University of Southampton

Faculty of Mathematical Studies

Mathematics

A Nonparametric Regression Approach to Prediction in Finite Populations

by

Kathleen Elizabeth Bennett

Thesis submitted for the degree of Doctor of Philosophy

This thesis was submitted for examination in July 1994.

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MATHEMATICAL STUDIES

MATHEMATICS

Doctor of Philosophy

A Nonparametric Regression Approach
to Finite Population Prediction
by Kathleen Elizabeth Bennett

Nonparametric regression provides an intuitive estimate of a regression function or conditional expectation without the restrictions imposed by parametric models. This is a particularly useful property since the rigidity of such parametric models is not always desirable. The application of nonparametric regression, in the univariate setting, is investigated in the context of predicting a finite population total. We propose, instead of parametric estimators of the finite population total, nonparametric regression estimators obtained by smoothing the data and interpolating the smooth to predict nonsample values. It is shown how such estimation can be more robust and efficient than inference tied to parametric regression models. The nonparametric regression estimators considered are classified as *operational* and *model-based.* They require the selection of a smoothing parameter which controls the smoothness of the resultant curve. Methods of choosing the smoothing parameter are discussed.

One important property that some of the estimators are shown to possess is 'total preservation'. Suppose $\hat{\mathbf{y}} = \mathbf{Sy}$, where $\mathbf{S}$ is a smoother matrix and $\hat{\mathbf{y}}$ and $\mathbf{y}$ are vectors of estimated and observed y respectively. Then an estimator is said to be 'total-preserving' if $\mathbf{1}^T\hat{\mathbf{y}} = \mathbf{1}^T\mathbf{Sy} = \mathbf{1}^T\mathbf{y}$. It is shown how, under repeated sampling, 'total preserving' nonparametric regression estimators are design-unbiased or approximately design-unbiased and how they remain more efficient than standard parametric methods, for a suitable choice of the smoothing parameter.

Contents

List of Tables

List of Figures

Acknowledgements

This research was carried out under a research studentship from the Scientific and Engineering Research Council.

I would like to thank Dr. David Firth, my supervisor over the last three years, for his support and without whom this thesis would not have been possible.

I would also like to express my gratitude to Prof. Fred Smith for his help during my research period.

Thanks also to Ray Brown and Peter Hubbard, for their continued computing assistance, and to friends Mandy, John and Sunita, Paula, Lorraine, Rory and Christine who have helped along the way.

A special thank-you to my parents and grandparents who in their own way have helped and encouraged me.

Chapter 1

Introduction

Nonparametric regression provides an estimate of a regression function or conditional expectation without the restriction imposed by parametric models. Nonparametric regression allows the data to decide on the best functional form. Methods for nonparametric regression range from simple ideas such as moving averages or kernel smoothing to the more complex spline smoothing, thus providing a flexible family of approaches to choose from.

In this thesis we consider the use of nonparametric regression methods in the context of finite population prediction. The data of a finite population are made up of units usually consisting of a response variable, y, which is the variable of interest, and possibly other auxiliary or explanatory variables, in some way related to the response variable. In the univariate setting, which we are mainly considering, there is just one explanatory variable, x, associated with the response. One may be interested in determining a population quantity of interest, which is typically some function of the response. Linear functions include the mean or population total and quadratic functions include the population variance. If a sample of the response variable is all that is available, because of possible cost or time constraints, an estimator of the population quantity of interest will be required. The estimator will typically rely on some implicit underlying parametric model, for example, the ratio estimator for the population total is based on a linear regression through the origin with the variance proportional to x. The parametric models used tend not to be robust to misspecification of the true underlying curve because of their rigidity to a specific form. If, instead, nonparametric regression is used to model the functional relationship between x and

y, the strict rigidity is removed often leading to more efficient estimation. Estimating 'parameters' of the finite population from the sample is now based on *smoothing* the sample data and interpolating this smooth to predict nonsample values. A description of the problem is given below.

1.1 Description of the problem

A population of interest consists of N units labelled $i = 1, \ldots, N$. Associated with unit k are two numbers (x_k, y_k): x_k known, y_k fixed but unknown for the whole population (usually known for a sample of the population only). Here x_k may be some measure of size, for example size of hospital k in a population of hospitals, and y_k some characteristic of interest, say the number of patients discharged from hospital k. The quantity of interest is the population total

$$T = \sum_{i=1}^{N} Y_i.$$

We assume that the explanatory variable, x, is in some way related to Y and is known for the entire population. However, the response variable Y is only known for a sample of n units from the population. A popular choice as an estimator of the population total has been the ratio estimator.

$$\hat{T}_{RE} = \sum_{i=1}^{N} x_i \frac{\sum_{j=1}^{n} y_j}{\sum_{j=1}^{n} x_j}.$$

This estimator could have been derived in two different ways:

1. Quasi-likelihood, a generalisation of the weighted least squares approach. This approach does not specify an error distribution for the observations, only a mean-variance relationship. In this case with variance proportional to the mean, we arrive at the ratio estimator. For example, in terms of weighted least squares estimation, the ratio estimator is the best linear unbiased estimator (B.L.U.E) of the population total, when we assume a superpopulation model with the following first and second moments:

$$E(y_i) = \mu_i = \beta x_i$$

$$\text{var}(y_i) = \sigma^2 x_i.$$

The estimator is best in the sense of minimum variance (Gauss-Markov Theorem).

2. Maximum likelihood estimation. Here we assume a likelihood for the data, in this case Poisson likelihood, with mean $\mu_i = \beta x_i$. The log-likelihood for the Poisson distribution can be written as :

$$l(\mu, y) = \sum_{i=1}^{n} (y_i \log \mu_i - \mu_i).$$

Maximising this gives

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i},$$

hence

$$\hat{T}_{RE} = \sum_{i=1}^{N} x_i \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}.$$

This approach applies to any generalised linear model. Generalised linear models are briefly introduced in Chapter 3, and are considered in more detail in McCullagh and Nelder (1989).

Royall and Herson (1973a) consider some robustness and efficiency properties of the ratio estimator. In particular they introduce i) *balanced* samples with respect to the population to overcome the problem of a misspecified superpopulation model, and ii) stratification on the size variable x_i. Let x_{ij}, y_{ij} be the values of x_i, y_i in stratum j. The stratification is based on the x variable. By estimating, for each stratum, a separate slope through the origin, $\hat{\beta}_j = \sum_{j=1}^{n} y_{ij} / \sum_{j=1}^{n} x_{ij}$, a *within stratum* estimator is derived as:

$$\hat{T}_j = \sum_{i=1}^{N_j} x_{ij} \hat{\beta}_j$$

where N_j is the number of population units in stratum j. The population total is then estimated by the *separate* or *stratified ratio estimator*

$$\hat{T}_{SRE} = \sum_{j=1}^{J} \hat{T}_j.$$

This approach provides some robustness to failure of the assumption of an underlying superpopulation model. Royall and Herson (1973b) also consider balanced sampling in ensuring robustness. Balanced samples are defined for a Pth degree polynomial as the pth sample moment of x's equalling the pth population moment, for all $p = 1, \dots, P$. When stratification is used in an optimal manner under balanced sampling it is more efficient than the simple balanced strategy. However, the global superpopulation model underlying this method is discontinuous at the boundaries of the strata, with consequent loss of efficiency when the true regression is a smooth function of x. The stratified approach will be suboptimal when the true curve is smooth, because of the discrete 'jumps' at the boundaries. Figure 1.1 has been included to illustrate this. Other estimators of the population total exist, with different underlying parametric regression models, to adapt to differing situations. Some of these are discussed in Chapter 2, and are referred to again in Chapter 6 when we address robustness within the sampling framework by incorporating the inclusion probabilities.

We propose to consider alternative estimators using the method of local likelihood estimation suggested by Hastie and Tibshirani (1986); see also Tibshirani and Hastie (1987) and Buja, Hastie and Tibshirani (1989). This new approach to the robustness problem allows a different coefficient vector $\hat{\beta}(\mathbf{x}_\mathrm{i})$ for each x_i in the population and is a generalisation of kernel smoothing to be introduced in Chapter 3. Here we replace the locally weighted means in kernel smoothing with locally weighted regression curves. This greatly reduces the well-known 'edge effect' bias present in ordinary kernel smoothing and removes the need to use boundary modifications.

One representation of a nonparametric regression estimator of the population total can be written as:

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \notin s} \hat{y}_i.$$

For example,

$$\hat{T}_{LWRE} = \sum_{i \in s} y_i + \sum_{i \notin s} x_i \hat{\beta}(x_i),$$

where

$$\hat{\beta}(x_i) = \frac{\sum_{j \in s} W_b(x_i, x_j) y_j}{\sum_{j \in s} W_b(x_i, x_j) x_j},$$

and $j \in s$ denotes the values in the sample, etc. The $\hat{\beta}(x_i)$ is derived from a set of 'local' quasi-likelihood estimating equations, assuming a model with a single explanatory variable, no intercept and variance proportional to the mean, as in the case of the ratio estimator. The motivation and derivation is given in Chapter 4 along with other possible nonparametric regression estimators. In this case $\hat{\beta}(x_i)$ is an estimate of a *local* slope at x_i and the weights $W_b(x_i, x_j)$ are derived from a suitable kernel, for example, the Gaussian kernel

$$W_b(x_i, x_j) = \exp\left(\frac{-(x_i - x_j)^2}{2b^2}\right).$$

Figure 3.1 gives a plot of a Gaussian kernel estimator with varying bandwidths, including $b = \infty$. The b appearing in the weight is the *smoothing parameter* or bandwidth. It determines how wide to make the 'window' for weighting observations around a target x_i. If the window is large, i.e. $b \to \infty$, then most or all of the sample observations will be included in the estimation at x_i, with approximately equal weight. As b becomes smaller more weight is given to observations close to x_i and less weight to those far away; the fitted curve becomes local. As $b \to 0$ the curve will come close to every data point thus becoming quite 'wiggly' in appearance. The role of b is therefore important, more so than the actual choice of kernel function used, since it determines how smooth the resulting curve will be.

This smoothing approach does not create the artificial discontinuities apparent in the separate ratio estimator. It assumes a smooth curve for the underlying superpopulation model which is more realistic, and in most practical situations will lead to increased efficiency, for an appropriate choice of smoothing parameter.

The local quasi-likelihood equations mentioned above can be extended to any number of auxiliary variables and any type of data, provided the first two moments of the response variable y are specified. Thus the scope for application as a general approach to smoothing is large; some of these applications are discussed in Chapter 7. For more on quasi-likelihood in general see McCullagh and Nelder (1989, Chapter 9) or Firth (1993).

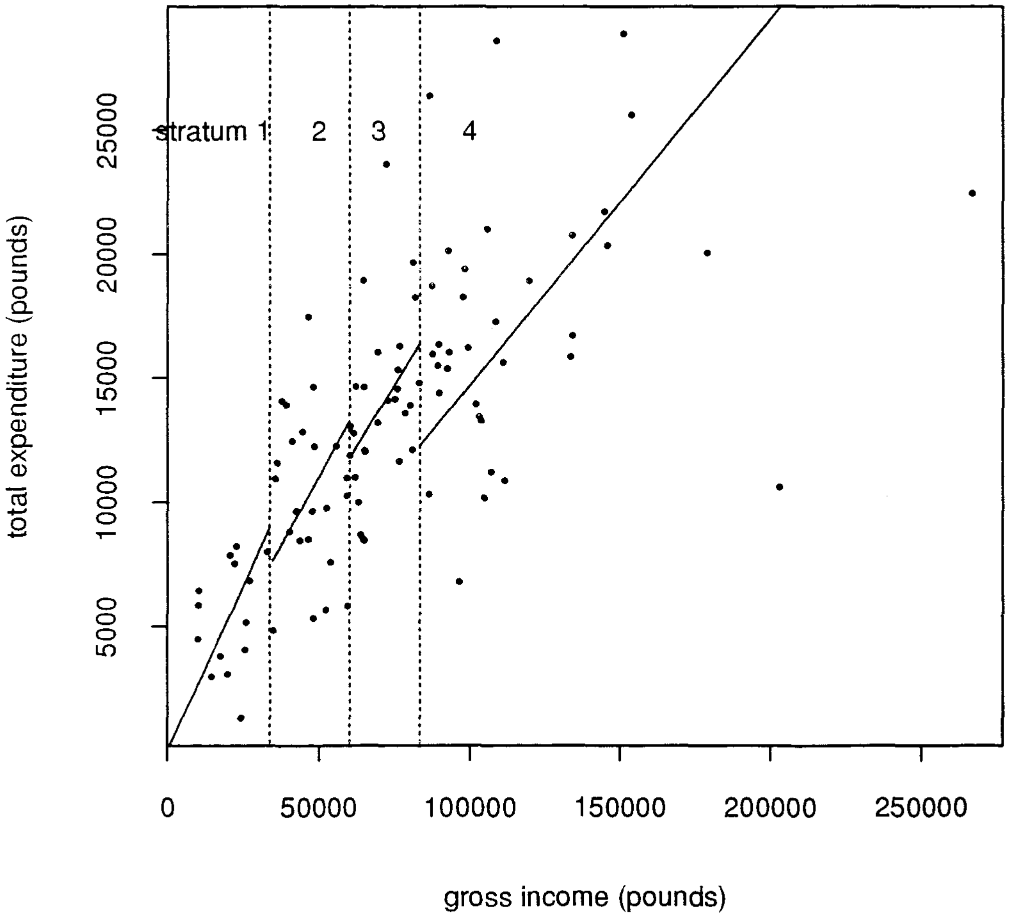


Figure 1.1: The separate ratio estimator, including discontinuties at the stratum boundaries

1.2 Notation

This section will explain the notation used throughout.

The random variables X and Y are the explanatory and response variables respectively. Observed values are in lowercase indexed by the unit number, for example (x_i, y_i). The sample size is denoted by n and the population size by N. The sampling fraction is denoted by $n/N = f$.

A vector of random variables is written in boldface as $\mathbf{y} = (y_1, \ldots, y_n)$. Matrices are also represented in boldface, for example $\mathbf{S}$, with elements denoted by S_{ij} for example. The trace of a matrix is denoted by tr(.). The marginal density of x is denoted by $f(x)$ and $\hat{f}(x)$ is an estimator for $f(x)$. The joint density of x and y is $f(x, y)$. The regression curve of Y on x is $m(x) = E(Y|X = x)$ and $\hat{m}(x)$ is an estimator of $m(x)$. Expectation is denoted by E: with a subscript ξ to denote with respect to the superpopulation model, and with subscript π to denote with respect to the sampling plan. The conditional variance of $Y|X = x$ is denoted by var$(Y|x) = \sigma^2(x) = [E(Y^2|X) - m^2(x)]$ and the bias by bias$(\hat{T})$=$E(\hat{T} - T)$. The mean squared error is given by MSE $= E\left[\hat{T} - \sum_{i=1}^{N} m(x_i)\right]^2$ and the predictive mean squared error by PMSE $= E\left[T - \hat{T}\right]^2$.

The distribution of Y is sometimes indexed by $\mu = E(Y)$ or $\eta = g(\mu)$, where g is a monotonic function known as the *link* function. The hat notation denotes an estimator of a population characteristic made from the sample. The various estimators of the population total are denoted by $\hat{T}$ with an abbreviated subscript to describe the type of estimator; T denotes the true population total. Observations in the sample are denoted by $j \in s$ and those not in the sample by $j \notin s$. Other notation used is explained wherever necessary.

1.3 Rest of the thesis

The rest of the thesis is organised as follows. In Chapter 2 an introduction to finite population prediction inference is given with emphasis on the two main approaches that are frequently used: the superpopulation model approach and the sampling theory approach. Some standard parametric estimators of the finite population are given and the role of π-weighting introduced. Chapter 3 will consider smoothing and nonparametric regression generally in order to give some background to this area, before looking specifically at nonparametric regression in predicting totals in Chapter 4. Also in Chapter 3 a brief introduction to generalised linear models is given. In Chapter 4 motivations and derivations for our class of nonparametric regression estimators are explained and an important 'total-preserving' property that some estimators have is introduced. This is a beneficial property for any estimator to have and it is later established in Chapter 6 that estimators with this property have good design-based properties. The nonparametric regression estimators described in this chapter are classified as operational (or automatic), and model-based estimators. The model-based estimators are based on generalised linear models with the introduction of nonparametric regression components, for example regression splines in the linear predictor. The bias, variance and mean squared error of these estimators, under the superpopulation model, are also given. In Chapter 5 the question of the choice of smoothing parameter is addressed: a review of some of the methods that already exist is given, including crossvalidation and methods based on asymptotics of estimators such as the Gasser-Müller and Nadaraya-Watson estimators. These are introduced with a view to modifying them to the specific problem of finite population prediction. Some asymptotic bias and variance properties, i.e. as $n \rightarrow \infty$, $b \rightarrow 0$, of the nonparametric estimators are also given. Examples are included to illustrate the use of crossvalidation as a bandwidth selector method and the application of approximate asymptotic methods. In Chapter 6 we introduce π_i, the ith inclusion probability, as a weight into the estimators. This ensures, in some cases, design unbiasedness or approximate design unbiasedness under repeated sampling. We also consider the approximate design-variance of one estimator, the locally weighted ratio estimator. Numerical simulation results confirm that π-weighting is beneficial in nonparametric regression estimators. When we include π-weighting it can be shown that nonparametric regression estimators are often more efficient than their parametric counterparts. Finally in Chapter 7 we state the conclusions from this work and indicate areas of possible further research. Where appropriate, examples are used throughout the thesis.

1.4 Datasets used

In this section details of the datasets used in this thesis are given, since some of the datasets are used more than once. The datasets described here have been associated with the ratio estimator in the published literature and make for a good comparison with the nonparametric regression estimators. We draw random samples from the population. The method of selecting these samples is explained in Chapter 2.

1.4.1 Hospitals

This is referred to in Royall and Herson (1973a) and Herson (1976). It is the population that fell into the January 1968 sample of the National Center for Health Statistics' hospital discharge survey, and consists of $N = 393$ short-stay hospitals as the population units. Each unit or hospital consists of two variables, x_k and y_k:

- x_k is the known number of beds in hospital k
- y_k is the number of patients discharged from hospital k in January 1968.

