Kernel probability estimates (apart from a normalizing factor) over two different samples namely the response sample and the full sample.*

Lemma 5.6.1 *The* $\lim_{b\to\infty} \hat{\phi}_k = \frac{r}{n}$.

Proof

$$\begin{aligned}
\lim_{b\to\infty} \hat{\phi}_k &= \lim_{b\to\infty} \frac{\sum_{k\in R} K\left(\frac{x_k - x_i}{b}\right)}{\sum_{i\in s} K\left(\frac{x_k - x_i}{b}\right)} \\
&= \frac{\sum_{k\in R} K(0)}{\sum_{i\in s} K(0)} \\
&= \frac{r}{n}.
\end{aligned}$$

■

Remark 5.6.2 *As* $b \to \infty$ *the estimate* $\hat{\phi}_k$ *is overly smoothed.*

Simonoff (1996) considers smoothing for categorical data sets. He also considers smoothing for a multinomial regression and Poisson regression setup. We only consider smoothing in a binary regression setup which is a special case of multinomial regression.

5.6.1 *Kernel Binary Regression*

Now for kernel regression we estimate the response probabilities ϕ_k by

$$\hat{\phi}_k^K = \sum_{k\in s} w_{ki} z_k. \tag{5.20}$$

A problem with the above formulation is that the error terms ε_k are heteroscedastic because the response variable Z_k is an indicator variable. One can easily show that

$$\sigma^2(\varepsilon_k) = \phi_k(1-\phi_k).$$

A weighted nonparametric regression model will provide efficient estimates when the error variance is unequal. Using weights

$$v_k = \frac{1}{\phi_k(1-\phi_k)},$$

is the proper approach to take but there exists a difficulty with this procedure because the ϕ_k are unknown. A way out of this predicament is to estimate ϕ_k in stages.

Step 1

Fit the regression model by nonparametric regression

$$\hat{\phi}_k = \sum_{k \in s} w_{ki} z_k,$$

where the function w_{ki} *is defined by the following:*

$$w_{ki} = \frac{K\left(\frac{x_k - x_i}{b}\right)}{\sum_{i \in s} K\left(\frac{x_k - x_i}{b}\right)} \quad i = 1 \ldots n,$$

and b *is the so called 'bandwidth' parameter.*

Step 2

Estimate the weights v_k *using the results of Stage 1*

$$\hat{v}_k = \frac{1}{\hat{\phi}_k \left(1 - \hat{\phi}_k\right)}.$$

An initial $\hat{\phi}_k$ *will be needed to start the process, a method for finding this value will be developed in Chapter 6.*

Step 3

The estimated weights are then used to transform the variables x_k *and* z_k *as:*

$$\begin{aligned} x_k^* &= \frac{x_k}{\hat{v}_k}, \\ z_k^* &= \frac{z_k}{\hat{v}_k}. \end{aligned}$$

Step 4

Fit the regression model by nonparametric regression

$$\tilde{\phi}_k = \sum_{k \in s} w^*_{ki} z^*_k,$$

where the function w^*_{ki} *is defined by the following:*

$$w^*_{ki} = \frac{K\left(\frac{x^*_k - x^*_i}{b}\right)}{\sum_{i \in s} K\left(\frac{x^*_k - x^*_i}{b}\right)} i = 1 \ldots n.$$

Step 5

Steps 2 to 4 are iterated till convergence i.e.

$$\left|\left|\tilde{\phi}^{t+1}_k\right|\right| - \left|\left|\tilde{\phi}^{t}_k\right|\right| \leq \varepsilon,$$

for some specified constant ε.

The estimate $\hat{m}_{smrq}$ is modified so as to incorporate the estimated response probabilities $\bar{\phi}_k$ as

$$\hat{m}_{smr\bar{q}} = \frac{1}{N} \sum\nolimits_{k \in U} \hat{\mu}(x_k) + \frac{1}{N} \sum\nolimits_{k \in R} \frac{e_k}{\bar{\phi}_k \pi_k}, \tag{5.21}$$

and the unbiased estimator of $V_p(\hat{m}_{sm})$ is also modified as

$$\begin{aligned} \hat{V}_p(\hat{m}_{smr\bar{q}}) &= \frac{1}{N^2} \sum_k \sum_{l \in r} \frac{1}{\pi_{kl}} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1\right) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \\ &\quad + \frac{1}{N^2} \sum_k \sum_{l \in r} \frac{1}{\bar{\phi}_{kl}} \left(\frac{\bar{\phi}_{kl}}{\bar{\phi}_k \bar{\phi}_l} - 1\right) \frac{e^*_k}{\bar{\phi}_k} \frac{e^*_l}{\bar{\phi}_l}. \end{aligned} \tag{5.22}$$

5.6.2 *Spline Smoothing*

Using the same setup as before we shall now describe the ***binary spline smoother***. The goal of this procedure is to minimize the penalized residual sum of squares which is

$$\sum_{k \in s} (z_k - \mu(x_k))^2 + \lambda \int \left(\mu''(t)\right)^2 dt,$$

over all functions $\mu(x_k)$ with continuous first and integrable second derivatives. As before the parameter λ represents the rate of exchange between the residual error and the roughness of the curve $\mu(.)$ and therefore is a smoothing parameter which has the same function as the bandwidth.

As we have seen the unique solution for the problem of is a cubic spline $\hat{\mu}(x_k)$. The cubic spline has the following properties:

a. *A cubic polynomial fits the data between two successive sampled x_k values.*

b. *At the sampled values x_k, $\hat{\mu}(x_k)$ and its two first derivatives are continuous.*

c. *At the boundary points $x_{(1)}$ and $x_{(n)}$ the second derivatives of $\hat{\mu}(x_k)$.*

Therefore we estimate the response probability under this setup as:

$$\hat{\phi}_k = \hat{\mu}(x_k) \text{ for all } x_k \in s.$$

The parameter λ is chosen as previously described by the cross-validating the sum of squares criterion

$$CV(\lambda) = \frac{1}{n} \sum\nolimits_{k \in s} \left(z_k - \hat{\mu}_\lambda^{-k}(x_k)\right)^2,$$

where $\hat{\mu}_\lambda^{-k}(x_k)$ denotes the fit at x_k by leaving out that data point. with the following.

We again have the same problem of heteroscedasticity of the ε_k. As before we estimate ϕ_k in stages.

Step 1

Fit the regression model by the binary spline smoother $\hat{\varphi}_k = \hat{\mu}(x_k)$ for all $x_k \in S$.

Step 2

Estimate the weights v_k *using the results of Stage 1*

$$\hat{v}_k = \frac{1}{\hat{\varphi}_k (1 - \hat{\varphi}_k)}.$$

An initial $\hat{\varphi}_k$ *will be needed to start the process a method for finding this value will be developed in Chapter 6.*

Step 3

The estimated weights are then used to transform the variables x_k *and* z_k *as:*

$$\begin{aligned} x_k^* &= \frac{x_k}{\hat{v}_k}, \\ z_k^* &= \frac{z_k}{\hat{v}_k}. \end{aligned}$$

Step 4

Fit the binary spline smoother

$$\tilde{\varphi}_k = \hat{\mu}(x_k^*) \text{ for all } x_k \in S.$$

Step 5

Steps 2 to 4 are iterated till convergence i.e.

$$||\tilde{\varphi}_k^{t+1}|| - ||\tilde{\varphi}_k^t|| \leq \varepsilon,$$

for some specified constant ε.

To insure that

$$0 \leq \tilde{\varphi}_k \leq 1,$$

we again re-scale any $\tilde{\varphi}_k > 1$ as $\tilde{\varphi}_k = 1$.

As before the estimate $\hat{m}_{smrq}$ is modify so as to incorporate the estimated response probabilities $\tilde{\varphi}_k$ as

$$\hat{m}_{smrq} = \frac{1}{N}\sum_{k\in U}\hat{\mu}(x_k) + \frac{1}{N}\sum_{k\in r}\frac{e_k}{\tilde{\varphi}_k\pi_k}, \tag{5.23}$$

and the unbiased estimator of $V_p(\hat{m}_{sm})$ is modified as

$$\begin{aligned}\hat{V}_p(\hat{m}_{smrq}) &= \frac{1}{N^2}\sum_k\sum_{l\in r}\frac{1}{\pi_{kl}}\left(\frac{\pi_{kl}}{\pi_k\pi_l}-1\right)\frac{e_k}{\pi_k}\frac{e_l}{\pi_l} \\ &\quad + \frac{1}{N^2}\sum_k\sum_{l\in r}\frac{1}{\tilde{\phi}_{kl}}\left(\frac{\tilde{\varphi}_{kl}}{\tilde{\varphi}_k\tilde{\varphi}_l}-1\right)\frac{e_k^*}{\tilde{\varphi}_k}\frac{e_l^*}{\tilde{\varphi}_l}.\end{aligned} \tag{5.24}$$

Statisticians make parametric assumptions with respect to the responding distribution $q(r|s)$. The conclusions that will be made with this modeling effort will only be as good as the assumptions made at the parametric stage of analysis. We have shown that it is possible to estimate the responding distribution $q(r|s)$ nonparametrically. Hence the problem of model misspecification will not be a problem in our analysis.

Chapter 6

Empirical Comparisons of Generalized Regression and Generalized Smoothing Estimators under a Simple Random Sampling Design in the Presence of Nonresponse

6.1 Introduction

This chapter will compare estimates of the population mean by contrasting parametric and nonparametric methods when nonresponse occurs in a sample. The parametric methods considered are the reduced generalized regression model and the generalized regression model incorporating the estimated response probabilities $\bar{\phi}_k$. For the nonparametric methods the reducedgeneralized smoothing estimator (normal kernel and spline) and thegeneralized smoothing estimator (normal kernel and spline) incorporating the es-

timated response probabilities $\bar{\phi}_{k.}$ The sampling design that will used throughout this chapter is *simple random sampling without replacement.* In order to understand the behavior of these different estimates the populations under investigation will have known point scatters.