These data were kindly supplied by Richard Royall.

1.4.2 India

This dataset appears in Hanurav (1967) and consists of population sizes of $N = 72$ districts in four states of India. The population unit is the district and for each unit there are two variables:

- x_k= population size of the district (rounded to nearest thousands) as per 1951 census.
- y_k= population size of the district as per 1961 census.

1.4.3 Wheat

This dataset was obtained from Sukhatme (1954). The population consists of $N = 32$ observations, and for each unit:

- x_k= number of wheat acres in 1936
- y_k=number of wheat acres in 1937.

1.4.4 Family Expenditure Survey, 1968-1983

This data is available from the ESRC Data Archive at Essex University. The original population consisted of $N = 7058$ observations; each unit in the population referred to a particular household in the survey. Several variables were recorded for each household: some useful measures include total expenditure, expenditure broken down by housing, fuel, food and transport and gross household income. For each population unit we have considered:

- x_k=known gross household income (pounds)
- y_k= total expenditure (pounds).

A random sample of 500 units was selected from the original population of 7058 units and this was used as the 'population'.

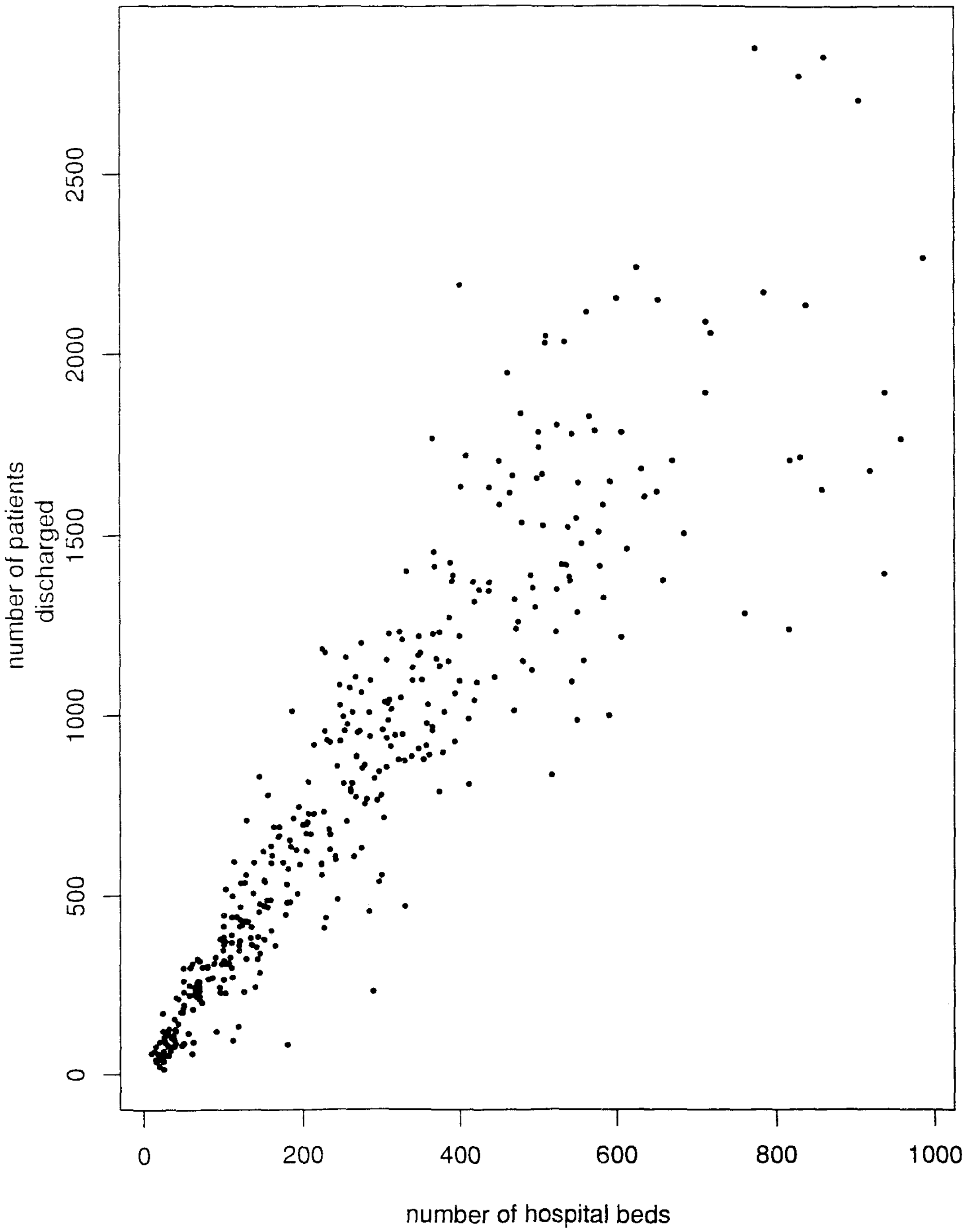


Figure 1.2: Scatter plot of hospital dataset

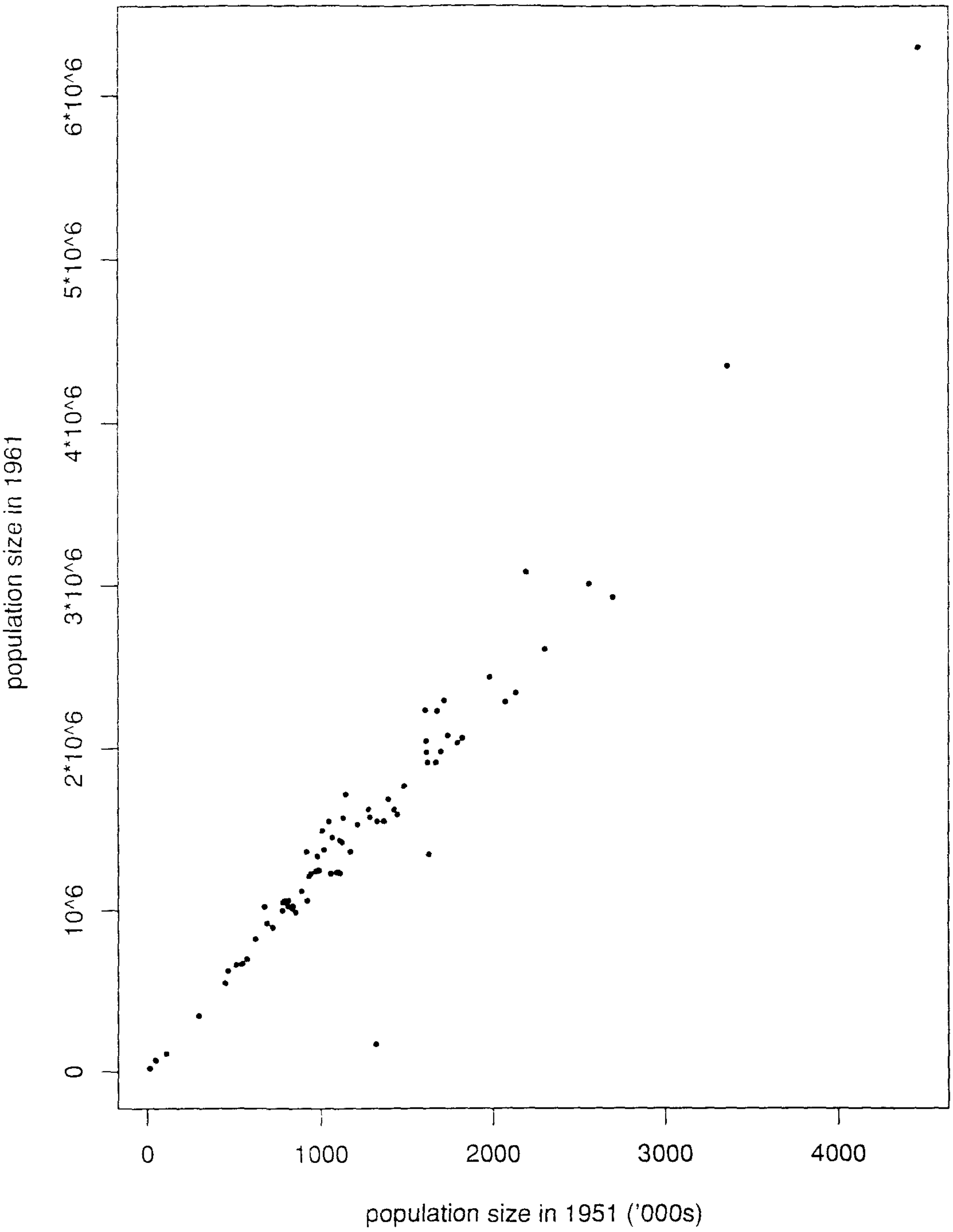


Figure 1.3: Scatter plot of India population

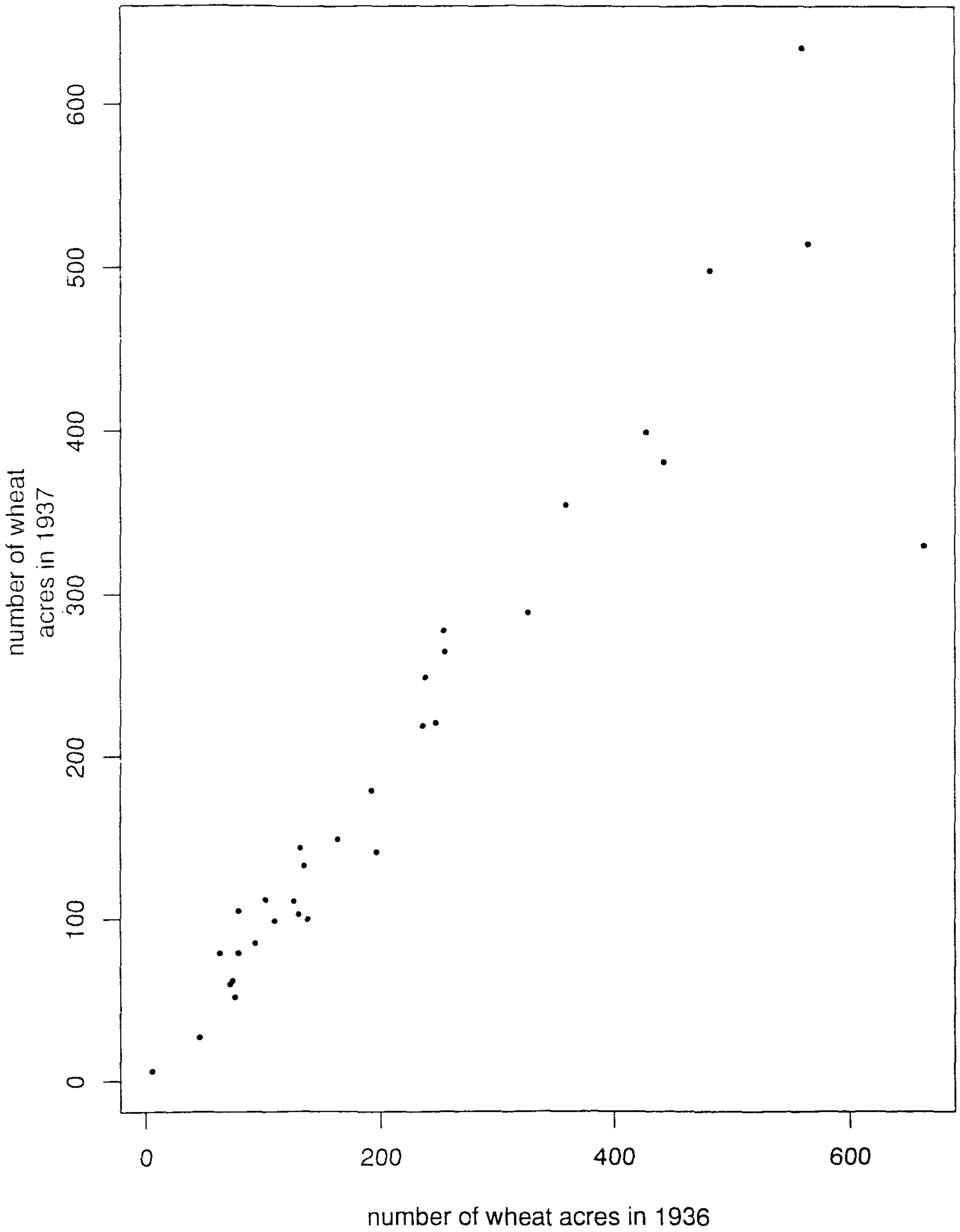


Figure 1.4: Scatter plot of wheat population

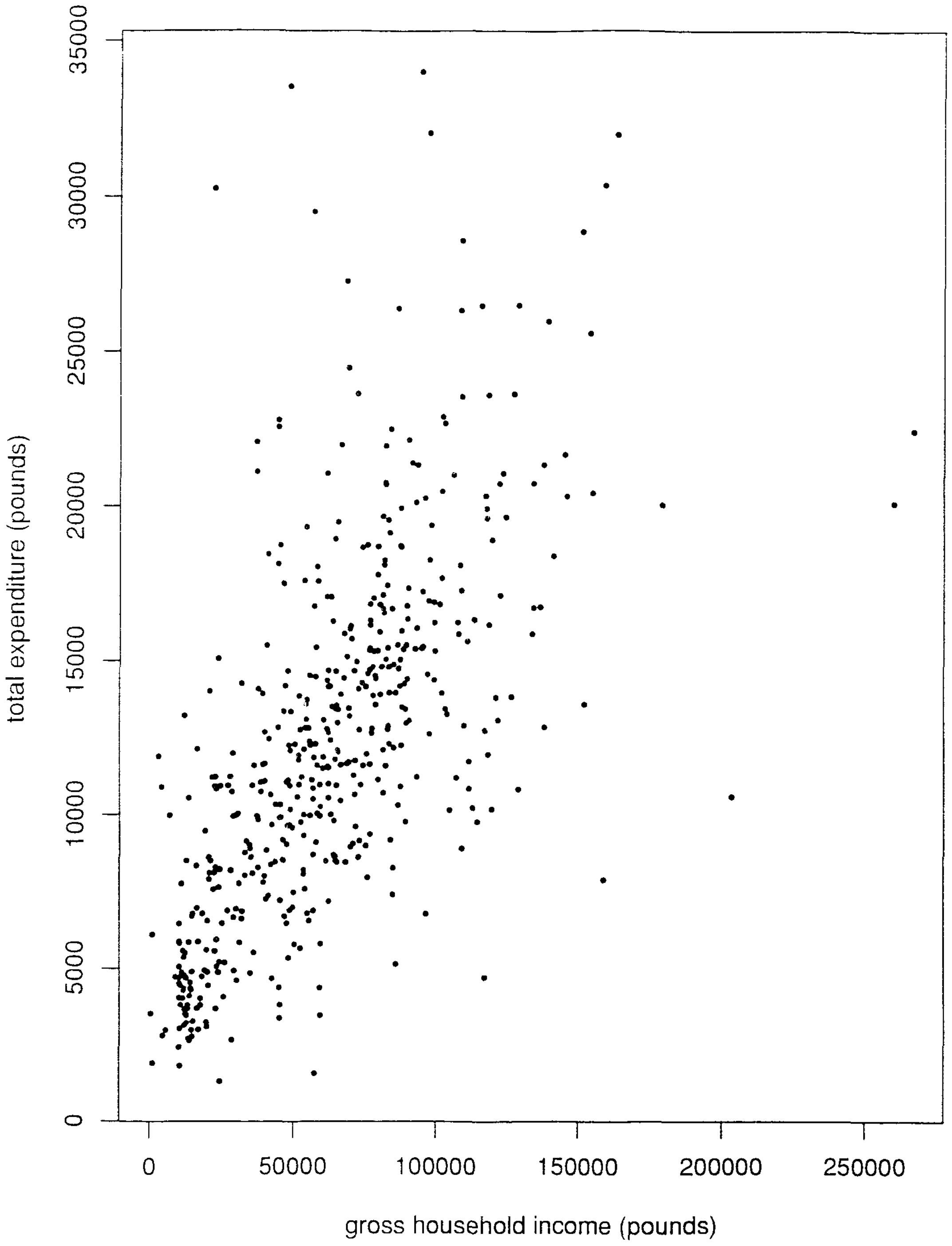


Figure 1.5: Scatter plot of data from the Family Expenditure Survey

Chapter 2

Basic ideas and principles of finite population prediction

2.1 Introduction

In this chapter we consider the different approaches to inference in finite populations in more detail. The aims are to introduce the principles behind the selection of a sample from a finite population and the inferences based on characteristics of interest of the target population. As an introduction, some basic terminology and notation that is used in this chapter is given.

A finite population is a collection of N units, where N is called the *size* of the population. Associated with the ith unit there may be one or more variables. The variable of interest for the population, y_i, $i = 1, \ldots, N$, is known as the response variable. There may be other variables, x_{ri}, $r = 1, \ldots, p$, $i = 1, \ldots, N$ associated with unit i, which may be used to describe the variable of interest by a particular regression function. These are known as *auxiliary* or *explanatory* variables. In order to gain information about a function $\theta(\mathbf{y})$ of $\mathbf{y} = (y_1, \ldots, y_N)$, a sample of size n ($\leq N$) is selected from the population and the y-values observed. The x variables are assumed known for the whole population.

Functions of the response variable that one may be interested in include linear functions, such as the population mean or total.

The design (or sampling plan) is represented as a probability function defined on the sample space of all possible samples s of size n. A wide variety of standard selection schemes are available including simple random sampling, stratified sampling, systematic sampling, cluster sampling, etc. A detailed discussion of the properties of designs can be found in Cassel et al. (1977).

After s has been chosen, we denote by y_s and y_r the observed and unobserved parts of y respectively. An estimator, $\hat{\theta}(y_s)$ is a function of y_s, thought to produce values that lie near unknown population quantities of interest for most samples. Estimators can be functions of study variables x and y. An *estimate* is produced from an estimator after a specific outcome from a sample has been observed. Properties of estimators such as expected value, variance and mean squared error can be found with respect to the sampling design. This is discussed further in Section 2.2.

2.2 Two methods of inference

2.2.1 Superpopulation models

This more recently developed *prediction* approach, e.g. as in Brewer (1963) and Royall (1970), views $y_1, \ldots, y_N$ as realisations of random variables $Y_1, \ldots, Y_N$. After the sample has been observed, an estimate of the population function requires predicting the function of the unobserved Y's. Relationships among the variables are expressed in an assumed model of their joint probability distribution, and predictions are made with reference to this model; randomisation probabilities play no role in this inference. The assumed model is known as the 'superpopulation model', and this approach is referred to as *model-based.*

The superpopulation model, from which the finite population was drawn, is denoted by ξ.

The model for the superpopulation may, for example, be a generalised linear model with

$$\eta_i = g(\mu_i) = \sum_{j=0}^{J} \beta_j x_i^j.$$

The link function, $g(.)$, is typically known and describes the transformation of the μ's to the linear predictor. The β_j are the parameters that are to be estimated in the model, e.g. in a polynomial regression model. A generalised linear model can be specified in terms of the mean and variance of the response variable:

$$E_\xi(Y_i) = \mu_i(\beta) = g^{-1}(\beta^T \mathbf{x}_i);$$

$$\text{var}_\xi(Y_i) = \phi V(\mu_i), \quad \text{cov}_\xi(Y_i, Y_j) = 0, \quad (i \neq j).$$

The parameters in the model are estimated by maximising the likelihood associated with the generalised linear model. The monograph by McCullagh and Nelder (1989) gives a detailed account of generalised linear models.

The representation, in terms of mean and variance, described above is also useful in defining a class of estimators. Parameter estimates obtained by maximum likelihood define a particular estimator, when their values are replaced or 'plugged back' into the model. When $V(\mu_i)$ does not depend on the mean μ_i, weighted least squares can be used to obtain parameter estimates and hence estimators from the specified model. When $V(\mu_i)$ depends on μ_i, then quasi-likelihood estimating equations can be used to obtain parameter estimates and hence an estimator.

The inference in this setting is tied to the particular sample s that was realised, and not to other samples.

Consider 'Hospitals' (Section 1.4), where x is the number of beds and y the number of patients discharged during one month. From the population of N=393 observations a sample of size n is selected and the y variable observed; the x variable is known for the whole population. To estimate the total number of patients discharged

$$T = \sum_{i=1}^{N} y_i$$

we write it first as

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \notin s} y_i,$$

where s is the sample of n hospitals. If y_i is a realisation of the random variable Y_i, $i = 1, \ldots, N$, then estimating T is equivalent to predicting $\sum_{i \notin s} Y_i$ of unobserved values. The first and second moments for the joint probability distribution of the Y's might be

$$E_\xi(Y_i) = \beta x_i, \qquad (i = 1, \ldots, N)$$

and

$$\text{cov}_\xi(Y_i, Y_j) = \begin{cases} \sigma^2 x_i & (i = j), \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

This seems a reasonable model for this set of data, and is the model most used in the literature. Robustness to failure of the working model is also an important consideration. This is usually understood to mean that inferences made from a sample are insensitive to violations of the assumptions that have been made. Royall and Herson (1973a,b) have given special attention to robustness and recommend the use of balanced samples. They also recommend disproportionate sampling with optimum allocation of sample to strata, to ensure bias-robustness. Simple random sampling (SRS) can be a valuable tool for choosing approximately balanced samples, and hence protecting against bias incurred when, for example, we have overlooked a regressor. If the x variable is used to divide the population into H strata, then a 'separate' estimator can be used. For example, suppose $\bar{x}_h$ is the average size of N_h hospitals in stratum h, and $\bar{y}_{sh}$ and $\bar{x}_{sh}$ are sample averages in stratum h; then an estimator of the population total is

$$\hat{T}_{SRE} = \sum_{h=1}^{H} \frac{\bar{y}_{sh}}{\bar{x}_{sh}} N_h \bar{x}_h.$$

This is known as the *separate ratio estimator* or *stratified ratio estimator*. Assuming model (1.1) described above, but with different parameters (β_h, σ_h^2) in each stratum, this estimator is the best linear unbiased estimator (BLUE) of the population total T. When the sample from each stratum is balanced, i.e. $\bar{x}_{sh} = \bar{x}_h$, $\hat{T}_{SRE}$ becomes simply $\sum_{h=1}^{H} N_h \bar{y}_{sh}$. Even when balance is not achieved within each stratum, stratification itself limits the degree of imbalance possible, so that $\hat{T}_{SRE}$ is protected from extreme bias by using piecewise linear models (see Figure 1.1).

Model based quantities such as bias, variance and mean squared error can be obtained. An estimator is said to be model-unbiased if, given s,

$$E_\xi\left(\hat{\theta}(\mathbf{y}_s)\right) - \theta(\mathbf{y}) = 0.$$

Quantities measuring variability are the model variance

$$\mathrm{var}_\xi\left(\hat{\theta}(\mathbf{y}_s)\right) = E_\xi\left(\hat{\theta}(\mathbf{y}_s) - E_\xi(\theta(\mathbf{y}))\right)^2$$

or the model mean squared error

$$\mathrm{MSE} = E_\xi\left(\hat{\theta}(\mathbf{y}_s) - \theta(\mathbf{y})\right)^2,$$

respectively. Model-based or prediction intervals can be derived if we use an estimator of variance or mean squared error obtained from the sample data.

In order to arrive at estimators of the bias and variance described above we may need to estimate parameters such as σ^2 and coefficients appearing in the superpopulation model. This is performed based on the sample observations, for example, an unbiased estimator of σ^2 under a model with $\text{var}_\xi(Y_i) \propto x_i$ is

$$\sigma^2 = \frac{1}{n-1} \sum_{k \in s} \frac{(y_k - \hat{y}_k)^2}{x_k}.$$

This can however be severely biased if some other model holds. Royall and Eberhardt (1975) consider robust model-based variance estimators for the ratio estimator when proportionality to x does not hold. Model-based properties such as these are given in Chapter 4 for a variety of finite population total estimators. Departures from the superpopulation model covariance matrix are also discussed in Royall (1988) but have not been considered further in the work presented in this thesis.

A thorough account of superpopulation models can be found in Cassel, Särndal and Wretman (1977).

2.2.2 The classical, design-based approach

A probability sampling plan or design is a scheme for choosing the sample such that there is a known probability $\pi(s)$ of selecting a subset of units s.

The properties of $\hat{\theta}(\mathbf{y}_s)$, the random quantity calculated from the sample, are expressed in terms of its expected value, variance and mean squared error with respect to the design $\pi(s)$. The *design bias* can be written as

$$E_\pi(\hat{\theta}(\mathbf{y}_s)) - \theta(\mathbf{y}) = \sum_{s \in \mathcal{S}} \pi(s)(\hat{\theta}(\mathbf{y}_s)) - \theta(\mathbf{y})$$

where summation is over all possible samples s. The *design variance* is similarly defined as

$$\text{var}_\pi\left(\hat{\theta}(\mathbf{y}_s)\right) = \sum_{s \in \mathcal{S}} \pi(s)[\hat{\theta}(\mathbf{y}_s) - E_\pi(\hat{\theta}(\mathbf{y}_s))]^2,$$

and *design mean squared error* as the sum of design variance and squared design bias.

Definition 2.1 *A design unbiased (or π-unbiased) estimator is one satisfying*

$$E_\pi(\hat{\theta}(\mathbf{y}_s)) = \theta(\mathbf{y}),\ \forall \mathbf{y}.$$

Design-unbiasedness guarantees robustness of inference to misspecification of the stochastic process underlying the population. Design unbiasedness and approximate design-unbiasedness are particularly good features for any estimator to have.

Definition 2.2 *The π_i, $i = 1, \ldots, N$ denote the inclusion probabilities of the ith unit of the population in the sample and depend on the given sampling design.*

If

$$I_i(s) = \begin{cases} 1 & i \in s \\ 0 & \text{otherwise,} \end{cases}$$

is a random indicator for whether a given unit is in the sample, then

$$\pi_i = E_\pi[I_i(s)] = \sum_{s \in \mathcal{S}} \pi(s) I_i(s).$$

The $I_i(s)$ is the *sample membership indicator* of unit i. We also have second order inclusion probabilities which are the probability of inclusion of the ith and jth unit of the population in the sample. These are defined similarly to the above. Let

$$I_{ij} = \begin{cases} 1 & i \in s, j \in s,\ j \neq i \\ 0 & \text{otherwise ;} \end{cases}$$

then

$$\pi_{ij} = E_\pi[I_{ij}(s)] = \sum_{s \in \mathcal{S}} \pi(s) I_{ij}(s).$$

The following relations hold:

$$\begin{aligned} \textstyle\sum_{i=1}^N \pi_i &= n, \\ \textstyle\sum_{i \neq j}^N \pi_{ij} &= (n-1)\pi_i, \\ \textstyle\sum_i \sum_{j>i}^N \pi_{ij} &= n(n-1)/2. \end{aligned}$$

Next we describe examples of two designs commonly used in survey sampling: simple random sampling without replacement (SRS) and stratified simple random sampling (SSRS).

Simple random sampling

In simple random sampling, each of the possible ${}^N C_n$ samples has an equal probability of selection. The sampling design is:

$$\pi(s) = \begin{cases} 1/({}^N C_n) & \text{if } s \text{ has } n \text{ units} \\ 0 & \text{otherwise.} \end{cases}$$

This type of design is known as an equal probability design, because the inclusion probabilities are all equal. For SRS the first and second order inclusion probabilities are

$$\pi_i = \frac{n}{N} = f \qquad (i = 1, \ldots, N),$$

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \qquad (i \neq j = 1, \ldots, N).$$

This design is often taken as a point of reference when comparing other designs.

Stratified SRS

Most designs in practice are unequal probability designs because they are often more efficient. For instance, stratified simple random sampling with proportional or optimal allocation is an unequal probability design. Here the population is divided into non-overlapping subpopulations, or strata. Stratification nearly always leads to a smaller variance for the estimator than the comparable SRS, if optimal allocation is used. This method of selection requires the population stratum sizes N_h to be known so that, once the strata have been determined, simple random samples of size n_h from N_h can be selected within strata independently.

Before stratification can be performed a number of questions are required to be answered:

1. Which stratification variable should be used?

2. How should the stratum boundaries be chosen?

3. How many strata should be used?

The answers to these are normally related and can depend on the precision of the estimates required, cost considerations and administrative restrictions.

If the sampling design is stratified sampling, then a sample n_h is selected from N_h according to a design $\pi_h(.)$, $h = 1, \ldots, H$ and selection in one stratum is independent of selections made in other strata. The N_h are assumed known and $\sum_{h=1}^{H} N_h = N$. The inclusion probabilities here are

$$\pi_{hi} = \frac{n_h}{N_h} \qquad (i = 1, \ldots, N_h),$$

$$\pi_{hij} = \frac{n_h(n_h-1)}{N_h(N_h-1)} \qquad (i \neq j = 1, \ldots, N_h).$$

If i and j belong to different strata then $\pi_{ij} = \pi_i \pi_j$. With stratification the variance is usually reduced and may be reduced even further if optimal allocation of sample units to strata is made. This allocation, also known as Neyman allocation, is achieved by minimising the variance subject to some fixed constraint such as cost or sample size. The form of the optimal allocation is

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

where S_{yh} is the standard deviation associated with y's in stratum h. It indicates that the larger the variation within a stratum, the larger the n_h allocated to that stratum should be.

An alternative allocation that is used is *proportional* or *equal* allocation, where

$$n_h = n \frac{N_h}{N}.$$

If stratum standard deviations are the same then proportional allocation is optimal.

2.2.3 Summary

The fundamental difference between these two approaches (model and design based) is that model-based inferences are based on the actual sample observed (conditional), and an assumed model of the superpopulation from which the finite population was drawn, while design based inferences are based on the randomisation plan used to select the sample s, and averaging over the set of samples from which s was drawn (unconditional).

The paper by Hansen, Maddow and Tepping (1983) gives an evaluation and contrast of the design and model-based approaches to inference in survey sampling. They give a summary of principles they think should be used in practice and an example where the model-based approach can lead to serious biases when the assumed superpopulation differs from the true population.

A useful introduction to finite population inference is given by Bolfarine and Zacks (1991), with emphasis on Bayesian methods, and by Särndal et al. (1992).

In this thesis both approaches described above are considered in the problem of predicting a finite population total. Properties such as bias, variance and mean squared error under an assumed superpopulation model are given for some alternative estimators in Chapter 4. There it is shown how the alternative, nonparametric regression estimators are often more efficient than the standard methods. In Chapter 6 we consider the design-based approach. Again we are interested in bias, variance and mean squared error but under repeated sampling. We highlight the gains to be found when using π-weighted estimators.

The model and design-based approaches are different and important and receive separate attention in this thesis. A recent paper by Smith (1994) reviews both these approaches.

2.2.4 Design-model based approach

It is also of some interest to consider the combined expectation under the design and superpopulation models. One aim might be to find an estimator which is design and model unbiased, i.e. such that

$$E_\pi \left[E_\xi(\hat{T} - T) \right] = 0,$$

or to find the expectation under repeated sampling of $\text{var}_\xi(\hat{T})$, the variance of $\hat{T}$ under some superpopulation model, in order perhaps to minimise it. This type of approach has been considered by Godambe and Joshi (1965), and more recently by Godambe and Thompson (1986). In their paper, Godambe and Thompson focus on estimating equations which define the target population quantity, assumed to be linear. They derive a design unbiased estimator, which is optimal under the joint superpopulation and design model. This estimating equation is just a sample based version of the population estimating equation, inversely weighted by the inclusion probabilities. An estimate $\hat{\theta}(\mathbf{y}_s)$, is then obtained by solving the π-weighted sample estimating equation. For more on the role of sampling weights see Pfeffermann (1993).

Robustness of model and design-based inference in complex surveys using smoothing is discussed in Njenga (1990) and Smith and Njenga (1992).

Joint design-model based approaches are referred to again in Section 6.6, when specific examples are considered.

2.3 Some standard estimators

2.3.1 Introduction

As mentioned in Section 1.2.1, estimators of population quantities are often based on a specific parametric model. We focus on estimators of the population total, but the same approach can be extended to any population quantity of interest.

The classical estimators used to predict the finite population total have been the expansion estimator, ratio estimator, separate and combined ratio estimators and regression estimators, all of which are described below. These estimators are based on implicit underlying parametric regression models, where the parameter estimates are obtained by weighted least squares or quasi-likelihood estimating equations. This is in contrast to the nonparametric regression models which are introduced in Chapter 3, where the 'parameter estimates' are based on 'local' quasi-likelihood estimating equations.

2.3.2 The simple location model

This model can be specified by:

$$E_\xi(Y_i) = \mu_i(\beta) = \beta; \qquad \mathrm{var}_\xi(Y_i) = \sigma^2,$$

where β is the location parameter. The expansion estimator

$$\hat{T}_E = N\bar{y}_s,$$

is the best linear unbiased estimator of T under this model. Under other models this estimator may be biased; in some cases the estimator may be made unbiased by choosing a balanced sample.

2.3.3 Linear regression through the origin

The model specifying this relation is:

$$E_\xi(Y_i) = \mu_i(\beta) = \beta x_i; \qquad \mathrm{var}_\xi(Y_i) = \sigma^2 x_i.$$

Under this model the ratio estimator,

$$\hat{T}_{RE} = \sum_{i=1}^{N} x_i \frac{\sum_{j\in s} y_j}{\sum_{j\in s} x_j},$$

is the best linear unbiased estimator (BLUE) of T. The relationship of this estimator with weighted least squares and quasi-likelihood was introduced in Chapter 1. The ratio estimator can be biased under other regression models. For example, the ratio estimator is biased for a polynomial regression model, unless the sample moments are balanced with respect to the population moments. Royall and Herson (1973a) describe balanced sampling with respect to the ratio estimator in more detail.

2.3.4 Separate linear regressions through the origin

Suppose now the population has been stratified into H strata of size N_h $(h = 1, \ldots, H)$. One model is to consider separate linear regressions through the origin within strata, i.e.

$$\begin{aligned} E_\xi(Y_{hi}) = & \ \mu_{hi}(\beta_h) = \ \beta_h x_{hi}; \\ \mathrm{var}_\xi(y_{hi}) = & \ \sigma^2 x_{hi} \qquad (h = 1, \ldots, H). \end{aligned} \tag{2.2}$$

The best linear unbiased estimator from this model, the separate ratio estimator, can be written as:

$$\hat{T}_{SRE} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} x_{hi} \frac{\sum_{j \in sh} y_{hj}}{\sum_{j \in sh} x_{hj}},$$

where sh denotes the sample values in stratum h. Royall and Herson (1973b) consider balanced sampling in conjunction with the separate ratio estimator.

2.3.5 Linear regression with an intercept

The relation between x_i and y_i may not be restricted to be linear or to go through the origin but may be based on linear or polynomial regression with an intercept term. For example, the linear regression with intercept model may be written as:

$$\begin{aligned} E(Y_i) &= \mu_i(\beta) = \beta_0 + \beta_1 x_i; \\ \mathrm{var}_\xi(Y_i) &= \sigma^2. \end{aligned} \tag{2.3}$$

The BLUE of T from model (1.3) is:

$$\hat{T}_{LR} = N\left(\bar{y} + \beta(\bar{X} - \bar{x})\right),$$

where β is an estimate of the change in y value when x is increased by one unit. The ratio estimator, in this case, corresponds to using $\bar{y}/\bar{x}$ for β. It is also possible to have separate regression estimators within each stratum and then to sum the separate estimates to obtain a separate regression estimator. These regression estimators can be extended to any class of polynomial regression equation.

Regression estimators are not covered extensively in this thesis, but many of the ideas that are considered can be applied to the regression setting quite easily.

2.4 The role of π-weighting

An important and widely used π-unbiased estimator of the population total is the Horvitz-Thompson estimator

$$\hat{T}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i},$$

due to Horvitz and Thompson (1952). When we assume simple random sampling, this π-weighted estimator is just the expansion estimator. A *π-weighted* estimator is defined as any estimator where the sample values appearing in the estimator are divided by their corresponding inclusion probabilities. This ensures good design properties such as design unbiasedness, or in some cases approximate design unbiasedness.