In section 6.2 we identify two populations from Chapter 3 that have a known point scatter. Section 6.2 also describes two response mechanisms that will be used in the analysis. Formulae developed in Chapter 5 are presented in section 6.3 when the sampling design is simple random sampling. In section 6.4 criteria are developed to judge the efficacy of the estimators. The results of the simulation procedure are presented in section 6.5 while in section 6.6 the conclusions are presented.

6.2 Populations

In order to understand the behavior of the parametric and nonparametric estimators under nonresponse two populations were chosen from the five populations in Chapter 3. The results of Chapter 3 show that under full response that the generalized regression estimate is optimal in population 1 while in population 5 either the normal kernel regression estimator or the spline regression estimator is optimal. We purposely choose these populations so as to verify the theory developed in Chapter 5. The following table summarizes the major characteristics of the two populations.

Table 6.1

Population

Summary Statistics

	Population 1		Population 5	
	$l = 1$, $g = 1.0$		$l = 0.5$, $g = 0.5$	
	$\beta = \sigma = 0.4$		$\beta = 4.0$, $\sigma = 0.4$	
	X	Y	X	Y
Mean	20.51	8.18	20.01	25.53
Standard Deviation	14.2	6.01	11.42	1.39
Skewness	1.27	1.39	0.15	.036
Kurtosis	2.20	2.91	2.48	2.60
Coefficient of Variation	0.69	0.74	0.14	0.05
Correlation	0.957		0.579	

In the analysis two nonresponse mechanisms are considered. The mechanisms are defined by independent Bernoulli (ϕ_k) trails such that the probability of nonresponse is ϕ_k for unit k and are given by:

Mechanism I: ϕ is constant and independent for all units in the population.

Mechanism II: $\phi_k = \exp(-\lambda x_k)$ is this case ϕ_k is a decreasing function of x_k. Rubin (1977) defined this type of nonresponse mechanisms '*ignorable*' while Rancourt, Lee and Särndal (1994) called this '*the unconfounded mechanism*'. Under this mechanism small units x_k respond more frequently than large units.

In the simulation study the response probabilities were set at the 70% and 90% level. For mechanism II a $\lambda = .019$ provides a mean response level of 70% while a $\lambda = .005$ provides a mean response level of 90%.

6.3 The Estimators

So as to have an unbiased estimate of the population mean in the presence of nonresponse, the statistician must model correctly either the regression function or the response probabilities. If the regression model is representative of the population point scatter, then the estimator is *model and design unbiased* whether the estimated response probabilities are correct or not. On the other hand if the regression model is misspecified then the response probabilities have to be correctly estimated in order that the estimate be *design unbiased (p unbiased).*

6.3.1 Generalized Regression Estimates

The first step in the analysis is to model the sample point scatter by a regression curve. We again consider a model ξ in which the point scatter $\left(x_k, \frac{y_k}{x_k}\right)$ is constant such that

$$E_\xi(y_k) = \beta x_k$$

with variance structure proportional to the x_k around the regression line

$$V_\xi(y_k) = \sigma^2 x_k.$$

The generalized regression estimator for the population mean under a simple random sampling design has the following form:

$$\hat{m}_{GRE} = \bar{X}_U \frac{\bar{y}}{\bar{x}}. \tag{6.1}$$

Now $\bar{X}_U$ is the population mean of the known x_k's while $\bar{y}$ and $\bar{x}$ are the sample means found with the n sampled units. Suppose that only r units from a sample of size n respond and let R represent the responding set, the above equation is now written as

$$\hat{m}_{yr} = \bar{X}_U \frac{\bar{y}_r}{\bar{x}_r} \tag{6.2}$$

where $\bar{y}_r$ and $\bar{x}_r$ are the sample means of the r respondents. We shall call $\hat{m}_{yr}$ the reduced generalized regression estimator. The estimate $\hat{m}_{yr}$ has as a variance estimator

$$\hat{V}_{srs}(\hat{m}_{yr}) = \left(\frac{\bar{X}_U}{\bar{x}_r}\right)^2 \hat{V}_0 \tag{6.3}$$

such that

$$\hat{V}_0 = \frac{1-f}{r} \frac{\sum_{k\in R}\left(y_k - \hat{\beta}x_k\right)^2}{r-1}$$

with $\hat{\beta} = \dfrac{\sum_{k\in R} y_k}{\sum_{k\in R} x_k}$.

6.3.2 Generalized Smoothing Estimators

Thegeneralized smoothing estimator

$$\hat{m}_{sm} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{N}\sum_{k\in s}\frac{e_k}{\pi_k},$$

was shown in Chapter 2 to be asymtotically design unbiased (*p unbiased)* and design-consistent for $\bar{Y}$. The above model is modified to

$$\hat{m}_{smr} = \frac{1}{N}\sum_{k\in U}\hat{y}_k + \frac{1}{N}\sum_{k\in R}\frac{e_k}{\pi_k}. \tag{6.4}$$

and will be named the reducedgeneralized smoothing estimator. The estimates $\hat{y}_k$ can be either a normal kernel or a spline estimates of y_k based on a responding set R instead of a sample set s.

6.3.3 Regression and Response Models

The response probabilities are now incorporated in the regression models. These will be estimated nonparametrically by binary regression. Initial response probabilities are needed before we can apply the algorithms developed in Chapter 5.

Since the vector of auxiliary variable x is know for the sample set s and the response set R the following steps will estimate the initial response probability ϕ_k :

1. Find the number of bins c for the sample set s with Doane's rule:

$$c = \left[\log_2 n + 1 + \log_2\left(1 + \hat{\gamma}\sqrt{\frac{n}{6}}\right)\right]$$

where n is the sample size of the full sample, $\hat{\gamma}$ is an estimate of the kurtosis of the sample values and [.] is the greatest integer function.

2. Calculate the bin width ϖ for the sampled set s:

$$\varpi = \frac{Range\,(x)}{c}.$$

3. Form the following intervals:

$$(m + (k-1)\varpi, m + k * \varpi] \text{ for } k = 1...c.$$

such that $m = \min(x \in s)$.

4. Count the number of x's that came from the sample set s that fall in the interval $(m + (k-1)\varpi, m + k * \varpi]$ suppose this value is n_k.

5. Counted the number of x's that came from the response set R that fall in the interval $(m + (k-1)\varpi, m + k * \varpi]$ suppose this value is r_k.

6. An initial estimate of the of the response probability for the interval $(m + (k-1)\varpi, m + k * \varpi]$ is given by

$$\hat{\theta}_k = \frac{r_k}{n_k}.$$

7. Transform $\hat{\theta}_k$ to

$$\ddot{\theta}_k = \log\left(\frac{\hat{\theta}_k}{1 - \hat{\theta}_k}\right)$$

for the subsequent analysis.

Remark 6.3.1 *All responding values have the same response probability $\hat{\theta}_k$ in the inter-*

val $(m + (k-1)\varpi, m + k * \varpi]$. ***The idea is similar to the response homogeneity group model but different in the sense that we stratify with the observed values from the full sample.***

The vectors $\ddot{\theta}$ and $x's$ do not have the same length. In order to overcome this difficulty a vector $\dot{\theta}$ is created such that $\ddot{\theta}_k$ repeats itself r_k times. We then perform a binary kernel regression using a normal kernel and a smooth spline regression as described in Chapter 5. In the binary regression the dependent variable is taken as the response vector $\dot{\theta}$ and as the explanatory variable the $x's$ from the sample set s.

Suppose that the estimated response vector $\bar{\theta}$ is found through either binary kernel regression or a smooth spline regression. The vector $\bar{\theta}$ is then transformed to

$$\hat{\phi} = \frac{\exp(\bar{\theta})}{1 + \exp(\bar{\theta})}$$

such that unit $k \in R$ has as a response probability $\hat{\phi}_k$. These response probabilities are now used in the estimation procedures developed in Chapter 5 section 6. Note that $\hat{\phi}_k$ can be estimated with either binary kernel regression or binary spine regression.

6.3.4 Generalized Regression Estimates with Estimated Response Probabilities

Assuming the model ξ, and suppose that only r units from a sample of size n respond. Let R represent the responding set, then under a simple random design the generalized regression estimator for the population mean has the following form when the response probabilities are estimated by binary kernel regression or binary spline regression:

$$\tilde{m}^{A}_{yrq} = \bar{X}_U \frac{\sum_{k \in R} \frac{y_k}{\hat{\phi}_k}}{\sum_{k \in R} \frac{x_k}{\hat{\phi}_k}}. \tag{6.5}$$

Now the estimated variance of $\tilde{m}^{A}_{yrq}$ is

$$\hat{V}_{srs}\left(\tilde{m}^{A}_{yrq}\right) = \left(\frac{\bar{X}_U}{\bar{x}_r}\right)^2 \hat{V}_0 + \frac{1}{r^2}\sum_{\substack{k,l\in U_t \\ k\neq l}}\sum\nolimits_{k,l\in R}\frac{1}{\hat{\phi}_{kl}}\left(\frac{\hat{\phi}_{kl}}{\hat{\phi}_k\hat{\phi}_l}-1\right)\frac{e_k}{\hat{\phi}_k}\frac{e_l}{\hat{\phi}_l} \tag{6.6}$$

such that

$$\hat{V}_0 = \frac{1-f}{r}\frac{\sum_{k\in R}\left(y_k-\hat{\beta}x_k\right)^2}{r-1}$$

and $e_k = y_k - \hat{y}_k$.

The ϕ_k can be estimated with a binary normal kernel or a binary spline regression. Therefore we let $\tilde{m}^{A}_{yrq} = \tilde{m}^{K}_{yrq}$ if the ϕ_k is estimated with a binary normal kernel regression and $\tilde{m}^{A}_{yrq} = \tilde{m}^{S}_{yrq}$ if the ϕ_k is estimated with a binary spline regression

6.3.5 Generalized Smoothing Estimators with Estimated Response Probabilities

The smooth estimates discussed in Chapter 5 are now modified so as to incorporate the binary kernel regression estimate $\hat{\phi}_k$. Under simple random sampling the smoothing estimate $\hat{m}_{smrq}$ has the following form:

$$\hat{m}_{smrq} = \frac{1}{N}\sum\nolimits_{k\in U}\hat{y}_k + \frac{1}{n}\sum\nolimits_{k\in R}\frac{e_k}{\phi_k}$$

where $\hat{y}_k$ is either a kernel smoother or a spline smoother. Therefore using $\hat{\phi}_k$ as an estimate of ϕ_k the above equation can be written as:

$$\hat{m}^{A}_{smrq} = \frac{1}{N}\sum\nolimits_{k\in U}\hat{y}_k + \frac{1}{r}\sum\nolimits_{k\in R}\frac{e_k}{\hat{\phi}_k} \tag{6.7}$$

Also an unbiased estimate of the variance of $\hat{m}^{K}_{smrq}$ is given by

$$\hat{V}_{srs}\left(\hat{m}^{A}_{smrq}\right) = \frac{1-f'}{r}s_e^2 + \frac{1}{r^2}\sum_{\substack{k,l\in U_t \\ k\neq l}}\sum\nolimits_{k,l\in r}\frac{1}{\hat{\phi}_{kl}}\left(\frac{\hat{\phi}_{kl}}{\hat{\phi}_k\hat{\phi}_l}-1\right)\frac{e_k}{\hat{\phi}_k}\frac{e_l}{\hat{\phi}_l} \tag{6.8}$$

The $\hat{y}_k$ is either a kernel smoother or a spline smoother and the ϕ_k can be estimated with a binary normal kernel or a binary spline regression. We let $\tilde{m}^A_{yrq} = \tilde{m}^{KK}_{yrq}$ if $\hat{y}_k$ is kernel smoother and the ϕ_k is estimated with a binary normal kernel regression, and $\tilde{m}^A_{yrq} = \tilde{m}^{KS}_{yrq}$ if $\hat{y}_k$ is kernel smoother and the ϕ_k is estimated with a binary binary spline regression. On the other hand we let $\tilde{m}^A_{yrq} = \tilde{m}^{SK}_{yrq}$ if $\hat{y}_k$ is a spline smoother the ϕ_k is estimated with binary normal kernel regression and finally $\tilde{m}^A_{yrq} = \tilde{m}^{SS}_{yrq}$ if $\hat{y}_k$ is a spline smoother the ϕ_k is estimated binary spline regression.

6.4 Simulation Procedure

In *mechanism I* the response rate is constant for all members of the population while for *mechanism II* the response rate ia different for all members of the population. Therefore the sampling mechanism needed to generate these samples are different. The simulation algorithm for both mechanisms are now presented along with the summary statistics calculated for each sample drawn. The analysis was only done for the simple random sampling design.

6.4.1 Mechanism I

Simulations were carried out in S-Plus Version 3.3. The simulation procedure for *mechanism I* is as follows:

1. A simple random sample of size $n = 25$ is drawn from Population 1 and Population 5 by random number generation.

2. From this random sample of size n a sample of size $r = P \times n$ is drawn by random number generation where P is set to be 70% and 90%.

3. Repeat steps 2 and 3, 1000 times.

When the sample is of size $n = 25$ the response rate is 100% the estimate $\hat{m}_{GRE}$ was calculated. For each sample of size r the following estimates of the finite population mean were calculated: $\hat{m}_{yr}$, $\hat{m}_{smr}$,and $\hat{m}_{spr}$. Thirdly for each sample the measures of precision,

relativity, efficiency and accuracy as defined in Chapter 3 were calculated so as to judge the behavior of $\hat{m}_{yr}$, $\hat{m}_{smr}$,and $\hat{m}_{spr}$.

6.4.2 Mechanism II

The simulation procedure for *mechanism II* is as follows:

1. A simple random sample of size $n = 25$ is drawn from Population 1 and Population 5 by random number generation.

2. For each unit k in the sample n the response probability is calculated with $\phi_k = \exp(-\lambda x_k)$, such that λ is set to 0.019 for the 70% mean response rate and to 0.005 for the 90% mean response rate.

3. *A Bernoulli trail is then performed for all $k \in s$ with probability ϕ_k for success (response) and $1 - \phi_k$ for failure (nonresponse).*

3. Repeat steps 2 and 3, 1000 times.

When the sample is of size $n = 25$ the response rate is 100% the estimate $\hat{m}_{GRE}$ was calculated. Since we know a priori the response probabilities ϕ_k for each sample of size r the estimate

$$\hat{m}_{yrq} = \bar{X}_U \frac{\sum_{k \in R} \frac{y_k}{\phi_k}}{\sum_{k \in R} \frac{x_k}{\phi_k}} \tag{6.9}$$

is used as a benchmark for the analysis. The analysis for samples of size r was made *with* and *without* the estimated response probabilities $\hat{\phi}_k$. For the analysis *with* the estimated response probabilities $\hat{\phi}_k$, the following six estimates of the finite population mean were found: $\hat{m}^K_{yrq}$, $\hat{m}^S_{yrq}$, $\hat{m}^{KK}_{smrq}$, $\hat{m}^{KS}_{smrq}$, $\hat{m}^{SK}_{smrq}$, and $\hat{m}^{SS}_{smrq}$. Now for the analysis *without* the estimated response probabilities $\hat{\phi}_k$ the following three estimates of the finite population mean were considered: $\hat{m}_{yr}$, $\hat{m}_{smr}$,and $\hat{m}_{spr}$. Finally for each sample the measures of precision, relativity, and accuracy as defined in Chapter 3 were calculated for each estimator.

The following measures of efficiency were computed for all the proposed estimators:

1. *The first measure of efficiency used as a base the generalized regression estimate:*

$$Eff_1\left(\hat{m}_{(.)}\right) = \frac{\bar{V}_{srs}\left(\hat{m}_{GRE}\right)}{\bar{V}_{srs}\left(\hat{m}_{(.)}\right)}. \tag{6.10}$$

2. *The second measure of efficiency used as a base $\hat{m}_{yrq}$ the generalized regression estimator adjusted with the known response probabilities ϕ_k*

$$Eff_2\left(\hat{m}_{(.)}\right) = \frac{\bar{V}_{srs}\left(\hat{m}_{yrq}\right)}{\bar{V}_{srs}\left(\hat{m}_{(.)}\right)}. \tag{6.11}$$

3. *The third and final measure of efficiency used as a base $\hat{m}_{yr}$ the generalized regression estimator reduced to size r*

$$Eff_3\left(\hat{m}_{(.)}\right) = \frac{\bar{V}_{srs}\left(\hat{m}_{yr}\right)}{\bar{V}_{srs}\left(\hat{m}_{(.)}\right)}. \tag{6.12}$$

Values of $Eff_{(.)}\left(\hat{m}_{(.)}\right) > 1$ imply that the proposed method of estimation is superior to that of the base estimator $\hat{m}_{GRE}$, $\hat{m}_{yrq}$ and $\hat{m}_{yr}$. If $Eff_{(.)}\left(\hat{m}_{(.)}\right) < 1$ this implies that the proposed method of estimation is inferior to the base estimator.

6.5 Simulation Results

The following pages contain an analysis of the Monte Carlo simulations. The simulations were performed on each population at and for both response mechanisms. The response rates were set at 70% and 90% for each mechanism.

6.5.1 Mechanism I

The following tables summarizes the Monte Carlo characteristics of the two populations considered for a sample size of $n = 25$ and mean response rates (R.R.) of 70% and 90% when response is at random. Table 6.2 contains the value of the population parameter $\bar{Y}$ and the average estimates of this parameter found by using $\bar{m}_{GRE}$, $\bar{m}_{yr}$, $\bar{m}_{sm}$ and $\bar{m}_{sp}$.

Table 6.3 contains the average bias of the estimators, while Table 6.4 presents the average absolute relative bias of the estimators.

Table 6.2

Average Estimate of Population Mean

	Population 1		**Population 2**	
Estimates	$\bar{Y} = 8.18$		$\bar{Y} = 25.53$	
$\bar{m}_{GRE}$	8.171	8.179	25.551	25.543
	R. R. = 70%	**R. R. = 90%**	**R. R. = 70%**	**R. R. = 90%**
$\bar{m}_{yr}$	8.190	8.189	25.575	25.522
$\bar{m}_{smr}$	7.991	7.894	25.520	25.496
$\bar{m}_{spr}$	8.165	8.196	25.541	25.531

Table 6.3

Average Bias of the Estimators

$\bar{B}(\hat{m}_{(.)})$

Estimates	**Population 1**		**Population 2**	
$\bar{m}_{GRE}$	−0.007	0.001	0.020	0.012
	R. R. = 70%	**R. R. = 90%**	**R. R. = 70%**	**R. R. = 90%**
$\bar{m}_{yr}$	0.011	0.012	0.043	−0.010
$\bar{m}_{smr}$	−0.284	−0.187	−0.036	−0.012
$\bar{m}_{spr}$	−0.018	0.012	0.009	−0.001

Table 6.4

Average Absolute Relative Bias of the Estimators

$\overline{RB}(\hat{m}_{(.)})$

Estimates	Population 1		Population 2	
$\bar{m}_{GRE}$	0.1%	0.0%	0.1%	0.0%
	R. R. = 70%	**R. R. = 90%**	**R. R. = 70%**	**R. R. = 90%**
$\bar{m}_{yr}$	0.1%	0.1%	0.2%	0.0%
$\bar{m}_{smr}$	3.5%	2.3%	0.1%	0.0%
$\bar{m}_{spr}$	0.2%	0.1%	0.0%	0.0%

The following patterns emerge from these tables:

1. For the estimators $\bar{m}_{yr}$, $\bar{m}_{smr}$ and $\bar{m}_{spr}$ the bias $\bar{B}(\hat{m}_{(.)})$ and relative bias $\overline{RB}(\hat{m}_{(.)})$ decreases as the response rate increases for the two populations.

2. In Chapter 3 it was shown that the finite general regression estimator $\bar{m}_{GRE}$ had the smallest bias in Population 1 because for this population this method of estimation is optimal. Therefore the finite reduced general regression estimator $\bar{m}_{yr}$ will also have the smallest bias in this population because it shares the same optimal properties.