The π-weighted ratio estimator with π_i's from stratified simple random sampling is known as the *combined ratio estimator* and is defined as:

$$\hat{T}_{CRE} = \sum_{i=1}^{N} x_i \frac{\sum_{h=1}^{H} \sum_{j \in sh} y_{hj}/\pi_{hj}}{\sum_{h=1}^{H} \sum_{j \in sh} x_{hj}/\pi_{hj}},$$

with $\pi_{hj} = n_h/N_h$, hence

$$\hat{T}_{CRE} = \sum_{i=1}^{N} x_i \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{\sum_{h=1}^{H} N_h \bar{x}_h}.$$

The difference between $\hat{T}_{CRE}$ and the separate ratio estimator, $\hat{T}_{SRE}$, lies in the way the sample strata x and y means are utilized. The $\hat{T}_{CRE}$ is based on the ratio of the sum of sample strata x means to the sum of the sample strata y means. The $\hat{T}_{SRE}$ is based on the ratio of the individual sample strata x and y totals, summed over the strata.

For more on these models and estimators also refer to Cochran (1977), or Bolfarine and Zacks (1991).

Some π-weighted estimators are described in Chapter 6, where we consider properties under repeated sampling in more detail.

Chapter 3

Smoothing and nonparametric regression

3.1 The idea of nonparametric regression

A regression curve describes a general relationship between an explanatory variable x and a response variable y. At each value of x, the average of Y is given by the regression function; the relationship can be written as

$$y = m(x) + \epsilon$$

with unknown regression function m and observation errors ϵ_i, having zero mean. Regression analysis aims to produce a reasonable approximation to the true regression function m. A parametric approach would be to give m in some prespecified form, described by a finite set of parameters; in nonparametric regression there is no specific form, thereby offering a flexible approach.

The related problem of nonparametric density estimation has received extensive attention in the statistical literature since the idea was introduced by Rosenblatt (1956). There the proposition was one of smoothing a histogram by averaging kernel functions. Since then and with the pioneering papers of Nadaraya (1964) and Watson (1964) there is now a growing literature on the problem of nonparametric estimation of an unknown regression function; see Härdle (1990a) or Hastie and Tibshirani (1990) for a review and key references in the literature.

The basic idea of nonparametric regression is similar to that in density estimation. For estimating the density at a given x, we consider points in a small neighbourhood around x, then we weight the frequencies in the neighbourhood. In regression fitting we are more interested in weighting the response y in a certain neighbourhood of x. We weight observations y_i depending on the distance of x_i to x, i.e. we use an estimator such as

$$\hat{m}(x) = \sum_{i=1}^{n} W_b(x; x_i)Y_i,$$

where $W_b(.)$ defines a weight function depending on b, the smoothing parameter and the sample of explanatory variables $x_1, \ldots, x_n$. Almost all nonparametric regression techniques are weighted averages of the response variable y.

If $m(x)$ is believed to be smooth then observations close to x should contain information about the value of $m(x)$ at x. Thus it should be possible to use something like a local average at x to construct an estimate of $m(x)$. This type of simple averaging of response values y having predictor values close to a target value is known as *smoothing* the data. Local averaging is performed in a neighbourhood around the target value and primarily depends on two things:

1. how to average the response values around the target values

2. how large to make the spread of contributing values around a target value.

The first decision is the *type* of smoother to choose, and here kernel smoothing is a popular choice because of its simplicity.

The second decision is rather more difficult. The size of the neighbourhood around the target x is typically expressed in terms of the *smoothing parameter*, sometimes known as the *bandwidth* or *span*. This determines among other things how far away observations are allowed to be to still significantly contribute to the estimation of $m(x)$, and governs the peakedness of the kernel in kernel smoothing. Large values of b produce estimates of low variance, which tend to have high bias, and which produce smoother estimates. The opposite is true for small values of b, which produce much 'wigglier' curves. Thus there is a fundamental trade-off between variance and bias governed by the smoothing parameter; there has been much literature on the choice of the best smoothing parameter and this is discussed in more detail in Chapter 5. Figure 3.1 gives a plot illustrating how the underlying smoothing parameter controls the amount of smoothing performed.

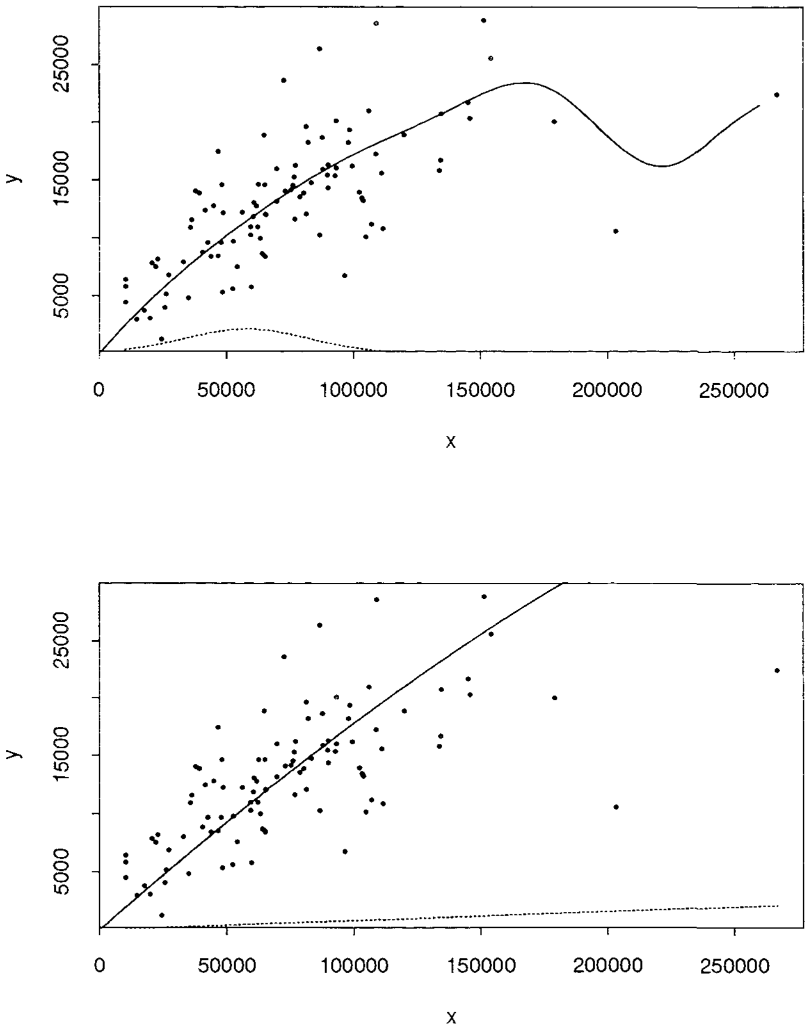


Figure 3.1: The Gaussian kernel smoother uses the Gaussian density function to assign weights to neighbouring points (dotted curve). The spread of this weight function determines the smoothness of the resultant regression curve (solid curve).

The best known estimators are:

1. Kernel estimators with subtypes
   - Priestley-Chao (1972), Gasser and Müller (1979), Cheng and Lin (1981), Benedetti (1977).
   - Nadaraya (1964) and Watson (1964) estimator for random designs.
2. K-nearest neighbour estimators, Buja, Hastie and Tibshirani (1989).
3. Local regression, Cleveland (1979), Fan (1992, 1993).
4. Smoothing splines, which are currently increasing in popularity, Eubank (1988), Green and Silverman (1994).

Most of the literature on nonparametric regression function estimation deals with kernel methods and their variants, which are considered in more detail in Section 3.3. Uses of nonparametric regression for model checking have been considered by Azzalini, Bowman and Härdle (1989) and further by Firth, Glosup and Hinkley (1991).

3.2 Basic principles of Generalised linear models

Kernel regression described above generalises further in generalised linear models; a brief introduction to these models is now given. The generalised linear model (Nelder and Wedderburn (1972), McCullagh and Nelder (1989)) consists of a *random* component, a *systematic* component and a *link* function, linking the two components. The random component describes the conditional distribution of Y given $X = x$ and for generalised linear models is assumed to be a member of an exponential family with probability density function

$$f_Y(y; \theta, \phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y, \phi)\},$$

for some $a(.)$, $b(.)$, and $c(.)$, canonical parameter θ and known ϕ, the dispersion or nuisance parameter. The mean and variance of Y are derived from the log-likelihood function

$$l(\theta, \phi; y) = \ln f_Y(y; \theta, \phi) \quad \text{as} \quad E(Y) = \mu = b^{'}(\theta) \quad \text{and} \quad \text{Var}(Y) = b^{''}(\theta)a(\phi),$$

where $b''(\theta) = V(\mu)$ is known as the *variance function* related to the mean function. The more important distributions of the above type include the Normal, Poisson, Binomial, Gamma and Inverse Gaussian distributions.

The systematic component of the model, denoted by η, relates the mean to the known covariates $\mathbf{x}_1, \ldots, \mathbf{x}_p$, in a linear predictor, *e.g.*

$$\eta = \sum_{i=1}^{p} \beta_j x_j.$$

The *link* function, g(.), between the systematic component and μ, the random component, may be any monotonic differentiable function:

$$\eta = g(\mu).$$

Some examples include the identity, log, logit, probit and complementary log-log links. When $\theta = \eta$, each of the distributions described above has what is called a *canonical link*; these simplify the algebra and algorithms used to find the parameter estimates. Given the random and systematic component, a link function, a vector of n observations and corresponding predictor variables $\mathbf{x}_1, \ldots, \mathbf{x}_p$, the maximum likelihood estimators of β, the parameters in the linear predictor, can be obtained. These are defined as the solution to the score equations

$$\sum_{j=1}^{n} x_{ir} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) V_i^{-1}(y_i - \mu_i) = 0, \qquad r = 1, \ldots, p,$$

where $V_i = \text{var}(Y_i)$. The Fisher scoring method or adjusted dependent variable regression are standard methods for solving these equations. See McCullagh and Nelder (1989) for more detail on this and other aspects of generalised linear modelling. When the random component is not of the generalised linear model type we have the important class of quasi-likelihood models due to Wedderburn (1974). Instead the random component is specified in terms of the first two moments only. A special case of a quasi-likelihood model is weighted least squares where the weights depend on the current parameter estimates.

3.3 Local likelihood and local quasi-likelihood estimation

The local likelihood method extends nonparametric regression techniques to likelihood based regression models. A simple illustration of likelihood based nonparametric regression is now given. Suppose, for each x in a population, a least squares lines is fitted to the data in a *neighbourhood* or *span* around the x value; here the local likelihood is based on normal distribution assumptions. The estimator derived from this method is known as a *running line* estimator.

3.3.1 The details of local likelihood

Tibshirani and Hastie (1987) extend smoothing ideas to other kinds of data whose relationship is expressible through a likelihood function. Consider the general likelihood setting and suppose $(x_i, y_i) \quad (i = 1, \ldots, n)$ are independent realisations of random variables X, Y and $(Y|X = x) \sim f(Y, \theta)$, where θ is a function of x. Then the corresponding likelihood is

$$L(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} f(y_i, \theta_i),$$

where $\theta_i = \beta_0 + \beta_1 x_i$ say after modelling. The likelihood or log-likelihood is then a function of β_0 and β_1. These parameters are estimated by maximising the log-likelihood.

Local likelihood assumes that θ_i is a smooth function of x_i i.e. $\theta_i = m(x_i)$. Here Hastie and Tibshirani define $\hat{m}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i$ where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ maximise the local likelihood

$$L_i(\beta_{0i}, \beta_{1i}) = \prod_{j \in N_i} f(y_j, \beta_{0j} + \beta_{1j} x_j),$$

and where N_i is a neighbourhood of points around a target x_i. This can be applied to the case of generalised linear models (Nelder and Wedderburn (1972)), where the linear predictor $g(\mu) = \beta_0 + \beta_1 x$ is generalised to $m(x)$. The estimation of the $\beta_i = (\beta_{0i}, \beta_{1i})$ can be performed using a Newton-Raphson search in each neighbourhood.

The method of local likelihood can be extended to local *quasi-likelihood*, with known variance function, V(μ) (McCullagh and Nelder (1989) Chapter 9). The special case V(μ) = 1 corresponds to least squares. In general the local quasi-likelihood estimating equations are defined as

$$\sum_{j \in N_i} \frac{(y_j - \mu_j)}{V(\mu_j)} \times \frac{\partial \mu_j}{\partial \beta_r} = 0, \qquad (r = 1, \ldots, p)$$

for the unknown parameters β in $\mu_j(\beta) = E(Y_j|x_j)$. For any specified value of x, a local model with parameter vector $\beta(x)$ may be estimated by solving the equations

$$\sum_{j=1}^{N} W_b(x, x_j) \frac{(y_j - \mu_j)}{V(\mu_j)} \frac{\partial \mu_j}{\partial \beta_r} = 0, \qquad (r = 1, \ldots, p),$$

where $W_b(x, x_j)$ is a kernel weight function. Several forms of weighting are discussed in Hastie and Tibshirani (1990). Fan (1993) and Fan and Gijbels (1993) use local likelihood in design-adaptive nonparametric regression. They derive a type of locally weighted linear regression which adapts to random and fixed design and without modification at the boundaries of the x value. Fan, Heckman and Wand (1992) also consider local polynomial kernel regression for generalised linear models, which has good properties provided the polynomial being used is of odd degree.

Fitting local lines appears to be worthwhile since it reduces bias at the endpoints, where fitted constants do not (Hastie and Loader, 1993).

3.4 Nonparametric regression smoothing techniques - Local means

3.4.1 Kernel smoothing

Kernel smoothers as local means can be derived from the method of local likelihood described above, assuming normal errors, an identity link and a constant model with no covariates present. The estimating equation would be :

$$\sum_{j=1}^{n} (y_j - \mu(x)) W_b(x, x_j) = 0.$$

Kernel smoothers used in regression problems have the common form

$$\hat{m}(x) = \sum_{j=1}^{n} W_b(x, x_j) Y_j,$$

a locally weighted average of the Y_j about a target point x. This local average is constructed in such a way that observations close to x contribute significantly, while those further away contribute less, because of their different means. If $K(.)$

is a smooth density function, some examples of weight functions constructed using $K(.)$ are :

(a)

$$W_b(x, x_j) = \frac{(x_j - x_{j-1})}{h} K\{(x - x_j)/b\},$$

where $x_1 < \ldots < x_n$,

(b)

$$W_b(x, x_j) = \frac{K\{(x - x_j)/b\}}{\sum_{j \in s} K\{(x - x_j)/b\}},$$

(c)

$$W_b(x, x_j) = \frac{1}{b} \int_{t_{j-1}}^{t_j} K\{(x - s)/b\} ds,$$

where $t_{j-1} < x_j < t_j$. The special case $t_j = x_j$ has been investigated by Cheng and Lin (1981).

The *kernel function* $K(.)$ may be chosen to be, for example, a symmetric unimodal probability density function. One example is the Gaussian kernel

$$K\{(x - x_j)/b\} = \exp\left(-(x - x_j)^2/2b^2\right),$$

but there are various other commonly-used types of kernel function, including the uniform, triangular, and Epanechnikov kernels. If $K(.)$ is unimodal then the heaviest weights are generally assigned to observations near the target x, and the least weights going to observations near the tails of the kernel function. The following moment conditions on the function $K(.)$ need to be satisfied

$$\int_{-1}^{1} K(u)\, du = 1, \tag{3.1}$$

$$\int_{-1}^{1} uK(u)\, du = 0, \tag{3.2}$$

$$\int_{-1}^{1} K(u)^2\, du < \infty. \tag{3.3}$$

Condition (3.1) is forcing the weights to sum to 1, under example (b) above the sum is exactly 1. Condition (3.2) is a symmetry condition that is automatically satisfied if $K(.)$ is symmetric about zero. Finally (3.3) is needed to ensure that the estimator has a finite asymptotic variance. The order of the kernel, κ, is defined by

$$\int K(u)u^j\, du = 0 \quad j = 1, \ldots, \kappa - 1, \quad \int_{-1}^{1} u^\kappa K(u)\, du = \alpha \neq 0; \tag{3.4}$$

kernels of order 2 are the most commonly used.

The weight function in (a) above has been used by Priestley and Chao (1972) in the fixed design setting. In a paper by Benedetti (1977) the asymptotics of this estimator and the optimal choices of the kernel function are considered. The weight function in (b) was introduced by Nadayara (1964) and Watson (1964) and is often referred to as the Nadaraya-Watson estimator. It is commonly used in the random design case. The weights sum to one exactly in this estimator, regardless of choice for K.

Finally, the weights in (c) were due to Gasser and Müller (1979), and were later generalised to estimating derivatives of regression functions in Gasser and Müller (1979, 1984). In these papers the authors also consider boundary value modifications in order to improve the estimation. Theoretical properties are discussed by Cheng and Lin (1981) and finite sample results by Gasser, Müller and Mammitzch (1985). Jennen-Steinmetz and Gasser (1988) is a useful reference on estimators closely related to those of Gasser and Müller above. Also Gasser and Engel (1990) discuss the advantages of using convolution type weights, such as these (see Clark, 1980 for more on convolution smoothing), in terms of minimax optimality, a property also discussed in Fan (1993).

The observations could have been generated in two ways.

1. The Random design setting where (x_i, y_i), $i = 1, \ldots, n$ are assumed independent, identically distributed random variables, joint density denoted by $f(x_i, y_i)$ and the marginal density of x_is denoted by $f(x_i)$. The regression curve is then defined by:

$$m(x_i) = E(Y_i|x_i).$$

2. The fixed design model where the variable is controlled, non-stochastic in nature, so

$$Y_i = m(x_i) + \epsilon_i.$$

   In many experiments the x_i are chosen to be equidistant on an interval $[a, b]$.

The weights (a) and (c) above are mostly used in the fixed design model. In the random design setting estimators based on these weights lead to a different

	Nadaraya-Watson	Gasser-Müller
Bias	$\frac{h^2(m''f+2m'f')(x)}{2f(x)}d_K$	$\frac{h^2m''}{2}d_K$
Variance	$\frac{\sigma^2(x)}{nhf(x)}c_K$	$\frac{3}{2}\frac{\sigma^2(x)}{nhf(x)}c_K$

$$\text{where } d_K = \int u^2K(u)du,$$
$$\text{and } c_K = \int K^2(u)du.$$

Table 3.1: Asymptotic bias and variance of Nadaraya-Watson and Gasser-Müller kernel smoothers

variance to that obtained by using weights in (b). Gasser and Engel (1990) give more detail on this.