3. In populations 5 for both response rates the bias $\bar{B}(\hat{m}_{(.)})$ and the absolute relative bias $\overline{RB}(\hat{m}_{(.)})$ of the reduced spline smoothing estimator is smaller than $\bar{m}_{yr}$ and $\bar{m}_{spr}$.

Table 6.5, 6.6 and 6.7 contain the average sample variance $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$, the average bias ratio $\overline{BR}\left(\hat{m}_{(.)}\right)$and finally the coverage ratio $C.R.\left(\hat{m}_{(.)}\right)$. When $\left|\overline{BR}\left(\hat{m}_{(.)}\right)\right| \leq 10\%$ the bias effect may be ignored because the coverage ratio is approximately equal to the nominal 95% coverage probability.

Table 6.5

Average Sample Variance

$\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$

Estimates	Population 1		Population 2	
$\bar{m}_{GRE}$	0.117	0.122	0.365	0.359
	R. R. = 70%	**R. R. = 90%**	**R. R. = 70%**	**R. R. = 90%**
$\bar{m}_{yr}$	0.305	0.167	0.916	0.507
$\bar{m}_{smr}$	0.775	0.413	.0123	0.052
$\bar{m}_{spr}$	0.568	0.220	0.208	0.088

Table 6.6

Average Absolute Bias Ratio

$\overline{BR}\left(\hat{m}_{(.)}\right)$

Estimates	Population 1		Population 2	
$\bar{m}_{GRE}$	0.2%	0.1%	3.3%	2.0%
	R. R. = 70%	**R. R. = 90%**	**R. R. = 70%**	**R. R. = 90%**
$\bar{m}_{yr}$	0.1%	0.1%	6.0%	3.0%
$\bar{m}_{smr}$	3.5%	2.5%	3.4%	1.2%
$\bar{m}_{spr}$	0.2%	0.1%	3.0%	0.2%

Table 6.7

Coverage Ratio

$C.R.\left(\hat{m}_{(.)}\right)$

Estimates	Population 1		Population 2	
$\bar{m}_{GRE}$	92.3%	94.5%	93.5%	93.9%
	R. R. = 70%	**R. R. = 90%**	**R. R. = 70%**	**R. R. = 90%**
$\bar{m}_{yr}$	90.0%	94.0%	90.0%	92.8%
$\bar{m}_{smr}$	94.2%	98.7%	90.0%	91.9%
$\bar{m}_{spr}$	89.6%	91.5%	91.7%	93.0%

The following conclusions can be inferred form these tables:

4. As the response rate increases the sample variance $\hat{V}_{srs}\left(\hat{m}_{(.)}\right)$and the average bias ratio $\overline{BR}\left(\hat{m}_{(.)}\right)$decreases for all the estimators $\bar{m}_{yr}$, $\bar{m}_{smr}$ and $\bar{m}_{spr}$.

5. The finite reduced general regression estimator $\bar{m}_{yr}$ has the smallest sample variance and bias ratio in population 1 but in population 5 the other two estimators have these properties.

6. The response rate has an effect on the coverage rate $C.R.(\hat{m}_{(.)})$. As the response rate increases the coverage rate $C.R.(\hat{m}_{(.)})$ approaches the nominal 95%, which confirms the asymptotic theory developed in section 2.9 of Chapter 2.

7. Table 6.6 demonstrates the effect that the bias can have on the coverage ratio Table 6.8. As the bias ratio decreases the coverage ratio approaches the 95% nominal level. Therefore the coverage rate $C.R.(\hat{m}_{(.)})$ is affected by the response rate and the bias ratio.

Tables 6.8 and 6.9 contain the mean squared error of simulation $\overline{mse}_{sim}\left(\hat{m}_{(.)}\right)$and the associated sample variance $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$. These estimators are consistent estimates of the design mean squared error $mse_{srs}\left(\hat{m}_{(.)}\right)$ and of the design variance $V_{srs}\left(\hat{m}_{(.)}\right)$.

Table 6.8

Mean Squared Error of Simulation

$\overline{mse}_{sim}\left(\hat{m}_{(.)}\right)$

Estimates	Population 1		Population 2	
$\hat{m}_{GRE}$	0.120	0.122	0.375	0.372
	R.R. = 70%	**R.R. = 90%**	**R.R. = 70%**	**R.R. = 90%**
$\hat{m}_{yr}$	0.334	0.159	0.816	0.527
$\hat{m}_{smr}$	0.760	0.314	0.135	0.082
$\hat{m}_{spr}$	0.756	0.264	0.280	0.106

Table 6.9

Variance of Simulation

$\overline{V}_{sim}(\hat{m}_{(.)})$

Estimates	Population 1		Population 2	
$\hat{m}_{GRE}$	0.120	0.122	0.372	0.375
	R.R. = 70%	**R.R. = 90%**	**R.R. = 70%**	**R.R. = 90%**
$\hat{m}_{yr}$	0.334	0.158	0.816	0.527
$\hat{m}_{smr}$	0.680	0.279	0.133	0.082
$\hat{m}_{spr}$	0.756	0.264	0.280	0.106

From Tables 6.8 and 6.9 the following can be inferred:

8. For all estimators as the response rate increases the mean squared error of simulation $\overline{mse}_{sim}(\hat{m}_{(.)})$and the sample variance $\overline{V}_{sim}(\hat{m}_{(.)})$ decreases.

9. For population 1 the $\overline{mse}_{sim}(\hat{m}_{yr})$ and $\overline{V}_{sim}(\hat{m}_{yr})$ is smaller than that of the other two estimators because $\hat{m}_{yr}$ is optimal.

10. In population 5 and for both response rates the $\overline{mse}_{sim}(\hat{m}_{yr})$ and $\overline{V}_{sim}(\hat{m}_{yr})$ is much larger than the other two estimators.

Table 6.10 demonstrates the efficiency of each estimator considered in the simulation for each of the populations studied.

Table 6.10

Efficiency

$Eff(\hat{m}_{(.)})$

Estimates	Population 1		Population 2	
$\hat{m}_{GRE}$	1.00	1.00	1.00	1.00
	R.R. = 70%	**R.R. = 90%**	**R.R. = 70%**	**R.R. = 90%**
$\hat{m}_{yr}$	0.36	0.77	0.46	0.71
$\hat{m}_{smr}$	0.18	0.44	2.81	4.56
$\hat{m}_{spr}$	0.16	0.46	1.34	3.52

The following can be inferred from Table 6.10:

11. The response rate has an effect on the efficiency of the estimator. As the response rate increases the efficiency increases.

12. The estimator $\hat{m}_{GRE}$ is used as a benchmark for both populations. In population 1 the optimal estimator is the generalized regression, because of this optimally the efficiency of the reduced generalized regression estimator $\hat{m}_{yr}$, is greater than the reduced smoothing kernel estimator or reduced smoothing spline estimator. In population 5 the generalized regression estimator is not optimal (as per Chapter 3), but at a response rate of 70% the smoothing kernel estimator or reduced smoothing spline estimator have greater efficiency than the $\hat{m}_{GRE}$ that has a response rate of 100%.

Finally Table 6.11 shows the relative accuracy of the variance estimator $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$ with respect to the simulation variance $V_{sim}\left(\hat{m}_{(.)}\right)$.

Table 6.11

Relative Accuracy

$RA(\hat{m}_{(.)})$

Estimates	Population 1		Population 2	
$\hat{m}_{GRE}$	0.96	1.02	0.96	0.98
	R.R. $= 70\%$	**R.R.** $= 90\%$	**R.R.** $= 70\%$	**R.R.** $= 90\%$
$\hat{m}_{yr}$	0.91	1.05	1.12	0.96
$\hat{m}_{smr}$	1.14	1.48	0.64	0.92
$\hat{m}_{spr}$	0.75	0.83	0.74	0.83

The above table demonstrates that:

13. As the response rate increases the relative accuracy of the sampled variance estimator $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$ approaches the sample variance $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$.

14. The relative accuracy was very good for the reduced generalized regression estimator. The other two estimators did not have the same accuracy as $\hat{m}_{yr}$.

6.5.2 Mechanism II

The following tables summarizes the Monte Carlo characteristics of the two populations considered for a sample size of $n = 25$ and mean response rates (R.R.) of 70% and 90% when response is a function of an auxiliary variable. Table 6.12 contains the value of the population parameter $\bar{Y}$ and the average estimates of this parameter. The $\bar{m}_{GRE}$ was calculated before nonresponse and used as a benchmark. A priori the response probabilities ϕ_k are known for each sample of size r the estimate $\bar{m}_{yrq}$ was also used as second benchmark in the simulation study. For the analysis *with* the estimated response probabilities $\hat{\phi}_k$, the following six estimates of the finite population mean were found: $\bar{m}^K_{yrq}$, $\bar{m}^S_{yrq}$, $\bar{m}^{KK}_{smrq}$, $\bar{m}^{KS}_{smrq}$, $\bar{m}^{SK}_{smrq}$, and $\bar{m}^{SS}_{smrq}$. And for the analysis *without* the estimated response probabilities $\hat{\phi}_k$ the following three estimates of the finite population mean were considered: $\bar{m}_{yr}$, $\bar{m}_{smr}$,and $\bar{m}_{spr}$. Table 6.13 contains the average bias of the estimators, while Table 14 presents the average absolute relative bias of the estimators.

Table 6.12

Average Estimate of Population Mean

Estimate	Population 1		Population 5	
	$\bar{Y} = 8.18$		$\bar{Y} = 25.53$	
$\bar{m}_{GRE}$	8.199	8.198	25.563	25.511
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	8.159	8.174	25.549	25.573
$\bar{m}^{K}_{yrq}$	8.160	8.174	25.548	25.574
$\bar{m}^{S}_{yrq}$	8.156	8.173	25.549	25.574
$\bar{m}^{KK}_{smrq}$	7.970	8.002	24.605	25.110
$\bar{m}^{KS}_{smrq}$	8.040	8.076	25.098	25.098
$\bar{m}^{SK}_{smrq}$	8.194	8.264	25.515	25.453
$\bar{m}^{SS}_{smrq}$	8.203	8.263	25.515	25.454
$\bar{m}_{yr}$	8.167	8.176	25.547	25.573
$\bar{m}_{smr}$	7.674	7.862	25.501	25.506
$\bar{m}_{spr}$	8.203	8.157	25.516	25.523

Table 6.13

Average Bias of the Estimators

$\bar{B}(\hat{m}_{(.)})$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	0.021	0.020	0.032	−0.021
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	−0.019	−0.004	0.042	0.018
$\bar{m}^{K}_{yrq}$	−0.019	−0.004	0.042	0.016
$\bar{m}^{S}_{yrq}$	−0.023	−0.005	0.043	0.017
$\bar{m}^{KK}_{smrq}$	−0.208	−0.176	−0.926	−0.422
$\bar{m}^{KS}_{smrq}$	−0.138	−0.102	−0.834	−0.434
$\bar{m}^{SK}_{smrq}$	0.015	0.086	−0.079	−0.017
$\bar{m}^{SS}_{smrq}$	0.025	0.085	−0.078	−0.016
$\bar{m}_{yr}$	−0.011	−0.002	0.041	0.016
$\bar{m}_{smr}$	−0.504	−0.316	−0.031	−0.025
$\bar{m}_{spr}$	0.025	−0.022	−0.015	−0.009

Table 6.14

Average Absolute Relative Bias of the Estimators

$\overline{RB}(\hat{m}_{(.)})$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	0.3%	0.2%	0.1%	0.1%
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	0.2%	0.1%	0.2%	0.1%
$\bar{m}^{K}_{yrq}$	0.2%	0.0%	0.2%	0.1%
$\bar{m}^{S}_{yrq}$	0.3%	0.1%	0.2%	0.1%
$\bar{m}^{KK}_{smrq}$	2.5%	2.1%	3.7%	1.7%
$\bar{m}^{KS}_{smrq}$	1.7%	1.2%	3.3%	1.7%
$\bar{m}^{SK}_{smrq}$	0.2%	1.0%	0.3%	0.1%
$\bar{m}^{SS}_{smrq}$	0.3%	1.0%	0.3%	0.1%
$\bar{m}_{yr}$	0.1%	0.0%	0.2%	0.2%
$\bar{m}_{smr}$	6.2%	3.9%	0.1%	0.1%
$\bar{m}_{spr}$	0.3%	0.3%	0.1%	0.0%

The following patterns are observed from these tables:

1. The bias $\bar{B}(\hat{m}_{(.)})$ and relative bias $\overline{RB}(\hat{m}_{(.)})$ decreased for all the estimators as the response rate increases for both populations.

2. Population 1 has a point scatter that is well represented by a generalized regression model therefore it is not surprising that the estimate $\bar{m}_{yr}$ had the smallest bias and bias ratio at both response rates. The adjusted estimators $\bar{m}^{K}_{yrq}$ and $\bar{m}^{S}_{yrq}$ have biases and relative biases close to $\bar{m}_{yr}$.

3. In Population 5 the smallest bias and bias ratio was observed for the reduced spline estimate $\bar{m}_{spr}$. The reduced kernel smoother $\bar{m}_{smr}$, the adjusted generalized regression estimates $\bar{m}^{K}_{yrq}$, $\bar{m}^{S}_{yrq}$, and the adjusted spline smoothers $\bar{m}^{KS}_{smrq}$ and $\bar{m}^{SS}_{smrq}$ had biases and relative biases close to $\bar{m}_{spr}$.

4. The adjusted smooth kernels $\bar{m}^{KK}_{smrq}$, $\bar{m}^{KS}_{smrq}$ have the largest bias and bias ratio for both response rates and in each population.

Tables 6.15, 6.16 and 6.17 contain the average sample variance $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$,the average bias ratio $\overline{BR}\left(\hat{m}_{(.)}\right)$ and the coverage ratio $C.R.\left(\hat{m}_{(.)}\right)$.

Table 6.15

Average Sample Variance

$\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	0.121	0.121	0.362	0.358
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	0.220	0.163	0.563	0.405
$\bar{m}^{K}_{yrq}$	0.219	0.162	0.554	0.409
$\bar{m}^{S}_{yrq}$	0.219	0.163	0.584	0.412
$\bar{m}^{KK}_{smrq}$	1.537	0.719	0.349	0.100
$\bar{m}^{KS}_{smrq}$	1.548	0.648	0.315	0.153
$\bar{m}^{SK}_{smrq}$	0.212	0.133	0.237	0.102
$\bar{m}^{SS}_{smrq}$	0.361	0.175	0.238	0.103
$\bar{m}_{yr}$	0.219	0.131	0.538	0.404
$\bar{m}_{smr}$	1.030	0.753	0.061	0.070
$\bar{m}_{spr}$	0.208	0.178	0.097	0.067

Table 6.16

Average Absolute Bias Ratio

$\overline{BR}(\bar{m}_{(.)})$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	6.0%	5.7%	5.3%	3.5%
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	4.1%	1.1%	6.5%	2.2%
$\bar{m}^{K}_{yrq}$	4.0%	1.0%	6.6%	2.2%
$\bar{m}^{S}_{yrq}$	4.8%	1.2%	6.6%	2.2%
$\bar{m}^{KK}_{smrq}$	24.5%	14.2%	14.5%	1.3%
$\bar{m}^{KS}_{smrq}$	17.2%	8.2%	15.7%	1.6%
$\bar{m}^{SK}_{smrq}$	23.5%	3.4%	24.8%	3.5%
$\bar{m}^{SS}_{smrq}$	20.3%	4.1%	24.3%	3.3%
$\bar{m}_{yr}$	2.4%	0.6%	6.5%	2.1%
$\bar{m}_{smr}$	58.1%	31.1%	29.5%	9.5%
$\bar{m}_{spr}$	5.4%	5.1%	4.9%	3.5%

Table 6.17

Coverage Ratio

$C.R.\left(\bar{m}_{(.)}\right)$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	95.0%	93.6%	94.2%	94.0%
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	91.0%	93.4%	91.0%	92.8%
$\bar{m}^{K}_{yrq}$	90.8%	93.4%	92.2%	93.2%
$\bar{m}^{S}_{yrq}$	90.8%	93.0%	90.2%	93.2%
$\bar{m}^{KK}_{smrq}$	60.6%	62.6%	57.0%	66.2%
$\bar{m}^{KS}_{smrq}$	69.0%	67.0%	62.0%	76.0%
$\bar{m}^{SK}_{smrq}$	84.0%	77.0%	93.8%	92.0%
$\bar{m}^{SS}_{smrq}$	85.0%	80.0%	93.8%	91.0%
$\bar{m}_{yr}$	91.0%	93.8%	92.4%	92.4%
$\bar{m}_{smr}$	93.0%	97.4%	93.2%	93.6%
$\bar{m}_{spr}$	84.2%	86.8%	93.8%	93.6%

The following can be inferred form these tables:

5. As the response rate increases the sample variance $\tilde{V}_{srs}\left(\bar{m}_{(.)}\right)$and the average bias ratio $\overline{BR}\left(\bar{m}_{(.)}\right)$decreases for all the estimators.

6. The finite adjusted general regression estimators $\bar{m}^{K}_{yrq}, \bar{m}^{S}_{yrq}$, the adjusted spline $\bar{m}^{SK}_{smrq}$, the reduced spline estimate $\bar{m}_{spr}$ and the reduced general regression estimator $\bar{m}_{yr}$ all have the same sample variance and bias ratio at the 70% response rate in population 1. At the 90% response rate the adjusted spline $\bar{m}^{SK}_{smrq}$ and the reduced general regression estimator $\bar{m}_{yr}$ have the smallest sample variance and bias ratio.

7. In population 5 the reduced smoothing estimator $\bar{m}_{smr}$ and the reduced spline estimate $\bar{m}_{spr}$ have essentially the same variance and bias ratio. The other estimators have larger sample variances and bias ratios.

8. The response rate has an effect on the coverage rate $C.R.(\hat{m}_{(.)})$. As the response rate increases the coverage rate $C.R.(\hat{m}_{(.)})$ approaches the nominal 95%, which again confirms the asymptotic theory developed in section 2.9 of Chapter 2.

9. Table 6.16 taken in conjunction with table 6.17 demonstrates the effect that the bias can have on the coverage ratio. As the bias ratio decreases the coverage ratio approaches the 95% nominal level. The estimators $\bar{m}^{KK}_{smrq}, \bar{m}^{KS}_{smrq}$ have less than favorably coverage rates in population1 and 5 because these estimates have large biases.

Tables 6.18 and 6.19 contain the mean squared error of simulation $\overline{mse}_{sim}(\hat{m}_{(.)})$and the associated sample variance $\overline{V}_{sim}(\hat{m}_{(.)})$of the estimators. These estimators are consistent estimates of the design mean squared error $mse_{srs}(\hat{m}_{(.)})$ and of the design variance $V_{srs}(\hat{m}_{(.)})$.

Table 6.18

Mean Squared Error of Simulation

$\overline{mse}_{sim}(\hat{m}_{(.)})$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	0.118	0.126	0.375	0.338
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	0.205	0.193	0.557	0.444
$\bar{m}^{K}_{yrq}$	0.204	0.194	0.548	0.447
$\bar{m}^{S}_{yrq}$	0.203	0.194	0.556	0.450
$\bar{m}^{KK}_{smrq}$	0.513	0.348	0.260	0.298
$\bar{m}^{KS}_{smrq}$	0.463	0.264	0.290	0.290
$\bar{m}^{SK}_{smrq}$	0.300	0.209	0.145	0.135
$\bar{m}^{SS}_{smrq}$	0.296	0.216	0.144	0.151
$\bar{m}_{yr}$	0.207	0.194	0.535	0.442
$\bar{m}_{smr}$	0.599	0.365	0.840	0.060
$\bar{m}_{spr}$	0.591	0.333	0.113	0.069

Table 6.19

Variance of Simulation

$\overline{V}_{sim}\left(\bar{m}_{(.)}\right)$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	0.118	0.126	0.374	0.338
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	0.205	0.193	0.557	0.442
$\bar{m}^{K}_{yrq}$	0.204	0.194	0.548	0.446
$\bar{m}^{S}_{yrq}$	0.203	0.194	0.556	0.449
$\bar{m}^{KK}_{smrq}$	0.470	0.317	0.260	0.120
$\bar{m}^{KS}_{smrq}$	0.444	0.254	0.290	0.102
$\bar{m}^{SK}_{smrq}$	0.300	0.202	0.144	0.129
$\bar{m}^{SS}_{smrq}$	0.295	0.209	0.144	0.145
$\bar{m}_{yr}$	0.207	0.194	0.535	0.441
$\bar{m}_{smr}$	0.344	0.265	0.084	0.059
$\bar{m}_{spr}$	0.590	0.332	0.112	0.069

10. For all estimators as the response rate increases the mean squared error of simulation $\overline{mse}_{sim}\left(\bar{m}_{(.)}\right)$and the sample variance $\overline{V}_{sim}\left(\bar{m}_{(.)}\right)$ decreases.