Asymptotic properties of these estimators (i.e. as $n \to \infty$, $h \to 0$, $nh \to \infty$) are given in Table 3.1 for the more frequently used Nadaraya-Watson and Gasser-Müller estimators (Gasser and Engel, 1990 and Härdle, 1990a).

Table 3.1 pertains to a random design. Note that the bias of the Nadaraya-Watson estimator is of a more complicated form than that of the Gasser-Müller estimator; as well as involving derivatives of the underlying regression function the bias also involves the design density and its first derivative. However, the variance of the G-M estimator is 1.5 times that of the N-W estimator which is not particularly desirable.

Chu and Marron (1991) compare and contrast these two popular choices of kernel estimators, by presenting a balanced discussion of the differences between the two estimators in terms of their asymptotic bias and variance. They conclude that both methods have important advantages and disadvantages which need to be taken into consideration when deciding which approach to use. Jones, Davies and Park (1994) give more detailed comparisons.

These smoothers generally perform much worse at the boundaries, where fewer observations contribute, due to the asymmetry of the data, to the estimation at that point. This introduces what is known as *boundary bias* and modifications are normally required in order to remove it. Modified kernel estimators for boundary bias have been investigated by Gasser and Müller (1979) and Rice (1984a). Gasser and Müller consider modifying the kernel to give a modified boundary kernel, while Rice (1984a) uses a generalised jackknife approach. Jones (1993) also gives a review of boundary kernels used in density estimation. Kernels adjusted for the boundaries have not been considered further here. The problem of boundary bias is less of an issue if local lines are used instead of local means (Hastie and Loader, 1993). Local lines are discussed in the next sections and in Chapter 4.

All of the kernel smoothers described above could have been derived from the local likelihood estimating equations. We look at this is more detail when we consider alternative estimators of the finite population total in Chapter 4.

3.4.2 Nearest neighbour functions

The construction of the estimators based on nearest neighbours differs from that of kernel estimators in that kernel estimators involve calculating a weighted average for a fixed bandwidth around x, while the nearest neighbour estimators are based on a weighted average with a varying bandwidth or neighbourhood. If we select an equal number of points, say $(k-1)/2$, to the left and right of x including x itself, i.e. a span of k points in all, then this is known as a symmetric nearest neighbourhood. The neighbourhood is 'nearest' in the sense that equal numbers of points either side of the target x are included in the span of points, based on their Euclidean distance from x. We may also chose the $(k-1)$ nearest x_i, $i = 1, \ldots, n$, values to x in terms of Euclidean distance, not necessarily symmetric; these are simply known as nearest neighbours. If an x value is not one of the sample values then an arbitrary choice may be to take the nearest sample value to it in terms of Euclidean distance. The smoothing parameter is the span k and controls the smoothness of the resulting estimate. Assuming the data are sorted by increasing x_i, a formal definition of the symmetric nearest neighbourhood is

$$N_i = \{\max(1, i-(k-1)/2), \ldots, \min(n, i+(k-1)/2)\}$$

and the k-NN smoother is defined as

$$\hat{m}(x_i) = \sum_{j \in N_i} y_j.$$

In an experiment where the x_i are from an equidistant grid the k-NN estimator is equivalent to a kernel approach using a uniform kernel function. The extreme case when $k = 1$ produces a regression function which jumps in a step-function fashion between two adjacent observations. The smoother fits each data point exactly. This running mean smoother produces a curve more jagged in appearance than that produced by a kernel regression smoother. However both curves have boundary bias problems because of the nature of the smoother.

3.5 Nonparametric regression smoothing techniques - Local regression

3.5.1 Locally weighted regression

Local regression was traditionally used for smoothing time series and scatterplots, Cleveland (1979) and later Cleveland and Devlin (1988) developed further aspects of this method by proposing a locally weighted regression, extending ordinary least squares to weighted least squares, and robustifying by iteratively reweighting. This method then combines the strict nature of running lines with the smooth kernel function weights and is often referred to as the locally weighted running line smoother (LOWESS).

Locally weighted regression smoothers can be extended to any class of local polynomial. Fan, Heckman and Wand (1992) consider a class of locally fitted generalised polynomials in which

$$\mu(\beta) = g^{-1}(\beta_1 + \beta_2 x + \cdots + \beta_p x^p)$$

for some specified link function g. These local polynomials are found to have some good properties including reduced bias near the boundaries of x.

A recent paper by Hastie and Loader (1993) also highlights this fact, by comparing these 'LOWESS' smoothers with the kernel smoothers. They find that the local regression smoothers adjust bias without the modifications of the kernel function. See also Fan and Gijbels (1992).

3.5.2 Running line smoother

The other type of k-NN approach we consider is that based on local regression, as described in Section 3.2 above. Suppose we have local linear regression within the k-NN setting.

Instead of averaging the response value in a neighbourhood around each x_i, the running line approach fits a least squares line to the data points in a neighbourhood. The value of the fitted line at x gives the smooth estimate and this is performed for each x in the population. This type of smoother reduces the end-point bias. However, the output of the running line also appears jagged because the weights at x are zero outside the neighbourhood.

Computationally the running mean and line smoothers are more efficient than corresponding kernel methods as the estimate at any x can be defined recursively, based on the estimate at the previous x value. This updating formula can be applied to any local polynomial fit.

Again these nearest-neighbour smoothers can be derived from the local likelihood estimating equations as examples of local least squares estimators; the running mean is obtained by assuming only an intercept term in the linear predictor, and the running line smoother by assuming a model with linear and intercept terms and an identity link. The parameters of the local fit are estimated and the value of the fitted line at x gives the smooth estimate.

3.5.3 Regression splines

These consist of less rigid forms of parametric fitting which are closer in spirit to the kernel smoothers mentioned above. Polynomial regression has limited appeal due to the global nature of the fit. It is often beneficial to work with polynomials of lower degree and divide the interval of interest into smaller pieces.

Regression splines represent the fit as local piecewise polynomials. Regions specifying the pieces are separated by knots or breakpoints (k knots). In addition it is useful if we force the piecewise polynomials to join smoothly at the knots, instead of having a discontinous function with jumps at the knots. Popular choices have been piecewise cubic polynomials.

The smooth at a point is computed by multiple regression on an appropriate set of basis vectors, for a given knot sequence. Basis vectors are functions representing the particular family of, say piecewise cubic polynomials, evaluated at the observed values. To smooth the data pairs (x_i, y_i) we would construct a regression matrix with $k + 4$ columns each corresponding to a function $A_j(x)$, linearly independent, evaluated at the n values of x. The smooth is taken as a linear combination of the $A_j(x)$ at a point x.

There are various forms for the basis functions representing the splines, the simplest is the truncated power series, however a B-spline basis provides a numerically superior alternative to this. A spline of order r with knots at $\xi_1, \ldots, \xi_k$ can be defined as any function of the form

$$s(t) = \sum_{i=0}^{r-1} \theta_i t^i + \sum_{i=1}^{k} \delta_i (t - \xi_i)_+^{r-1},$$

where + denotes the value if positive and 0 if negative or zero. This is equivalent to the specification that

1. s is a piecewise polynomial of order r on any sub-interval $[\xi_i, \xi_{i+1})$,

2. s has $r - 1$ continuous derivatives and

3. s has an $(r-1)$st derivative that is a step function with jumps at $\xi_i, \ldots, \xi_k$.

Let $s^r(\xi_1, \ldots, \xi_k)$ denote the set of all functions of the above form. This forms a *basis* of dimension $k + r$.

B-Splines

The B-splines are themselves piecewise polynomials and we need k+(degree)+1 of them if we want to span the space. Their algebraic definition can be written in terms of divided differences of Green's functions and is referred to in more detail in Schumaker (1981) and de Boor (1978). B-splines are usually defined recursively and as a result cannot be calculated directly.

The difficulty with using splines is in choosing the number and position of knots. Several methods for this are known. We can select the number of knots for each estimator and then place them uniformly over the range of values, or we could place the knots at quantiles of the predictor variable. The number of knots, k, is related to the degrees of freedom of fit, to be introduced in Section 3.7, so that if we fix the degrees of freedom we can determine k.

Number and position of knots

Choosing the number and position of knots is similar to choosing a smoothing parameter in nonparametric regression. The number of knots to be chosen can be derived directly by fixing the degrees of freedom of the fit. Increasing the degrees of freedom increases the number of knots, and the regression curve becomes more wiggly. Decreasing the degrees of freedom decreases the number of knots and the curve is more smooth. The important factor here is the placement of knots.

We discuss linear B-splines, defined as divided differences of truncated power functions as follows:

$$A_j(x_i) = \sum_{m=j}^{k} \left[\prod_{l=j, l \neq m}^{k} \frac{1}{(\xi_l - \xi_m)} \right] (x_i - \xi_m)_+^k,$$

where ξ_j is the jth knot for $j = 1, \ldots k$. The basis functions can also be represented in linear form as

$$A_j(x_i) = \sum_{j=1}^{d} \sum_{l=1}^{l_j} \alpha_{jl}(x_i - \xi_j)_+^{k-l}.$$

Regression splines are incorporated as a component of the linear predictor in a generalised model in Chapter 4 when we introduce 'model-based' estimators.

3.5.4 Penalised likelihood

The smoothing spline is an example of a smoother which minimises a penalised least squares criterion. A natural measure of fidelity to the data for a curve m is the residual sum of squares

$$\sum_{i=1}^{n} (y_i - m(x_i))^2 .$$

The spline introduces a term which penalises too much local variation in m. One convenient measure is the square of the L_2 norm of the second derivative of m, and the spline smoother is defined as the minimiser of

$$S_\lambda(m) = \sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int \left(m''(x)\right)^2 ,$$

where $\lambda > 0$ controls the smoothness. Penalised least squares can also be generalised to penalised likelihood (see Green and Silverman, 1994). A relation between kernel regression and spline smoothing is detailed in Silverman (1984), and Härdle (1990a, §3.4); other references on splines include Wahba (1975, 1990), Rice and Rosenblatt (1983), Eubank (1988), the review paper by Wegman and Wright (1983) and more recently Green and Silverman (1994). Smoothing splines are not considered in much detail in this thesis, but we do consider another class of splines already described, regression splines, when we introduce model-based estimators of the finite population total in Chapter 4.

3.5.5 Other smoothing methods

The methods discussed by Müller (1987) are in a similar vein to kernel methods already described. Instead of assuming a fixed bandwidth, the bandwidth is allowed to vary in proportion to the underlying design density to a power of $-\alpha$, $0 < \alpha \leq 1$.

Other smoothers include orthogonal series estimators, the regressogram, convolution smoothing, median smoothing and many others. For a more comprehensive list see Härdle (1990a) or the review article by Collomb (1981).

3.6 Smoother matrices and equivalent kernels

A linear smoother is special in that $\hat{\mathbf{y}}$ can be written in the form $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ where $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{S}$ is an $n \times n$ *smoother matrix*, which depends on the x_i and the smoothing parameter b. All of the above smoothers are examples of linear smoothers of y. Because of their independence from $\mathbf{y}$ linear smoother matrices have some useful properties, some of which are discussed in Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990).

One way to compare various smoothers is in plots of rows of their smoother matrices against x. These are known as plots of *equivalent kernels*, and help determine which x values are having an influence on the fit at the target value. Essentially they give the form of the neighbourhood and the weighting scheme at any target x value.

3.7 Degrees Of Freedom

In parametric linear regression, we have the notion of *degrees of freedom* associated with the fit to the data. It would be useful to have a similar measure for nonparametric smooth fits to the data. Buja, Hastie and Tibshirani (1989) gave three definitions of degrees of freedom analagous to the linear regression case:

1. degrees of freedom=tr($\mathbf{S}\mathbf{S}^T$).
2. degrees of freedom=tr($2\mathbf{S} - \mathbf{S}\mathbf{S}^T$).
3. degrees of freedom=tr($\mathbf{S}$).

The motivation for the first method comes about because, for the linear model,

$$\sum_{j \in s} \operatorname{var}(\hat{y}_j) = p\sigma^2$$

where p is the degrees of freedom or number of parameters fit, tr($\mathbf{S}\mathbf{S}^T$). The more parameters we fit, the rougher the function and higher its variance.

The second form is motivated by the expectation of the residual sum of squares:

$$\text{RSS} = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$$

which has expectation

$$E(\text{RSS}) = [n - \operatorname{tr}(2\mathbf{S} - \mathbf{S}^T\mathbf{S})]\sigma^2 + \mathbf{m}(\mathbf{x})^T(\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S})\mathbf{m}(\mathbf{x}).$$

The last term measures squared bias. The first term defines

$$df_{err} = n - \operatorname{tr}(2\mathbf{S} - \mathbf{S}^T\mathbf{S}),$$

since this is $n - p$ in linear regression. If we were smoothing noise, then $\sigma^2(n - \text{df}_{err})$ is the expected drop in RSS due to overfitting. Hence Buja, Hastie and Tibshirani (1989) are motivated to use tr($2\mathbf{S} - \mathbf{S}^T\mathbf{S}$).

Finally the third method of tr($\mathbf{S}$) can be motivated as the Mallows (1973) C_p correction for RSS, to make it unbiased for the PMSE by adding the quantity $2\operatorname{tr}(\mathbf{S})\hat{\sigma^2}/n$. This definition of degrees of freedom is popular in smoothing spline literature, when $\mathbf{S}\sigma^2$ emerges as the posterior covariance of $\hat{y}$ after appropriate Bayesian assumptions are made.

Of the three definitions df $=$ tr($\mathbf{S}$) is the easiest to compute, since it only requires finding diagonal elements of $\mathbf{S}$. This provides a reasonable way of calibrating smoothing parameters across a class of smoothers, and gives a choice without using alternative methods of smoothing parameter selection (see Chapter 5).

Hastie and Tibshirani (1989, 1990) also use these notions of degrees of freedom when comparing two fits, although this is not discussed further here. For the nonparametric regression estimators to be introduced, an alternative method for computing the degrees of freedom is given in the next chapter. This is specific to the context of predicting a population total, and is motivated in a similar way to Hastie and Tibshirani's methods. Finally, the degrees-of-freedom forms described above depend on the explanatory values and the smoothing parameter; the smoothing parameter is the major determinant for the degrees of freedom.

3.8 Multiple regression and generalised additive models

The methods described so far assume only one explanatory variable for the response Y. We may be interested in considering several explanatory variables simultaneously, in a multiple regression context. Generalised additive modeling is multiple regression using smoothing techniques so that dependence on each explanatory variable is described by a smooth curve and the sum of these smooth functions gives us the 'additive model'. This can be seen as a further generalisation of the generalised linear model due to Nelder and Wedderburn (1972). The additive model may be written as

$$\eta_i = \sum_{j=1}^{d} m_j(x_{ij}),$$

where $m_j(x_{ij})$ are separate nonparametric regression functions. This is referred to as a generalised additive model (GAM) with a specified link function.

Another multivariate smoothing technique is projection pursuit (Friedman and Stuetzle 1981).

We concentrate on the one explanatory variable case in this thesis, but many of the methods described could be extended to a multivariate setting also.

3.9 Bibliography

Introductory references that have been useful include Eubank (1988), covering material on kernel smoothers and in particular smoothing splines. Also Härdle (1990a, 1990b) covers nonparametric regression in some detail with particular reference to kernel smoothers. Silverman's (1985) paper on smoothing splines gives an introduction with reference to some applications of these methods. The forthcoming book by Jones and Wand (1994) is an introduction to kernel density and regression estimation.

The review paper on nonparametric regression by Collomb (1981) includes an introduction, the regressogram, kernel and analogous methods, splines, sequential methods, use in discriminant analysis and general comments. An earlier paper by Stone (1977) also introduces nonparametric regression.

Monographs by Eubank (1988), Müller (1987), Härdle (1990a, 1990b) and Wahba (1990) give a large variety of interesting real data examples. Green and Silverman (1994) concentrate on the roughness penalty approach to nonparametric regression.

Linear smoothers and the properties of their smoother matrices are discussed in some detail in Buja, Hastie and Tibshirani (1989) and again in Hastie and Tibshirani (1990), as are other topics such as the degrees of freedom of a smoother. Two papers by Fan (1992, 1993) look at local lines in more detail. Extension of smoothing ideas to multivariate data has been covered by Hastie and Tibshirani (1990) and Tibshirani and Hastie (1987) using generalised additive models.

Chapter 4

Smoothing in predicting a finite population total

4.1 Introduction

In the first part of this chapter we focus attention on a class of alternative estimators of the finite population total to those described in Chapter 2. The alternative estimators are based on 'local' quasi-likelihood estimating equations and are examples of nonparametric regression estimators. These can be defined as either *total preserving* or as *non-total-preserving* estimators, a property which is discussed further in Section 4.3. The nonparametric regression estimators that are proposed are often more efficient than existing parametric methods. The latter part of the chapter is devoted to properties of these estimators under a superpopulation model. Design-based properties are given in Chapter 6.

4.2 Motivation and derivation of some alternative estimators

In Chapter 3, nonparametric regression was introduced in its simplest form as a weighted mean, with reference to the Nadaraya-Watson estimator. This estimator can be derived from an estimating equation such as

$$\sum_{j=1}^{n}(y_j - \mu(x))W_b(x, x_j) = 0,$$

an example of a 'local likelihood' equation as discussed in Tibshirani and Hastie(1987). These ideas can extend to various kinds of data, whose relationship is expressible through a likelihood function. Here, a simple parametric function like $\beta_0 + \beta_1 x_j$ appearing in the likelihood is replaced with a smooth function, $\mu(x)$, and $\mu(x)$ is estimated locally. When the mean and variance are specified, 'local' quasi-likelihood estimating equations are:

$$\sum_{j=1}^{n} \frac{(y_j - \mu_j)W_b(x, x_j)x_{jr}}{V(\mu_j)g'(\mu_j)} = 0 \qquad (r = 1, \ldots, p),$$

where μ_j stands for $\mu_j(\hat{\beta}(x))$, and $\hat{\beta}(x)$ is a different coefficient vector for each $\mathbf{x}$, obtained by solving the above estimating equations. Here, $W_b(x, x_j)$ is a weight derived from a suitable kernel function.

This method is used in deriving a new class of estimators which are examples of *operational* estimators. They are operational in the sense of being automatic estimators except, perhaps, for the choice of a smoothing parameter. The automatic nature of these estimators is achieved through the local averaging, so once we decide which weights to use and the type of estimator, i.e. a local regression or ratio estimator, etc., the estimator computes the predicted finite population total. Local regression involves fitting a local curve with attached kernel weights. Local here means within a neighbourhood of a target x. Important types of operational estimators include the following:

1. The locally weighted ratio estimator (abbreviated to LWRE). A single explanatory variable, no intercept, identity link and variance proportional to mean (x_i in this case) are defined locally as:

$$E(y_i) = \mu_i = x_i\beta(x), \qquad \text{var}(y_i) = \sigma^2 x_i \quad (\sigma^2 \text{ constant}).$$

The estimating equation for $\beta(x)$ is then

$$\sum_{j=1}^{n} \frac{(y_j - x_j\beta(x))W_b(x, x_j)x_j}{x_j} = 0,$$

which leads to

$$\hat{\beta}(x) = \frac{\sum_{j\in s} W_b(x, x_j)y_j}{\sum_{j\in s} W_b(x, x_j)x_j}$$

as the explicit solution of the local quasi-likelihood equations. The predicted value is then $\hat{y} = x\hat{\beta}(x)$ and thus estimation of the population total is given by

$$\hat{T} = \sum_{i=1}^{N} \hat{y}_i = \sum_{i=1}^{N} x_i \hat{\beta}(x_i).$$

This new estimator is just a locally weighted version of the familiar ratio estimator. In this particular case, with mean and variance specified as above, weighted least squares is equivalent to Poisson maximum likelihood and so local least squares estimation is local likelihood estimation.

2. The locally weighted regression estimator. Simple linear regression with an intercept, identity link and constant variance are defined locally as:

$$E(y_i) = \mu_i = \beta_0(x) + \beta_1(x)x_i, \qquad \mathrm{var}(y_i) = \sigma^2.$$

Now the estimating equations are

$$\sum_{j=1}^{n}(y_j - (\beta_0(x) + \beta_1(x)x_j))W_b(x, x_j) = 0,$$

and

$$\sum_{j=1}^{n}(y_j - (\beta_0(x) + \beta_1(x)x_j))x_j W_b(x, x_j) = 0.$$

Solving these equations simultaneously is similar to solving the estimating equations for the parametric regression case, except now weights have been included, and each parameter estimate is a local parameter. Fan, Heckman and Wand (1992) use an approach similar to this based on local polynomial regression within a generalised linear modelling framework, but not for prediction purposes. In their paper, Fan, Heckman and Wand (1992) develop asymptotic theory and discuss bandwidth selection in terms of the 'plug-in' approach, to be discussed further in Chapter 5.

Other types of locally weighted estimator can be derived if we know the mean and variance functions, since any quasi-likelihood is entirely specified by these functions. A thorough account of quasi-likelihood inference is given in McCullagh and Nelder (1989, Chapter 9). For easy implementation of such estimators, we must ensure the parameter estimation is non-iterative, so that the parameters can be explicitly defined as in the two examples given above. This, however, restricts the class of models we can use.

Some examples of weight functions we consider are the Gaussian kernel function

$$W_b(x_i, x_j) = \exp\left(-\frac{1}{2}\left(\frac{x_i - x_j}{b}\right)^2\right)$$

and the Uniform k-nearest neighbour function, where k is the span

$$W_b(x_i, x_j) = \begin{cases} 1, & |i - j| < k \\ 0, & \text{otherwise.} \end{cases}$$

The x_i are assumed ordered. Any estimator based on the Uniform k-nearest neighbour function is known as a *running* estimator. For example, if the Nadaraya-Watson estimator had been used with this weight we would have a running mean. If the locally weighted ratio estimator had been used, we would have a running ratio estimator, etc.

The nearest neighbour estimators are particularly useful since they have the ability to adapt their bandwidth to the local density of the predictor variable, something the fixed-bandwidth kernels cannot do. This may be overcome by introducing a variable bandwidth into the kernel function which is allowed to vary with the local density and possibly other factors, such as local variance (see Chapter 5).

4.3 Total-preserving estimators

4.3.1 Introduction

For the purpose of estimating a population total only non-sample values are predicted, since the sample total is already known. We write

$$\hat{T} = T_s + \hat{T}_r,$$

where T_s is the known sample total and $\hat{T}_r$ is an estimator of the unknown non-sample total.

Examples of operational estimators have already been introduced. Although these have the property of being automatic estimators, not all possess the additional property of being 'total-preserving'. Next a definition of 'total-preserving' is given. Recall the definition of the smoother matrix in relation to the fitted values:

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}.$$

Definition 4.1 *If* $\mathbf{1}^T\hat{\mathbf{y}}_s = \mathbf{1}^T\mathbf{S}\mathbf{y}_s = \mathbf{1}^T\mathbf{y}_s$*, then* $\mathbf{S}$*, the linear smoother matrix, is said to be 'total-preserving'. This smoother is then associated with a 'total-preserving' estimator.*

Estimators with this property include the expansion estimator, ratio estimator and combined and separate ratio estimators. It becomes a useful property when properties of these estimators under repeated sampling are considered in Chapter 6. There it is shown that estimators violating this property can have serious bias problems under repeated sampling. Total-preserving estimators are shown to have bias approximately zero. Two types of 'total-preserving' estimator are described in Section 4.3.2, these being operational and model-based estimators respectively.

Estimators which are not total-preserving include the operational type already mentioned, with weight functions such as the Gaussian kernel or Uniform k-nearest neighbour. They can still be more efficient than existing parametric methods because of the implicit underlying smooth function. Estimators which are not total-preserving also have the property of being automatic estimators and are easily interpreted as locally weighted estimators.

4.3.2 Total-preserving operational estimators

These estimators involve taking the average of k running total-preserving estimators at x_i. Examples of estimators we may consider in the average running estimator include means, ratio estimators and linear regression estimators. For example, the averaged running ratio estimator is

$$\hat{T}_{ARRE} = \sum_{i=1}^{N} x_i \frac{1}{k}\left(\hat{\beta}_{i-k+1}(k) + \cdots + \hat{\beta}_i(k)\right)$$

where

$$\hat{\beta}_i(k) = \frac{\sum_{l(i)}^{l(i)+(k-1)} y_j}{\sum_{l(i)}^{l(i)+(k-1)} x_j},$$

such that $l(i) = \{m : min|x_i - x_m|, m \in s\}$. The span k, is the smoothing parameter associated with this estimator.

All of the estimators mentioned so far can be written in terms of their smoother or 'hat' matrix. Smoother matrices are introduced in Section 3.5. In the case of the averaged running estimators, we can write

$$\hat{y}_i = x_i \frac{1}{k} \sum_{j=i-k+1}^{i} \hat{\beta}_j(k) = \frac{1}{k} \sum_{j=i-k+1}^{i} \mathbf{S}_{ij}\mathbf{y}$$

where $\mathbf{S}_{ij}$ is the jth row of the smoother matrix, $\mathbf{S}_i$ with the following total-preserving property, $\frac{1}{k}\sum_{i=1}^{n} \mathbf{1}_\mathbf{k}^\mathbf{T}\mathbf{S}_{i}\mathbf{y} = \mathbf{1}_n^T\mathbf{y}\ \forall\mathbf{y}$. Here $\mathbf{S}_i$ is a $k \times n$ smoother matrix for each target x_i, and $\mathbf{1}_k$, $\mathbf{1}_n$ are vectors of 1's of length k and n respectively and $\mathbf{y}$ a vector of sample y values. For example, the jth row of $\mathbf{S}_i$ for the averaged running estimator is

$$\mathbf{S}_{ij} = \left(0, \ldots, 0, S_{(i,l(i)-k+j)}, \ldots, S_{(i,l(i)-k+j)}, 0 \ldots, 0\right),$$

where the first nonzero value is at position $\min(l(i)-k+j, 1)$ and the last element at position $\max(l(i)+j-1, n)$, for $j = 1, \ldots, k$. Each element $S_{(i,l(i)-k+j)}$ in the vector above is an estimate of $\hat{y}_i$ at a target x_i based on elements in the window:

$$\min(l(i)-k+j, 1) \quad \text{up to } \max(l(i)+j-1, n).$$

The averaging may be extended to running generalised linear models provided a canonical link is used and an intercept term included in the linear predictor. This ensures the total-preserving property holds since

$$\sum_{j=1}^{n}(y_j - \mu_j).1 = 0.$$

The averaged running estimator is always total-preserving, provided the running estimator considered is total-preserving itself. This is shown next. Now

$$\hat{T}_{ARRE} = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{n}\mathbf{S}_{ij}\mathbf{y},$$

where $\mathbf{S}_{ij}$ is as defined above. If we fix j and sum over the i then we have

$$\sum_{j=1}^{n}\mathbf{S}_{ij}\mathbf{y} = \mathbf{S}_{.j}\mathbf{y}.$$

If $\mathbf{S}_{.j}\mathbf{y} = \mathbf{1}^T\mathbf{y}$, i.e. the smoother matrix used in the *running* estimator is total-preserving, then it follows that averaging over all $j = 1, \ldots, k$ will also be total preserving.

For the running mean it is noted that, at the boundary, there may be some bias because of the nature of the estimator. This is rectified to a certain extent by using local lines instead of local means. The averaged running mean estimator is not a good estimator of the finite population total because of this boundary bias.

The averaged running estimators have some good properties relative to parametric estimators. They compare favorably with the separate ratio estimator but often have less variability (see Table 5.1 and Tables 6.1-6.3).

One question not addressed here but covered further in Chapter 5 is the choice of smoothing parameter or span in this case. Since a smoother matrix can be defined for these estimators, it is possible to find a value of the span relating to some value of the degrees of freedom (or number of strata if stratification is used). Other methods such as crossvalidation are computationally cumbersome in this case.

4.3.3 Model-based estimators

The emphasis here is more on the underlying superpopulation model. Generalised linear models, as described by McCullagh and Nelder (1989), are now considered in order to find a class of model-based estimators. A large class of total-preserving generalised linear model estimators exist. For example, a model for y_i with variance proportional to the mean and regression through the origin can be represented as :

$$E(Y_i) = \log(\mu_i) = \log(x_i) + \beta_0 + P(x_i),$$

$$Var(Y_i) = \sigma^2 \mu_i,$$

where $\log(x_i)$ is an offset term, β_0 the intercept, and $P(x_i)$ may be any function, e.g. a polynomial or regression spline at x_i. With just the intercept in the model the usual ratio estimator results. A model with a linear explanatory variable and intercept term is entirely parametric. Nonparametric regression components are introduced in $P(x_i)$ to make the model semiparametric, for example, $P(x_i)$ as a regression spline. Green and Yandell (1987) describe semiparametric generalised linear models in more detail. Regression splines are discussed in Section 3.4.

Including regression splines

Changing the knot sequence in the regression spline makes no difference to the total-preservation of the estimator, since this will just lead to a different set of basis functions from which to compute the B-spline. The total-preserving property holds because of the fact that we include an intercept term in our linear predictor and not because of the form of the regression spline.

In this work B-splines were fitted, initially based on fixing the degrees of freedom of the fit, and choosing the knots at appropriate equally-spaced quantiles of the x distribution. When stratification was used, this led to estimators that consistently under-predicted the true population total. The stratum boundaries were found to be a more appropriate knot sequence. The reason for this is briefly outlined below. If we recall the definition for the B-spline basis functions is :

$$A_j(x_i) = \sum_{m=j}^{k} \left[\prod_{l=j,l\neq m}^{k} \frac{1}{(\xi_l - \xi_m)} \right] (x_i - \xi_m)_+^k,$$

where ξ_j is the jth knot for $j = 1, \ldots, k$ and for x_i sorted in ascending order, $i = 1, \ldots, N$. The basis functions can also be represented in linear form as

$$A_j(x_i) = \sum_{j=1}^{d} \sum_{l=1}^{l_j} \alpha_{jl}(x_i - \xi_j)_+^{k-l}.$$

A matrix of the basis funtions can be constructed as:

$$\mathbf{A} = \begin{pmatrix} A_{11} & \ldots & A_{11} & 0 & 0 & \ldots & 0 \\ A_{12} & \ldots & A_{12} & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{1N_{\xi_1}} & \ldots & A_{1N_{\xi_1}} & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & A_{k1} & \ldots & A_{k1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & 0 & A_{kN_{\xi_k}} & \ldots & A_{kN_{\xi_k}} \\ & & & & & & \end{pmatrix},$$

where

$$A_{ji} = \sum_{m=j}^{k} \left[\prod_{l=j,l\neq m}^{k} \frac{1}{(\xi_l - \xi_m)} \right] (x_i - \xi_m)_+^k$$

for $j = 1, \ldots, k$ and $i = 1, \ldots, N$. The structure of the B-spline basis function is very similar to that of the smoother matrix of the separate ratio estimator (or any separate estimator). It appears, therefore, that a natural choice for the knots of the regression spline are the stratum boundaries, if stratified SRS is employed. The main difference between the separate estimator and the piecewise polynomials is what happens at the boundaries. The piecewise polynomials are constrained to join smoothly at the knots where the separate estimator is derived from a function discontinuous at the boundaries.

If the sample data had been chosen using optimal stratification then assigning equally spaced knots at the quantiles of the x distribution will generally not ensure optimal results. The units in the sample will have been chosen such that the larger the variation within a stratum the larger the n_h allocated to that stratum. If the knots (boundaries) are defined elsewhere the resulting estimator will be sub-optimal. This is particularly true if the variation in the data increases with the x values, say. Optimal stratification will assign fewer observations to the first stratum and more to the last. Fixing the stratum so that equal numbers of observations are in each will increase the numbers of observations in the first stratum and decrease the numbers in the last stratum, where they are needed most.

If the knot sequence is chosen as the boundaries used in stratified simple random sampling then the above smoother matrix is very similar to the stratified ratio estimator, or any separate estimator, smoother matrix. The position of the knots is, therefore, very important particularly if a design other than simple random sampling, for example, stratified simple random sampling is used. An example to illustrate this is given below. When simple random sampling was considered, the position of the knots was less restrictive, and the estimator based on a regression spline with knots chosen at equally-spaced quantiles of the x distribution performed satisfactorily.

The total-preserving property, for the model-based estimator with linear B-spline, does not hold within strata, as it does for the separate ratio estimator. This is because any set of linear basis functions used leads to a particular set of parameter estimates (using IRLS) defining the piecewise polynomial, and there will be only one such polynomial. This unique function does not ensure the within-strata total-preservation holds. The way in which the parameters are estimated could be modified, by constraining on the residuals within each stratum

Estimator	Mean	St. dev.	MSE/10^7
Ratio estimator	314095	6612	8.023
Stratified ratio estimator	319863	6692	4.486
GLM with B-spline (knots=equal)	310860	6402	12.70
GLM with B-spline (knots=strata)	321875	6309	4.282

Table 4.1: Table comparing estimators including the GLM with B-spline estimator based on equal and stratum boundary placed knots.

summing to zero. This might be performed using a penalised likelihood approach, possibly solving iteratively using a scheme such as in Green (1985). This has not been pursued further here.

Example 1

To illustrate the advantage of using stratum boundaries over quantiles an example is included.

Data is taken from the hospital dataset of Section 1.4. From the population of 393 observations 100 stratified simple random samples of size 100 were selected optimally based on 3 strata. The true population total is 320139.

Table 4.1 gives the means, standard deviations and mean squared errors of the ratio estimator, separate ratio estimator, GLM with regression spline based on equally spaced knots and knots placed at the stratum boundaries, over the 100 samples. Figure 4.1 has also been included to show how the predicted totals compare with the true total (solid line) in all the estimators mentioned.

The ratio estimator and GLM with B-spline based on equally placed knots perform the worst in terms of minimising the design mean squared error. They are both more variable than the GLM based on knots at the stratum boundaries and both underestimated the true population total quite considerably. The stratified ratio estimator is the next best, with the predicted values closer to the true value but with a larger overall variation. The GLM with stratum boundary knots performs the best, getting close to the true population total and with reduced variability on the stratified ratio estimator. This confirms the importance of appropriate knot placement for the sampling design in question when the GLM with regression spline estimator is being used.