11. In population 1 at both response rates, the mean squared error of simulation and the sample variance for the finite adjusted general regression estimators $\bar{m}^{K}_{yrq}, \bar{m}^{S}_{yrq}$ and the reduced generalized regression $\bar{m}_{yr}$ are identical and minimal.

12. In population 5 the reduced kernel smoother $\bar{m}_{smr}$ and the reduced spline smoother $\bar{m}_{spr}$ have the smallest mean squared error when the response rate was 90%. When the response rate is reduced to 70% the reduced spline smoother $\bar{m}_{spr}$ has the smallest mean squared error.

Tables 6.20, 6.21 and 6.22 demonstrates the efficiency of each estimator using a different benchmark. In table 20 the generalized regression estimator $\bar{m}_{GRE}$ is used as a

benchmark because for this estimator the response rate is 100%. In table 21 the adjusted generalized regression estimator with known a priori response probabilities $\bar{m}_{yrq}$ is used for the comparisons Finally in table 6.22 all the estimators are all compared to the reduced generalized regression estimator $\bar{m}_{yr}$.

Table 6.20

Efficiency with respect to $\bar{m}_{GRE}$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	1.00	1.00	1.00	1.00
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	0.58	0.65	0.67	0.76
$\bar{m}^{K}_{yrq}$	0.58	0.65	0.69	0.76
$\bar{m}^{S}_{yrq}$	0.58	0.65	0.67	0.75
$\bar{m}^{KK}_{smrq}$	0.23	0.36	2.59	1.13
$\bar{m}^{KS}_{smrq}$	0.26	0.48	2.60	1.16
$\bar{m}^{SK}_{smrq}$	0.39	0.60	1.44	2.50
$\bar{m}^{SS}_{smrq}$	0.40	0.58	1.29	2.24
$\bar{m}_{yr}$	0.57	0.65	0.70	0.76
$\bar{m}_{smr}$	0.20	0.35	3.33	7.45
$\bar{m}_{spr}$	0.20	0.38	4.46	4.87

Table 6.21

Efficiency with respect to $\bar{m}_{yrq}$

Estimate	Population 1		Population 5	
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	1.00	1.00	1.00	1.00
$\bar{m}^{K}_{yrq}$	1.01	1.00	1.02	0.99
$\bar{m}^{S}_{yrq}$	1.01	1.00	1.00	0.99
$\bar{m}^{KK}_{smrq}$	0.40	0.56	3.85	1.49
$\bar{m}^{KS}_{smrq}$	0.44	0.73	3.67	1.53
$\bar{m}^{SK}_{smrq}$	0.68	0.92	2.14	3.28
$\bar{m}^{SS}_{smrq}$	0.69	0.90	1.92	2.94
$\bar{m}_{yr}$	0.99	1.00	1.04	1.00
$\bar{m}_{smr}$	0.34	0.53	4.95	7.46
$\bar{m}_{spr}$	0.35	0.58	6.62	6.39

Table 6.22

Efficiency with respect to $\bar{m}_{yr}$

Estimate	Population 1		Population 5	
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}^{K}_{yrq}$	1.01	1.00	0.98	0.99
$\bar{m}^{S}_{yrq}$	1.02	1.00	0.96	0.98
$\bar{m}^{KK}_{smrq}$	0.40	0.56	2.06	2.48
$\bar{m}^{KS}_{smrq}$	0.45	0.73	1.92	2.52
$\bar{m}^{SK}_{smrq}$	0.70	0.92	2.60	3.27
$\bar{m}^{SS}_{smrq}$	0.69	0.90	2.60	2.94
$\bar{m}_{yr}$	1.00	1.00	1.00	1.00
$\bar{m}_{smr}$	0.35	0.53	6.36	7.44
$\bar{m}_{spr}$	0.35	0.58	4.75	6.38

13. The response rate has an effect on all three efficiencies. As the response rate increases the efficiencies increase.

The following can be inferred from Table 6.20:

14. In population 1 at both response rates the optimal estimator is the generalized regression $\bar{m}_{GRE}$, because of this optimally the efficiency of the finite adjusted general regression estimators $\bar{m}^K_{yrq}, \bar{m}^S_{yrq}$ and the reduced generalized regression $\bar{m}_{yr}$ is larger than the efficiency of the other estimators.

15. In population 5 at both response rates the generalized regression estimator is not optimal (as per Chapter 3), the reduced smoothing kernel estimator $\bar{m}_{smr}$ or reduced smoothing spline estimator $\bar{m}_{spr}$ have the largest efficiencies. The adjusted smoothing kernel estimator $\bar{m}_{smrq}$ and the adjusted reduced smoothing spline estimator $\bar{m}_{sprq}$ have larger efficency than the adjusted reduced generalized regression estimator but smaller than $\bar{m}_{smr}$ and $\bar{m}_{spr}$.

The following can be inferred from Table 6.21:

16. In population 1 at both response rates the finite adjusted general regression estimators $\bar{m}^K_{yrq}, \bar{m}^S_{yrq}$ and the reduced generalized regression $\bar{m}_{yr}$ have efficiencies equal to the adjusted generalized regression estimator with known a priori response probabilities $\bar{m}_{yrq}$. The other estimators have much smaller efficiencies.

17. In population 5 at both response rates the reduced smoothing kernel estimator $\bar{m}_{smr}$ or reduced smoothing spline estimator $\bar{m}_{spr}$ have the largest efficiencies. The efficiency of these estimators are greater than the efficiency of the generalized regression estimator with known a priori response probabilities $\bar{m}_{yrq}$. The adjusted smoothing kernel estimator $\bar{m}_{smrq}$ and the adjusted reduced smoothing spline estimator $\bar{m}_{sprq}$ also have larger efficency than the generalized regression estimator with known a priori response probabilities $\bar{m}_{yrq}$.

The following can be inferred from Table 6.22:

18. In population 1 at both response rates the finite adjusted general regression estimators $\bar{m}^K_{yrq}$ and $\bar{m}^S_{yrq}$ have efficiencies equal to the reduced generalized regression

$\bar{m}_{yr}$. The other estimators have much smaller efficiencies.

19. In population 5 at both response rates the reduced smoothing kernel estimator $\bar{m}_{smr}$ or reduced smoothing spline estimator $\bar{m}_{spr}$ have the largest efficiencies. The efficiency of these estimators are greater than the efficiency of the reduced generalized regression estimator $\bar{m}_{yr}$. The adjusted smoothing kernel estimator $\bar{m}_{smrq}$ and the adjusted reduced smoothing spline estimator $\bar{m}_{sprq}$ also have larger efficency than the reduced generalized regression estimator $\bar{m}_{yr}$.

Finally Table 6.23 shows the relative accuracy of the variance estimator $\bar{V}_{srs}\left(\hat{m}_{(.)}\right)$ with respect to the simulation variance $V_{sim}\left(\hat{m}_{(.)}\right)$.

Table 6.23

Relative Accuracy

$RA(\hat{m}_{(.)})$

Estimate	Population 1		Population 5	
$\bar{m}_{GRE}$	1.03	0.96	0.97	1.06
	RR. = 70%	RR. = 90%	RR. = 70%	RR. = 90%
$\bar{m}_{yrq}$	0.84	1.07	0.92	1.01
$\bar{m}^{K}_{yrq}$	0.84	1.08	0.92	1.01
$\bar{m}^{S}_{yrq}$	0.84	1.08	0.92	1.05
$\bar{m}^{KK}_{smrq}$	1.53	4.85	0.58	0.83
$\bar{m}^{KS}_{smrq}$	1.46	6.10	0.78	1.50
$\bar{m}^{SK}_{smrq}$	0.71	0.84	0.79	1.64
$\bar{m}^{SS}_{smrq}$	0.66	1.22	0.71	1.65
$\bar{m}_{yr}$	0.83	1.06	0.92	1.01
$\bar{m}_{smr}$	2.19	3.88	0.73	1.19
$\bar{m}_{spr}$	0.35	0.54	0.86	0.96

The above table demonstrates that:

20. As the response rate increases the relative accuracy of the sampled variance estimator $\tilde{V}_{srs}\left(\hat{m}_{(.)}\right)$ approaches the sample variance $\overline{V}_{sim}\left(\hat{m}_{(.)}\right)$.

21. The relative accuracy of the adjusted generalized regression estimators and the reduced generalized regression estimator were not as good as the generalized regression estimator. The response rate and the variability of the estimated response probability have an effect on the relative accuracy.

22. The relative accuracy of the kernel smoother and the spline smoother were not as favorable as the relative accuracy the generalized regression estimator. The response rate and the variability of the estimated response probability have an effect on the relative accuracy. Moreover research is still being conductd on finding optimal methods to estimate the varinace of these estimators.

6.6 Conclusions

The results confirm Särndal's and Hui's (1981) conclusions. *If the regression model is representative of the population point scatter, then the estimator is model and design unbiased even if the response probabilities are wrongly estimated.* The chapter has shown that it is possible to estimate a finite population mean with a nonparametric model in the presence of nonresponse by making no assumption about the underlying point scatter and the response mechanism that created the nonresponse.