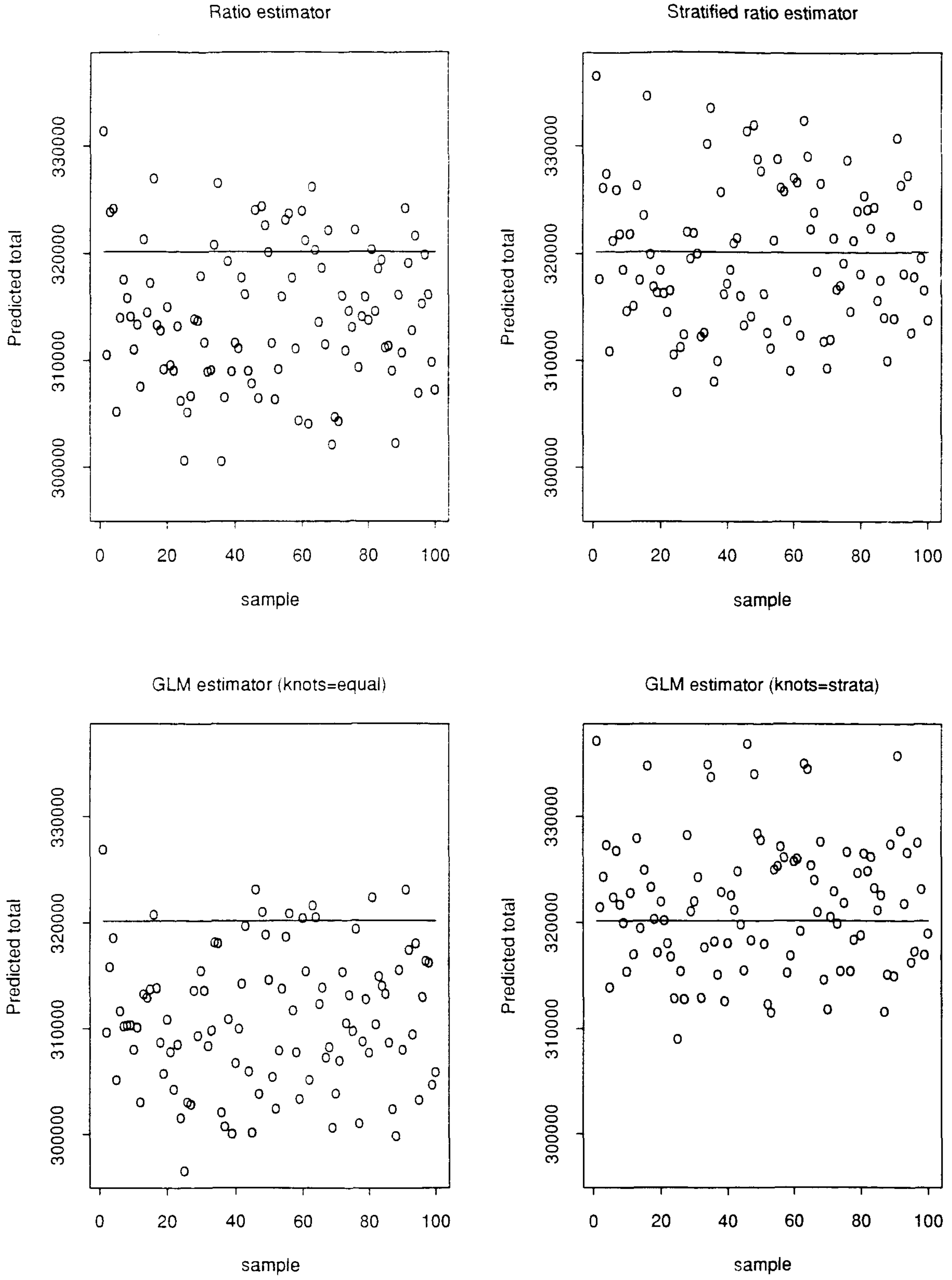
4.4 Properties of some nonparametric regression estimators

The properties considered involve taking expectations with respect to some underlying superpopulation model ξ. Properties under the other, design-based, approach are deferred until Chapter 6. The superpopulation model is assumed as follows:

$$Y_i = m(x_i) + \epsilon_i,$$

where ϵ_i are independent random errors with $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2(x_i)$. Using the known data a smooth curve, $m(x_i)$, can be traced. Our estimators utilise this smooth curve by interpolating (or extrapolating) the smooth to the unknown sample values.

Interest is in the bias, variance and predictive mean squared error under the superpopulation model. These are given below for the operational type of estimator mentioned in Section 4.2, and are summarised for the other estimators in Tables 4.2-4.5 below. Similar properties for the ratio and separate ratio estimators are found in Royall and Herson (1973a, b).



Ratio estimator

Stratified ratio estimator

GLM estimator (knots=equal)

GLM estimator (knots=strata)

Figure 4.1: Plot of predicted totals for 100 samples based on the ratio estimator, separate ratio estimator and GLM with B-spline (equally placed knots and knots at the stratum boundaries)

An example of how the bias, variance and mean squared error are derived for the averaged running ratio estimator is given below. The tables of results for the other nonparametric regression estimators are derived in a similar way.

Bias under a superpopulation model The averaged running ratio estimator is written as:

$$\hat{T}_{ARRE} = \sum_{i=1}^{N} x_i \sum_{m=1}^{k} \frac{1}{k} \frac{\sum_{j=1}^{n} W_{ijm} y_j}{\sum_{j=1}^{n} W_{ijm} x_j},$$

for some weight function W_{ijm}.

Under the general superpopulation model ξ, the bias can be written as

$$E_\xi\left(\hat{T}_{ARRE} - T\right) = E_\xi\left[\sum_{i=1}^{N} x_i \hat{\beta}(x_i) - \sum_{i=1}^{N} y_i\right]$$

$$= \sum_{i=1}^{N} x_i \sum_{m=1}^{K} \frac{1}{k} \frac{\sum_{j\in s} W_{ijm} m(x_j)}{\sum_{j=1}^{n} W_{ijm} x_j} - \sum_{i=1}^{N} m(x_i).$$

Variance The variance is given by

$$\begin{aligned}
\mathrm{var}_\xi(\hat{T}_{ARRE}) &= \mathrm{var}_\xi\left(\sum_{i=1}^{N} x_i \sum_{m=1}^{k} \frac{1}{k} \frac{\sum_{j=1}^{n} W_{ijm} y_j}{\sum_{j=1}^{n} W_{ijm} x_j}\right) \\
&= \mathrm{var}_\xi\left(\sum_{j=1}^{n} \sum_{i=1}^{N} \frac{x_i}{k} \sum_{m=1}^{k} \frac{W_{ijm} y_j}{\sum_{j=1}^{n} W_{ijm} x_j}\right) \\
&= \sum_{j=1}^{n} \sigma^2(x_j) \left\{\sum_{i=1}^{N} \frac{x_i}{k} \sum_{m=1}^{k} \left(\frac{W_{ijm}}{\sum_{j=1}^{n} W_{ijm} x_j}\right)\right\}^2 .
\end{aligned}$$

For example, when $\sigma^2(x_j) = \sigma^2 x_j$ and $b = \infty$, then $W_{ijm} = 1$ for all i, j and we have the ratio estimator. Then

$$\mathrm{var}_\xi(\hat{T}_{RE}) = \sigma^2 \frac{(\sum_{i=1}^{N} x_i)^2}{\sum_{j\in s} x_j}.$$

Mean squared error

The mean squared error is

$$\begin{aligned}
\mathrm{MSE}(\hat{T}) &= E_\xi(\hat{T} - \sum_{i=1}^{N} m(x_i))^2 \\
&= \mathrm{var}_\xi(\hat{T}) + (E_\xi(\hat{T} - T))^2.
\end{aligned}$$

The predictive mean squared error can be written as :

$$\text{PMSE}(\hat{T}) = E_\xi(\hat{T} - T)^2. \tag{4.1}$$

A general relationship between the PMSE and the MSE is given below for all ratio estimators considered. Recall that

$$\begin{aligned}
\text{PMSE}(\hat{T}) &= E_\xi(\hat{T} - T)^2 \\
&= E_\xi \left(\sum_{i \notin s} x_i \frac{\sum_{j=1}^n W_b(x_i,x_j)y_j}{\sum_{j=1}^n W_b(x_i,x_j)x_j} - \sum_{i \notin s} y_i \right)^2 \\
&= E_\xi \left(\left(\sum_{i \notin s} \frac{x_i}{\sum_{j=1}^n W_b(x_i,x_j)x_j} \sum_{j=1}^n W_b(x_i, x_j)y_j \right)^2 \right. \\
&\left. -2 \sum_{i \notin s} \frac{x_i}{\sum_{j=1}^n W_b(x_i,x_j)x_j} \sum_{j=1}^n W_b(x_i, x_j)y_j \sum_{i \notin s} y_i + \left(\sum_{i \notin s} y_i\right)^2 \right)
\end{aligned}$$

Evaluating the terms above gives

$$\begin{aligned}
\text{PMSE}(\hat{T}) &= \left(\sum_{i \notin s} x_i \frac{\sum_{j=1}^n W_b(x_i,x_j)m(x_j)}{\sum_{j=1}^n W_b(x_i,x_j)x_j} - \sum_{i \notin s} m(x_i) \right)^2 \\
&+ \left(\sum_{i \notin s} \frac{x_i W_b(x_i,x_j)}{\sum_{j=1}^n W_b(x_i,x_j)x_j} \right)^2 + \sum_{i \notin s} \sigma^2(x_i) \\
&= \text{MSE}(\hat{T}_r) + \sum_{i \notin s} \sigma^2(x_i)
\end{aligned}$$

Note the extra term that appears in the PMSE.

When all the weights are equal, *i.e.* $W_b(x_i, x_j) = 1$, then the above expressions agree with the results given by Royall and Herson (1973a) for the ratio estimator. They show that, in the case of a polynomial $m(x_i)$, the bias term is zero whenever the sample is *balanced* with respect to the population. What this means is that for a Pth degree polynomial, the pth sample moment of the x's is equal to the pth population moment, for all $p = 1, \ldots, P$.

The various estimators described and their properties are given in Tables 4.2-4.5. The smoother matrix is written in general form below:

$$\begin{pmatrix}
S_{11} & S_{12} & S_{13} & \cdots & S_{1n} \\
S_{21} & S_{22} & S_{23} & \cdots & S_{2n} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
S_{N1} & S_{N2} & S_{N3} & \cdots & S_{Nn}
\end{pmatrix}$$

For each estimator the elements S_{ij} are also given.

Estimator $\hat{T}_{SRE}$	$\sum_{h=1}^{H} N_h \bar{x}_h \frac{\sum_{k\in sh} y_{hk}}{\sum_{k\in sh} x_{hk}}$
$E_\xi(\hat{T}_{SRE} - T)$	$\sum_{h=1}^{H} N_h \bar{x}_h \left\{ \frac{\sum_{k\in sh} m(x_{hk})}{\sum_{k\in sh} x_{hk}} - \frac{\sum_{1}^{N_h} m(x_{hk})}{\sum_{1}^{N_h} x_{hk}} \right\}$
$\text{var}_\xi(\hat{T}_{SRE})$	$\sigma^2 \left\{ \sum_{h=1}^{H} \frac{(\sum_{k=1}^{N_h} x_{hk})^2}{\sum_{k\in sh} x_{hk}} \right\}$
PMSE($\hat{T}_{SRE}$)	$\text{var}(\hat{T}_{SRE}) + \text{bias}(\hat{T}_{SRE})^2 + \sum_{h=1}^{H} \sum_{k\notin sh} \sigma^2(x_{hk})$
Smoother matrix	$\begin{pmatrix} S_{11} & \dots & S_{11} & 0 & 0 & \dots & 0 \\ S_{12} & \dots & S_{12} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{1N_1} & \dots & S_{1N_1} & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & S_{h1} & \dots & S_{h1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & S_{hN_h} & \dots & S_{hN_h} \end{pmatrix}$

where $S_{hk} = \dfrac{x_{hk}}{\sum_{k=1}^{n_h} x_{hj}}$

Table 4.2: Separate ratio estimator and model-based properties

Estimator $\hat{T}_{LWRE}$	$\sum_{j\in s} y_j + \sum_{i\notin s} x_i \frac{\sum_{j=1}^{n} W_b(x_i,x_j)y_j}{\sum_{j=1}^{n} W_b(x_i,x_j)x_j}$ where e.g. $W_b(x_i, x_j) = \exp(-(x_i - x_j)^2/2h^2)$
$E_\xi(\hat{T}_{LWRE} - T)$	$\sum_{i\notin s} x_i \left\{ \frac{\sum_{j=1}^{n} W_b(x_i,x_j)m(x_j)}{\sum_{j=1}^{n} W_b(x_i,x_j)x_j} - \frac{\sum_{i\notin s} m(x_i)}{\sum_{i\notin s} x_i} \right\}$
$\mathrm{var}_\xi(\hat{T}_{LWRE})$	$\sum_{j\in s} \sigma^2(x_j) \left(\sum_{i\notin s} \frac{W_b(x_i,x_j)x_i}{\sum_{k\in s} W_b(x_i,x_k)x_k}\right)^2$
PMSE($\hat{T}_{LWRE}$)	$\mathrm{var}_\xi(\hat{T}_{LWRE}) + \mathrm{bias}(\hat{T}_{WLRE}(h))^2 + \sum_{i\notin s} \sigma^2(x_i)$
Elements of smoother matrix	$S_{ij} = \frac{x_i W_b(x_i,x_j)}{W_b(x_i,x_j)x_j}, \qquad j \notin s$ $S_{ij} = \left\{ \begin{array}{ll} 1, & \text{if } (i = j) \\ 0, & \text{if } (i \neq j), \quad j \in s \end{array} \right.$

Table 4.3: Locally weighted ratio estimator with kernel weight and model-based properties

Estimator $\hat{T}_{RRE}$	$\sum_{i\in s} y_i + \sum_{i\notin s} x_i \frac{\sum_{j\in s} W_{ji} y_j}{\sum_{j\in s} W_{ji} x_j}$ $W_{ji} = \{ \begin{array}{ll} 1, & \text{if } \lvert i-j\rvert < (k-1)/2 \\ 0, & \text{otherwise} \end{array}$
$E_\xi(\hat{T}_{RRE} - T)$	$\sum_{i\notin s} x_i \left\{ \frac{\sum_{j\in s} W_{ji} m(x_j)}{\sum_{j\in s} W_{ji} x_j} - \frac{\sum_{i\notin s} m(x_i)}{\sum_{i\notin s} x_i} \right\}$
$\text{var}_\xi(\hat{T}_{RRE})$	$\sum_{j\in s} \sigma^2(x_j) \left(\sum_{i\notin s} \frac{W_{ji} x_i}{\sum_{j\in s} W_{ji} x_j} \right)^2$
$\text{PMSE}(\hat{T}_{RRE})$	$\text{var}_\xi(\hat{T}_{RRE}) + \text{bias}(\hat{T}_{RRE}(k))^2 + \sum_{i\notin s} \sigma^2(x_i)$
Element of smoother matrix	$S_{ij} = \frac{x_i W_{ji}}{\sum_{j\in s} W_{ji} x_j}, \quad j \notin s$ W_{ji} as above $S_{ij} = \{ \begin{array}{ll} 1, & \text{if } (i = j) \\ 0, & \text{if } (i \neq j), \quad j \in s \end{array}$

Table 4.4: Running ratio estimator and model-based properties

Estimator $\hat{T}_{ARRE}$	$\sum_{i=1}^{N} \frac{x_i}{k} \left\{ \sum_{m=1}^{k} \frac{\sum_{j \in s} W_{jim} y_j}{\sum_{j \in s} W_{jim} x_j} \right\}$ $W_{jim} = \begin{cases} 1, & \text{if } \lvert i-j \rvert < (k-m)/2 \\ 0, & \text{otherwise} \end{cases}$
$E_\xi(\hat{T}_{ARRE} - T)$	$\sum_{i=1}^{N} \frac{x_i}{k} \left\{ \sum_{m=1}^{k} \left(\frac{\sum_{j \in s} W_{jim} m(x_j)}{\sum_{j \in s} W_{jim} x_j} - \frac{\sum_{i=1}^{N} m(x_i)}{\sum_{i=1}^{N} x_i} \right) \right\}$
$\text{var}_\xi(\hat{T}_{ARRE})$	$\sum_{j \in s} \sigma^2(x_j) \left\{ \sum_{i=1}^{N} \frac{x_i}{k} \left(\sum_{m=1}^{k} \frac{W_{jim}}{\sum_{l \in s} W_{lim} x_l} \right) \right\}^2$
$\text{PMSE}(\hat{T}_{ARRE})$	$\text{var}_\xi(\hat{T}_{ARRE}) + \text{bias}(\hat{T}_{ARRE}(k))^2 + \sum_{i \notin s} \sigma^2(x_i)$
Element of smoother matrix	$S_{ij} = \frac{x_i}{k} \left(\sum_{m=1}^{k} \frac{W_{jim}}{\sum_{l \in s} W_{lim} x_l} \right),$ W_{jim}as above

Table 4.5: Averaged running estimator and model-based properties

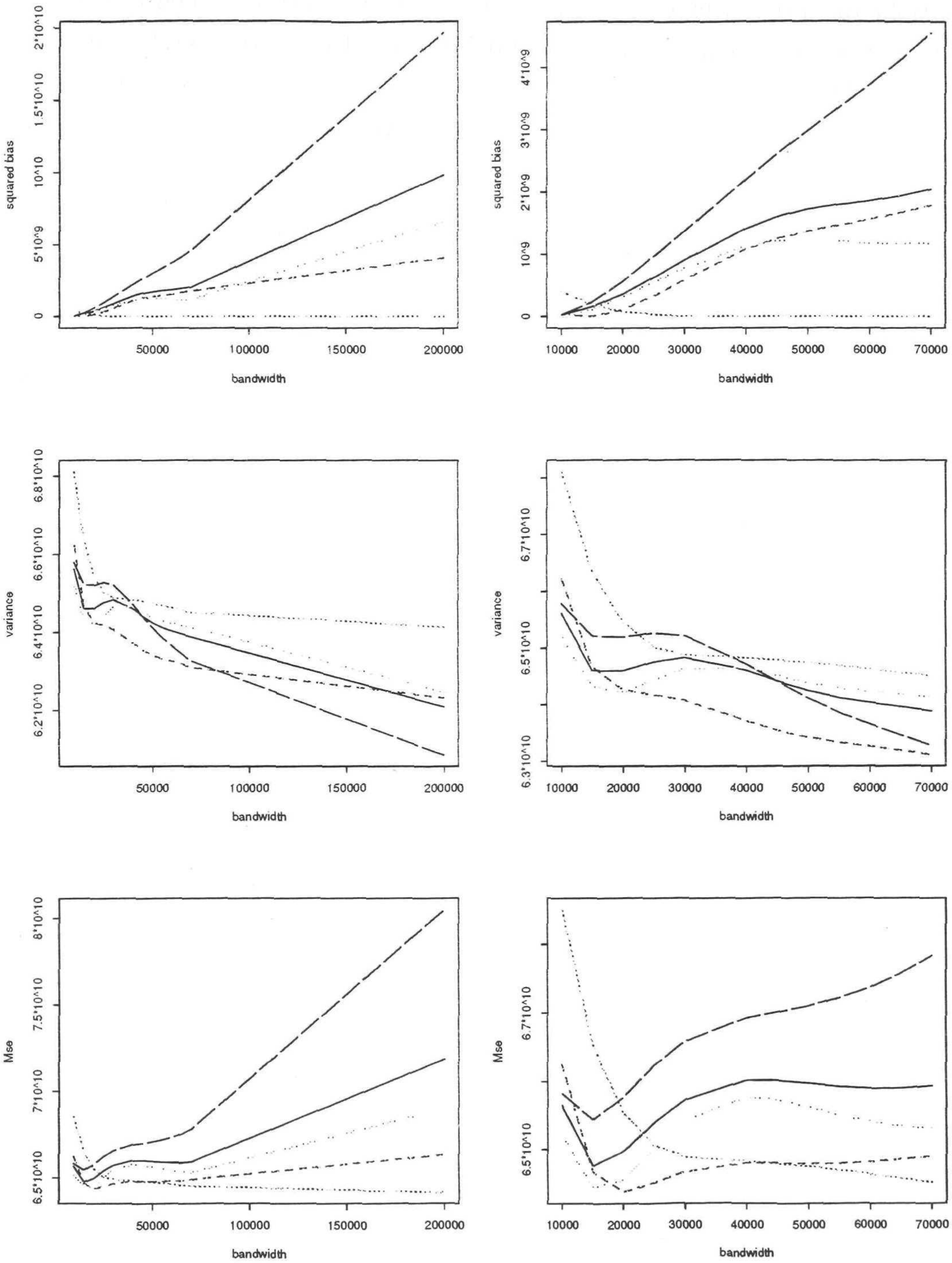


Figure 4.2: Bias, variance and MSE plots against smoothing parameter for the locally weighted ratio estimator with Gaussian kernel weight function

In Figure 4.2, squared bias, variance and mean squared error for the locally weighted ratio estimator with Gaussian kernel weight function are plotted against smoothing parameter for five different random samples. The assumed underlying superpopulation model is a quadratic with a linear term. The bias increases with increasing smoothing parameter. In this example, as the prediction moves away from fitting points closely, the bias increases. If the estimator fitted the points too closely, we would have an increased variance which is indicated in the variance plot. Small smoothing parameter values lead to a larger variance; the ratio estimator always gives the smallest variance. Thus the mean squared error or the predictive mean squared error trades off bias and variance, to arrive at a compromise value of the smoothing parameter. The difficulty is that the population mean squared error cannot be computed if the population units are not known. This is where methods for choosing the smoothing parameter such as crossvalidation are of some use. These methods try to estimate the mean squared error based on the sample values only and are discussed further in Chapter 5.