Chapter 7

Further Topics for Research

7.1 Introduction

The theory and results of the previous chapter have demonstrated that nonparametric regression is a powerful and useful tool that should be in the toolbox of all survey statisticians. The normal kernel regression method and the spline smoother studied in the thesis are not the only nonparametric regression available to estimate the population mean or total. The dissertation used only one auxiliary variable x associated with the response variable y but in practice the response variable may have many predictors. Also a binary regression model was used to estimate the response probabilities ϕ_k but other methods exist to estimate these probabilities. This chapter can be taken as a launch pad for further research.

Different approaches to nonparametric regression will be the focus of section 7.2. In particular the *local linear regression model* and the *locally weighted regression smoother* will be presented in this section. In section 7.3 we will introduce the multivariate kernel regression estimator, the local multivariate regression estimator, generalized additive model, as generalizations of the multivariate regression model. Section 7.4 will introduce the smooth logistic regression model which can be used to estimate the response probabilities ϕ_k.

7.2 Nonparametric Regression

The normal kernel regression method exposed in the previous chapters is called the *local mean regression* estimator. The estimator seeks to average the values of the response variable locally. The *local linear regression* approach is an alternative to the *local mean regression* estimator. The local linear regression

$$\mu(x_k) = \alpha + \beta(x_k - x_i) + \varepsilon_k$$

is the least squares solution to

$$\min_{\alpha,\beta} \sum\nolimits_{i\in s} (y_i - \alpha - \beta(x_k - x_i))^2 K\left(\frac{x_k - x_i}{b}\right). \tag{7.1}$$

Any of the kernels defined in chapter 2 can be used as a weight in the local linear regression model. The solution of the above least squares is

$$\hat{\mu}(x_k) = \frac{1}{n}\sum\nolimits_{i\in s} \frac{(s_2(x_k;b) - s_1(x_k;b))(x_k - x_i) w_{ki} y_i}{s_2(x_k;b)s_0(x_k;b) - s_1(x_k;b)^2} \tag{7.2}$$

where b is the bandwidth, w_{ki} is defined by a kernel weight

$$w_{ki} = \frac{K\left(\frac{x_k - x_i}{b}\right)}{\sum_{i\in s} K\left(\frac{x_k - x_i}{b}\right)}$$

and

$$s_j(x_k;b) = \sum\nolimits_{i\in s} (x_k - x_i)^j w_{ki},\ j = 0,1,2. \tag{7.3}$$

The local linear regression was introduced by Cleveland (1979). Fan and Gijbels (1992) showed that the local linear regression had smaller bias near the boundaries of the covariate space than the local mean regression method.

The following procedure due to Cleveland (1979) has received much attention in estimating $\mu(x_k)$. The method is known as the *locally weighted regression smoother* (**loess**).

A locally weighted smooth $\hat{\mu}(x_k)$ using q nearest-neighbors is calculated as follows:

a. The q nearest-neighbors of x_k are denoted by $Q(x_k)$. The number of neighbors of x_k, q is specified as a percentage of the total number of sampled values. This percentage is called a span which has the same role as the bandwidth b in the kernel smoothing methodology.

b. Calculate the largest distance between x_k and another point in $Q(x_k)$

$$\Delta(x_k) = \max_{Q(x_k)} |x_k - x_j| .$$

c. Weights w_i are assigned to each point in $Q(x_k)$ using the tri-cube weight function

$$w\left(\frac{|x_k - x_j|}{\Delta(x_k)}\right)$$

where

$$w(u) = \left(1 - u^3\right)^3 I\,(0 \leq u < 1)$$

d. $\hat{\mu}(x_k)$ is the fitted value at x_k from the weighted least squares fit of y_k to x_k on the neighborhood $Q(x_k)$.

In locally weighted regression smoothing, the span is constant over the entire sampled x_k values. Now if either the curvature of $\mu(x_k)$ or the error variance $\sigma^2 v(x_k)$ varies over the range of the x_k, a constant span will not produce an optimal fit. The optimal fit can be found with the *supersmoother* which uses a local cross validation and chooses a span for the x_k values by leaving out one at a time the x_j and estimating the w_i on the remaining $n - 1$ points, Simonoff (1996).

The *local linear regression* and *locally weighted regression smoothing* methods are alternative procedures to the *local mean regression model.* Further studies on these two methods within a survey sampling design would enrich the on going research. Preliminary studies have been done with these procedures when the response rate was 100 % and the initial results have been encouraging. One draw back that I have encountered with these

methods is the large cost of computer resources to find the estimates.

7.3 Multivariate Nonparametric Regression

The local mean regression model on a single predictor generalizes in a straightforward way to multiple predictors. For a multidimensional predictor variables $\mathbf{X}_k = (X_{1k}, ...X_{dk})^T$ one uses a multidimensional product kernel function

$$K^*(\mathbf{X}_1, ..., \mathbf{X}_d) = \prod_{j=1}^{d} K(\mathbf{X}_j) \tag{7.4}$$

where $K(\mathbf{X}_j)$ is a kernel for each predictor The kernel weights are now defined as

$$\mathbf{W}(\mathbf{X}_k) = \frac{K^*(\mathbf{X}_1, ..., \mathbf{X}_d)}{\sum_{k \in s} K^*(\mathbf{X}_1, ..., \mathbf{X}_d)} \tag{7.5}$$

Then a multivariate version of the local mean regression estimator is

$$\hat{\mu}(\underline{\mathbf{x}}) = \sum_{k \in s} \mathbf{W}(\mathbf{X}_k) y_k \tag{7.6}$$

which is the multivariate fitted regression surface.

The local linear regression model also generalizes in a straightforward way to multiple predictors and is called *the local multivariate regression estimator.* The local multivariate regression estimator

$$y_i = \beta_0 + \beta_1 (\mathbf{X}_1 - x_{1i}) ... + \beta_d (\mathbf{X}_d - x_{di}) + \varepsilon_i$$

is the least squares solution of

$$\mu_d(\underline{\mathbf{x}}) = \min_{\alpha, \beta} \sum_{i \in s} (y_i - \beta_0 - \beta_1 (\mathbf{X}_1 - x_{1i}) ... - \beta_d (\mathbf{X}_d - x_{di}))^2 K_d \left(\mathbf{H}^{-1}(\mathbf{X}_i - \underline{\mathbf{x}})\right) \tag{7.7}$$

such that $K_d(\mathbf{H}^{-1}(\mathbf{X}_i - \mathbf{x}))$ is a multivariate kernel function with bandmatrix $\mathbf{H}$ and $\underline{\mathbf{x}}$ is a $d \times 1$ vector corresponding to the predictor variables. We now define the design matrix $\mathbf{X}_x$ and the weight matrix $\mathbf{W}_x$ so as to find estimates $\hat{\beta}_0$ and $\hat{\beta}_j$ $(j = 1...d))$. Let

$$\mathbf{X}_x = \begin{pmatrix} 1 & \mathbf{X}_1 - x_{11} & \cdot & \cdot & \cdot & \mathbf{X}_d - x_{1n} \\ \cdot & & & \cdot & \cdot & \\ \cdot & & & \cdot & \cdot & \\ \cdot & & & \cdot & \cdot & \\ 1 & \mathbf{X}_1 - x_{1n} & \cdot & \cdot & \cdot & \mathbf{X}_d - x_{dn} \end{pmatrix}$$

and

$$\mathbf{W}_x = diag(K_d(\mathbf{H}^{-1}(\mathbf{X}_1 - \underline{\mathbf{x}}) ... K_d(\mathbf{H}^{-1}(\mathbf{X}_n - \underline{\mathbf{x}})).$$

If the matrix $\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x$ is invertible then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{y}. \tag{7.8}$$

The estimator $\hat{\mu}_d(\underline{\mathbf{x}})$ is the intercept term $\hat{\beta}_0$ or

$$\hat{\mu}_d(\underline{\mathbf{x}}) = \mathbf{e}_1^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{y} \tag{7.9}$$

where $\mathbf{e}_t$ is the $(p+1)$ vector having the value 1 in the $t'th$ entry and zero elsewhere. Properties of this estimator can be found in Ruppert and Wand (1991).

As the dimension of the regressor hyperplane increases, estimates of the nonparametric multivariate regression gets progressively more difficult. An important consequence of this pattern is the somewhat paradoxical fact that in high dimensions, local neighborhoods are almost surely empty, and neighborhoods that are not empty are almost surely not local. This has been called the *curse of dimensionality.* A way around these difficulties is to restrict the form of the multivariate regression estimator. Friedman and Stuetzle (1981) and Hastie and Tibshirani (1985) proposed the *generalized additive model*

to overcome the curse of dimensionality.

The multivariate regression model is now written as

$$y_i = \mu_d(\underline{\mathbf{x}}) = \beta_0 + \mu_1(\mathbf{x}_{1i}) + ... + \mu_d(\mathbf{x}_{di}) + \epsilon_i \tag{7.10}$$

where μ_j $(j = 1...d)$denote functions whose shapes are unrestricted, apart from an assumption of smoothness and conditions such as $\sum_{i=1}^{n} \mu_j(\mathbf{X}_{ji}) = 0$ for each j. The additivity assumption allows all of the one dimensional smoothing methods to be used in the multivariate context. The intercept parameter β_0 can be estimated by the mean of the respondents $\bar{y}$ because of the restriction that each additive component sums to zero. Hastie and Tibshirani (1990) P. 99 proposed the *backfitting algorithm* to estimate the generalized additive model. In the sequel S_j is an arbitrary scatter plot smoother (kernel or spline).

a. Initialize $\hat{\beta}_0 = \bar{y}$, $\mu_j(\mathbf{x}_j) = \mu_j^0$ *for* $j = 1...d$.

b. Cycle $j = 1...d, 1...d, ...$

$\mu_j(\mathbf{x}_j) = S_j\left(\mathbf{y} - \beta_0 - \sum_{k \neq j} \mu_k(\mathbf{x}_k)\right)$

c. Continue (b.) until the individual functions don't change.

Let $\hat{\mu}_j(\mathbf{x}_j)$be the estimated value of $\mu_j(\mathbf{X}_j,)$therefore the multivariate regression model is estimated by

$$\hat{y}_i = \hat{\mu}_d(\underline{\mathbf{x}}) = \hat{\beta}_0 + \hat{\mu}_1(\mathbf{x}_{1i}) + ... + \hat{\mu}_d(\mathbf{x}_{di}). \tag{7.11}$$

All the concepts presented in this dissertation are based on a univariate model. Further studies and analysis of these multivariate methods within a survey sampling design framework would hopefully extend the ideas presented in this thesis. Preliminary studies have been done and were encouraging, with generalized additive model using spline smoothers. More research has to be done in this area with different types of smoothers and applied to the areas of full sample response and partial sample response.

7.4 Response Probabilities

In chapter 5 we used binary kernel and binary spline regression models to estimate the response probabilities ϕ_k. The local linear and the locally weighted regression models can be modified to binary regression models so as to estimate the probabilities ϕ_k.

An alternative procedure to estimate the probabilities ϕ_k is to use a smooth logistic regression. The log-likelihood has the form

$$l(\beta_0, \beta_1) = \sum\nolimits_{k \in s} l_k(\beta_0, \beta_1) K\left(\frac{x_k - x}{b}\right) \tag{7.12}$$

where $l_k(\beta_0, \beta_1)$is the contribution to the usual log-likelihood from the kth observation in other words

$$l_k(\beta_0, \beta_1) = y_k \log\left(\frac{\phi_k}{1-\phi_k}\right) + \log(1-\phi_k). \tag{7.13}$$

The ϕ_k denotes the response probability at the sampled value x_k. The logit link function is defined as

$$\log it(\phi_k) = \log\left(\frac{\phi_k}{1-\phi_k}\right) = \beta_0 + \beta_1 x_k \tag{7.14}$$

and is assumed to link the response probability ϕ_k to the auxiliary variable x_k.

The log-likelihood defines a *local likelihood*. The local likelihood is found by summing the contribution of each l_k for each observation, weighted by the distance between the corresponding x_k and the point of estimation x.

Maximization of $l(\beta_0, \beta_1)$ provides local estimates $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. The local estimates then can be used to find the estimated response probability $\hat{\phi}_k$ at x_k. The $\hat{\phi}_k$ is found with the following relationship

$$\hat{\phi}_k = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_k\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_k\right)}. \tag{7.15}$$

The generalized additive model also provides another method to find the response probabilities ϕ_k. The advantage of this method is that it can be used if one has one or many predictors for the response probability.

The methods proposed in this section would provide estimates of the response probabilities. Are these estimates of ϕ_k any better than the binary regression estimates of chapter 5? Does the nonresponse bias decrease sufficiently by using these estimates of ϕ_k? These questions among others are important topics for further research!

Bibliography

[1] Azzalini, A. Bowman, A. and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*, **79**, 1-11.

[2] Bailar, J. and Bailar, B. (1978). Comparison of Two Procedures for Imputing Missing Survey Values. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

[3] Bailar, J., Bailey, B. and Corby, C. (1977). A Comparison of Some Adjustment and Weighting Procedures for Survey Data. *Symposium of Survey Sampling and Measurement, University of North Carolina.*

[4] Bowman, A. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis The Kernel Approach with S-Plus Illustrations.* Clarendon Press, Oxford.

[5] Cassel, C. G., Särndal, C. E. and Wretman, J. H. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

[6] Cassel, C. G., Särndal, C. E. and Wretman, J. H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys* (eds. W.G. Madows and I. Olkin), **3**, 143-160. Academic press, New York.

[7] Clevland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829-36.

[8] Cochran, W. G. (1977). *Sampling Techniques, 3rd ed.*, Wiley, New York.

[9] Doane, D. P. (1976). Sesthetic Frequency Classifications. *Amer. Statist.* **30**, 181-183.

[10] Devroye Wagner (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, 231-9.

[11] Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *J. R. Statist. Soc. B*, **31**, 195-224.

[12] Ernst, L. (1978). Weighting to Adjust for Partial Nonrespnse. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

[13] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Deker, New York.

[14] Fan, J. and Gijbels I. (1992). Variable bandwidth and local linera regression smoothers. *Ann. Statist.*, **20**, 2008-36.

[15] Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, **76**, 817-23.

[16] Gasser, T., Stroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biomerika*, **73**, 625-633.

[17] Giommi, A. (1985). On estimation in nonresponse sistuations. *Statistica*, **1** 57-63.

[18] Giommi, A. (1987). Nonparametric Methods for Estimating Individual Response Probabilities. *Survey Methodology* **13**, 127-133.

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

[20] Godambe, V. P. and Joshi, V. M. (1965). Admissibility and Bayes estimation in sampling finite populations I. *Ann. Math. Statist.*, **36**, 1707-22.

[21] Griliches, Z. , Hall, B. H. and Hausman, J. A. (1977). Missing Data and Self-Selection in Large Panels. *Discussion Paper No. 573, Harvard Institute of Economic Research.* Harvard University, Cambidge, USA.

[22] Hájek, J. (1960). Limiting distributions in Simple Random Sampling from a Finite Population. *Periodica Mathmatica*, **5**, 361-374.

[23] Hall, P. and Marron, J. S. (1988). On variance estimation in nonparametric density estimation and regression. *J. Multivariate Anal.*, **27**, 228-54.

[24] Hall, P. , Kay, J. and Titterington, D. M. (1990) Asymptotically optimal difference based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521-528.

[25] Hansen, M. H., and Hurwitz, W. N. (1943). On the theory of sampling from finite population. *Ann. Math. Statist.* **35** 1491-1523.

[26] Hansen, M. H., and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *J.Amer. Statist. Assoc.*, **41**, *517-529.*

[27] Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983). An evaluation of model dependent and probability-sampling inferences in sample surveys. *J.Amer. Statist. Assoc.*, **78**, *776-793.*

[28] Härdle, W. (1989). Asymptotic maximal deviation of M- smoothers. . *Multivariate Anal.*, **28**, 374-390.

[29] Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

[30] Hastie, T. and Tibshirani, R. (1985). Generalized additive models. *Technical, Report,* **98**, Dept. of Statistics, Stanford University.

[31] Hastie, T. and Tibshirani, R. (1995). *Generalized additive models.* Chapman & Hall, London.

[32] Hausman, J. and Spence, M. (1977). Non-random Missing Data. *Working Paper*, No. 200, Department of Economics. M.I.T.

[33] Issaki, C. T. and Fuller, W. A. (1981). Survey design under a superpopulation model. *Current Topics in Survey Sampling* 199-226. New York Academic Press.

[34] Issaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J.Amer. Statist.* **77,** *89-96.*

[35] Little, R. J. A. (1983). Superpopulation models for nonresponse. In *Incomplete Data in Sample Surveys* (eds. W.G. Madows and I. Olkin), **2**, 337-413. Academic press, New York.

[36] Madow, W.G. Olkin, I. and Rubin D. B. (1983). *Panel on Incomplete Data in Sample Surveys P.I.D.S.S.* **Vol 1, 2 and 3**, Academic Press, New York.

[37] Nadaraya, E. A. (1964). On estimating regression. Theory Probab. Appl., **10,** 189-190.

[38] Nargundkar, M. S., and Joshi, G. B. (1975). Nonresponse in sample surveys. 40th Session of the International Statistical Institute, Warsaw, Contributed papers, 626-628.

[39] Neyman, J. (1934). On two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. J. *R. Statist. Soc.* **97**, 558-625.

[40] Niyonsenga, T. (1994). Nonparametric Estimation of Response Probabilities in Sampling Theory. *Survey Methodology* **20,** 177-184.

[41] Nordheim, E. V. (1979). Discriminant Ananlysis with Nonrandomly Missing Data. *Technical Report* **No. 556**, Department if Statistics, University of Wisconsin-Madison, USA.

[42] Politz, A. and Simmons,W. (1940). An attempt to get not-at-homes into the sample without callbacks. *J.Amer. Statist.* **44**, 9-31.

[43] Rancourt, E., Lee, H. and Särndal, C. E. (1994). Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse. *Survey Methodology* **20**, 137-147.

[44] Raz, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. *J. Amer. Statist. Assoc.*, **85**, 132-138.

[45] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12**, 233-243.

[46] Roussas, G. (1997). *A Course in Mathematical Statistics. Second Edition.* New York Academic Press.

[47] Royall, R. M. (1968). An old approach to finite population sampling theory. *J. Amer. Statist. Assoc.*, **63**, 1269-1279.

[48] Royall, R. M. (1970a). On finite population sampling theory under certain linear regression models. *Biometrica* **57**, 377-387.

[49] Royall, R. M. (1970b). Finite population sampling theory - on labels in estimation. *Ann. Math. Statist.* **41**, 1774-1779.

[50] Robinson P. M., and Särndal C. E. (1983). Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling. *Sankhyā B*, **45**, 240-248.

[51] Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Amer. Statist. Assoc.*, **72**, 538-543.

[52] Rubin, D. B. (1983). Conceptual issues in the presence of nonresponse. In *Incomplete Data in Sample Surveys* (eds. W.G. Madows and I. Olkin), **2**, 123-142. Academic press, New York.

[53] Särndal, C. E. (1972). Sample survey theory vs. general statistical theory: Estimation of the population mean. *Rev. Int. Statist. Inst.*, **40**, 1-12.

[54] Särndal, C. E. (1976). On uniformly minimum variance estimation in finite populations. *Ann. Statist.*, **4**, 993-997.

[55] Särndal, C. E. and Hui, T. K. (1981). Estimation for Nonresponse Situations: To What Extent Must We Rely on Models? *Current Topics in Survey Sampling* 227-246. New York Academic Press.

[56] Särndal, C. E., Swensson, B. Wretman J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

[57] Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci.*, **52**, 947-950.

[58] Scott, A., and Wu, C. F. (1981). On the Asymptotic distribution of ratio and regression estimators. *J. Amer. Statist. Assoc.*, **76**, 98-102.

[59] Silverman, B. W. (1986). *Density Estimation for Statistics abd Data Analysis.* Chapman & Hall, London.

[60] Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer-Verlag, New York.

[61] Wahba, G. and Wold, S. (1975). A completely automatic French curve: fitting spline functions by cross validation. *Commmun. Statist.*, **4**, 1-17.

[62] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā A*, **26**, 359-372.