4.5 Degrees of freedom

In Section 3.7 the notion of degrees of freedom based on smoother matrices as given in Tibshirani and Hastie (1987) was introduced. These degree of freedom formulations provide one way of calibrating between various estimators in terms of their smoothing parameter. It is particularly useful for comparing estimators with the separate ratio estimator, as a smoothing parameter can be found corresponding to the number of strata.

Here a new formula for the degrees of freedom is introduced, based on smoother matrices and derived from a modified residual sum of squares. This method requires some knowledge about the variance function, $\sigma^2(x_i)$, for y_i.

Recall that $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ where $\mathbf{S}$ is an $n \times n$ smoother matrix based on the known sample x values alone. As an alternative motivation for the degrees of freedom consider

$$E_\xi(\mathrm{MRSS}) = E_\xi\left(\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\sigma^2(x_i)}\right),$$

modified in order to take account of the variance not being a constant in this case. Suppose $y_i^{'} = y_i/\sigma(x_i)$, and $\hat{y}_i^{'} = \hat{y}_i/\sigma(x_i)$ then,

$$\begin{aligned} E_\xi(\mathrm{MRSS}) &= E_\xi(\mathbf{y}^{'} - \hat{\mathbf{y}}^{'})^T(\mathbf{y}^{'} - \hat{\mathbf{y}}^{'}) \\ &= E_\xi(\mathbf{Ay} - \mathbf{A}\hat{\mathbf{y}})^T(\mathbf{Ay} - \mathbf{A}\hat{\mathbf{y}}) \qquad (4.2) \end{aligned}$$

where $\mathbf{A} = \mathrm{diag}(1/\sigma(x_i))$.

Theorem 4.1 *From the expected value of the modified residual sum of squares*

$$\sum_{j=1}^{n} \frac{(y_j - \hat{y}_j)^2}{\sigma^2(x_j)}$$

the derived degrees of freedom formula is:

$$\mathrm{df} = 2\,\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{S}^T\mathbf{V}(\mathbf{X})^{-1}\mathbf{S}\mathbf{V}(\mathbf{X})),$$

where $\mathbf{S}$ *is the smoother or hat matrix of the estimator and* $\mathbf{V}(\mathbf{X}) = \mathit{diag}(\sigma^2(x_i))$.

Proof of Theorem

We can write

$$\begin{aligned} E_\xi(\mathrm{MRSS}) &= E_\xi(\mathbf{Ay} - \mathbf{ASy})^T(\mathbf{Ay} - \mathbf{ASy}) \\ &= E_\xi(\mathbf{A}(\mathbf{I} - \mathbf{S})\mathbf{y})^T(\mathbf{A}(\mathbf{I} - \mathbf{S})\mathbf{y}) \\ &= E_\xi(\mathbf{y}^T(\mathbf{I} - \mathbf{S})^T\mathbf{A}^T\mathbf{A}(\mathbf{I} - \mathbf{S})\mathbf{y}) \\ &= E_\xi(\mathbf{y})^T(\mathbf{I} - \mathbf{S})^T\mathbf{A}^T\mathbf{A}(\mathbf{I} - \mathbf{S})E_\xi(\mathbf{y}) \\ &\quad + \mathrm{tr}\left((\mathbf{I} - \mathbf{S})^T\mathbf{A}^T\mathbf{A}(\mathbf{I} - \mathbf{S})\,\mathrm{var}_\xi(\mathbf{y})\right) \qquad (4.3) \end{aligned}$$

The last line was obtained using the result $E_\xi(\mathbf{z}^T\mathbf{Az}) = E_\xi(\mathbf{z})^T\mathbf{A}E_\xi(\mathbf{z}) + \mathrm{tr}(\mathbf{A}\Sigma)$ where $\Sigma = \mathrm{var}_\xi(\mathbf{z})$. The first term is the squared bias and the second term is the analogue of $(n-p)\sigma^2$ in linear regression.

To derive a formula for the degrees of freedom, consider the second part of $E_\xi(\mathrm{MRSS})$ in more detail:

$$\begin{aligned} \mathrm{tr}\left((\mathbf{I} - \mathbf{S})^T\mathbf{A}^T\mathbf{A}(\mathbf{I} - \mathbf{S})\,\mathrm{var}_\xi(\mathbf{y})\right) &= \mathrm{tr}\left((\mathbf{A} - \mathbf{AS})^T(\mathbf{A} - \mathbf{AS})\,\mathrm{var}_\xi(\mathbf{y})\right) \\ &= \mathrm{tr}\left((\mathbf{A}^T\mathbf{A} - (\mathbf{AS})^T\mathbf{A} - \mathbf{ASA}^T + (\mathbf{AS})^T\mathbf{AS})\,\mathrm{var}_\xi(\mathbf{y})\right). \end{aligned}$$

Here $\mathrm{var}_\xi(\mathbf{y}) = \mathrm{diag}\left(\sigma^2(x_i)\right)$,

$$\begin{aligned} \mathrm{tr}(\mathbf{A}^T\mathbf{A}\,\mathrm{var}_\xi(\mathbf{y})) &= \mathrm{tr}(\mathbf{I_n}) &&= n \\ \mathrm{tr}((\mathbf{AS})^T\mathbf{A}\,\mathrm{var}_\xi(\mathbf{y})) &= \mathrm{tr}(\mathbf{S}^T\mathbf{A}^T\mathbf{A}\,\mathrm{var}_\xi(\mathbf{y})) &&= \mathrm{tr}(\mathbf{S}^T) \\ \mathrm{tr}((\mathbf{AS})\mathbf{A}^T\,\mathrm{var}_\xi(\mathbf{y})) &= \mathrm{tr}(\mathbf{SA}^T\,\mathrm{var}_\xi(\mathbf{y})\mathbf{A}) &&= \mathrm{tr}(\mathbf{S}) \end{aligned}$$

So equation (4.3) above simplifies to :

$$\begin{aligned}\mathrm{tr}((\mathbf{I}-\mathbf{S})^T\mathbf{A}^T\mathbf{A}(\mathbf{I}-\mathbf{S})\,\mathrm{var}_\xi(\mathbf{y})) &= n-2\,\mathrm{tr}(\mathbf{S})+\mathrm{tr}((\mathbf{AS})^T\mathbf{AS}\,\mathrm{var}_\xi(\mathbf{y}))\\ &= n-2\,\mathrm{tr}(\mathbf{S})+\mathrm{tr}(\mathbf{S}^T\mathbf{A}^T\mathbf{AS}\,\mathrm{var}_\xi(\mathbf{y}))\\ &= n-2\,\mathrm{tr}(\mathbf{S})+\mathrm{tr}(\mathbf{S}^T\mathbf{V}(\mathbf{X})^{-1}\mathbf{SV}(\mathbf{X}))\end{aligned} \tag{4.4}$$

where $\mathbf{V}(\mathbf{X}) = \mathrm{diag}\,(\sigma^2(x_i))$. Again this is the analogue to the $(n-p)\sigma^2$ term in linear regression. Hence, the degrees of freedom formula can be written as:

$$\mathrm{df} = 2\,\mathrm{tr}(\mathbf{S}) - \mathrm{tr}(\mathbf{S}^T\mathbf{V}(\mathbf{X})^{-1}\mathbf{SV}(\mathbf{X})).\Box$$

For the ratio estimator, $\sigma^2(x_i) = x_i$, and the smoother matrix elements consist of $x_i/\sum_{j=1}^n x_j$ for rows $i = 1,\ldots,n$. From this the new form the degrees of freedom gives df=tr(S)=1. Similarly, for the stratifed ratio estimator the degrees of freedom formula leads to df=H, the number of strata.

4.6 An example

An example is included to compare the different types of estimator.

Data is taken from the Family Expenditure Survey (FES) of Section 1.4. A random sample of 100 observations are selected from the population using stratified simple random sampling. To compare our estimators with the separate ratio estimator, 4 strata are assumed and samples are selected in two ways :

1. Optimal allocation. Assuming equal numbers of population units in each stratum (125), optimal allocation chooses $n_h \propto N_h(\bar{x})^{1/2}$. In this case the sample stratum consisted of (14,23,28,35) units in each stratum respectively.

2. Proportional allocation. Equal numbers of sample values in each stratum; 25 sample units were selected from each population stratum.

4.6.1 Results

An underlying superpopulation model was obtained by fitting models in *GLIM* using the population data. The final model was assumed quadratic with a linear and intercept term $E(y_i) = 3872 + 0.1665x_i - 3.972\,10^{-7}x_i^2$, and $V(y_i) = \sigma^2 x_i$ with multiplying constant $\hat{\sigma}^2 = 465$ calculated from the data using $\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{((N-1)x_i)}$. As the intercept was small in relation to the large range of the x and y values, it was not omitted from the equation. Below are some of the results.

(a) Optimal allocation

	Ratio Est.	Strat. RE	LWRE $b = 30000$	Running RE $k = 25$	Averaged RRE $k = 51$
Bias	-163548	-50290	-30799	-36333	-61940
Variance $\times 10^{10}$	6.0833	6.5184	6.5606	6.4339	6.3021
MSE $\times 10^{10}$	8.7581	6.7713	6.6655	6.5660	6.6858

Each of the smoothing parameters used in the locally weighted ratio estimator with Gaussian kernel and Uniform k-nearest neighbour weight functions and the averaged running ratio estimator are based on 4 degrees of freedom. The degrees of freedom formula used is based on the result from Theorem 4.1 of Section 4.5.

In this example, the running ratio estimator and Gaussian kernel weighted local ratio estimator appear the best in reducing bias. The ratio estimator has the worst bias as expected, since it is not taking account of the quadratic feature in the data. The separate ratio estimator and the averaged running ratio estimator are comparable in this instance. The separate ratio estimator is expected to do well because of the optimal allocation. All of the estimators are comparable in terms of variance, except the ratio estimator which has the smallest variance as expected. These results are only for one sample, and each sample will be different. Optimal allocation is like using a *variable bandwidth* in the kernel smoothing context. It allocates more observations to strata where there is more variability, in other words the larger the x values become. Global and variable bandwidths are considered further in Chapter 5.

Now consider the case when we do not assume optimal allocation.

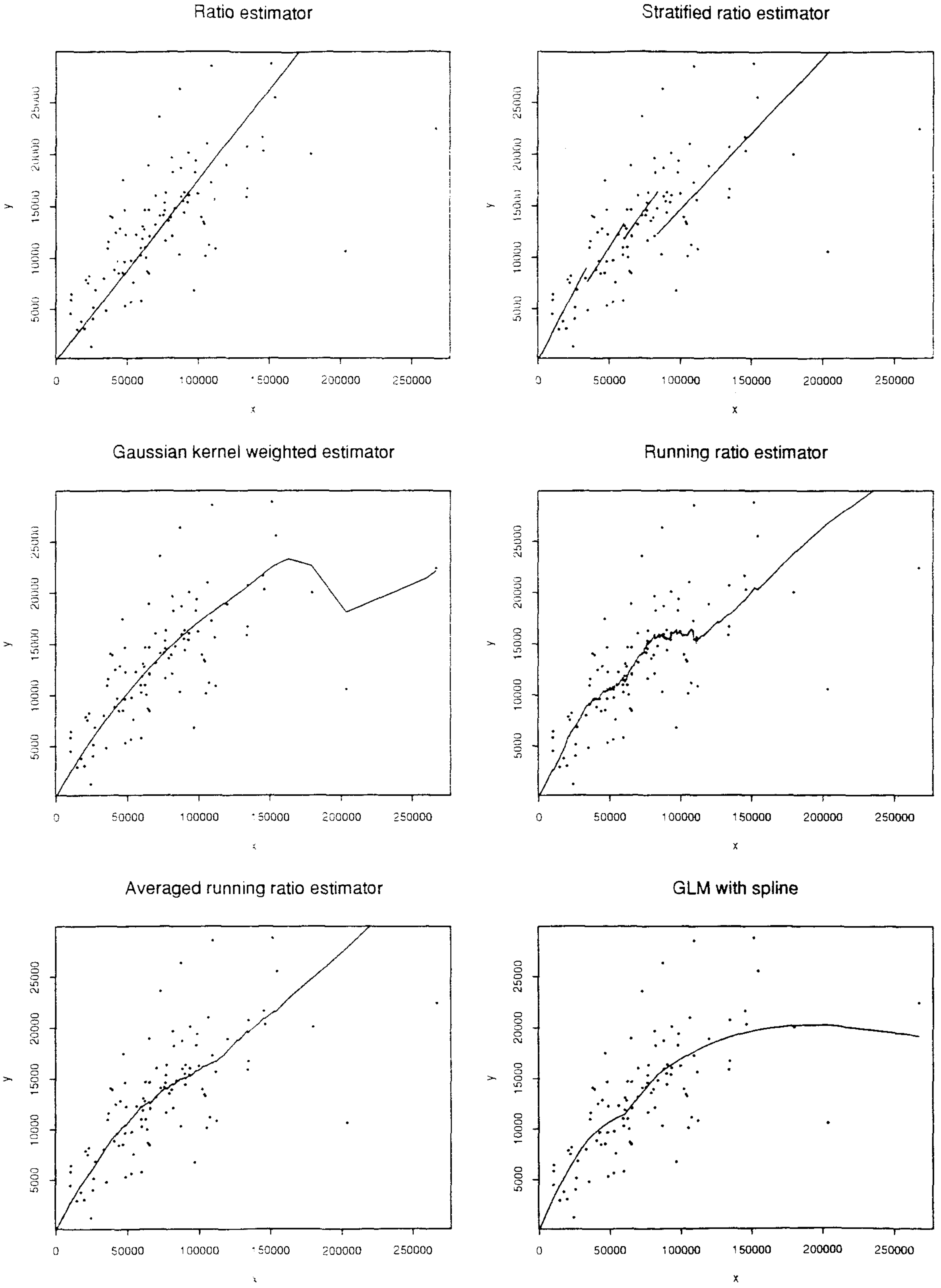
(b) Equal/Proportional allocation

	Ratio Est.	Strat. RE	LWRE $b = 35000$	Running RE $k = 27$	Averaged RE $k = 51$
Bias	−169626	−171748	−111714	−122457	−123236
Variance $\times 10^{10}$	6.962	6.680	7.449	7.147	6.977
MSE $\times 10^{10}$	9.841	9.9293	8.6973	8.6466	8.4958

The second example gives different results. Here the locally weighted ratio estimator with Gaussian kernel weight gives a smaller bias than the running ratio estimator and the averaged running ratio estimator, the ratio and separate ratio estimator giving the worst results in terms of bias under the superpoulation model.

The ratio estimator is the best linear unbiased estimator (BLUE), when the sample drawn is balanced with respect to the population, or when the underlying superpopulation is in fact linear through the origin. It tends to get worse as the data deviate from these ideal conditions, and this is when an alternative nonparametric regression based estimator performs better. Alternative estimators of the population total are therefore considered, in an effort to improve on results given by Royall and Herson (1973a, b). The best alternative estimator depends on the value of the smoothing parameter chosen. Figure 4.3 is a plot of the various estimators for one sample, calibrated so that each is based on 4 degrees of freedom (except the ratio estimator). Note the discontinous nature of the separate ratio estimator, the smooth curve based on the Gaussian Kernel weight and the jagged appearance of the running ratio estimator. The averaged running ratio estimator and generalised linear model with regression spline give predicted values close to the true values. Both are smooth total-preserving estimators.

Figure 4.4 has been included to illustrate the equivalent kernels of the various estimators already mentioned. Equivalent kernels are introduced in Section 3.4, where smoother matrices and degrees of freedom are discussed. Equivalent kernels indicate the shape of the weight function used to weight the response variables y. The separate ratio estimator has been included; the ratio estimator is very similar in that a constant weight is applied to all y's. Also included are the locally weighted ratio estimator with Gaussian kernel weight, the running ratio estimator



Ratio estimator

Stratified ratio estimator

Gaussian kernel weighted estimator

Running ratio estimator

Averaged running ratio estimator

GLM with spline

Figure 4.3: Various smoothers calibrated to 4 degrees of freedom; ratio estimator based on 1 degree of freedom.

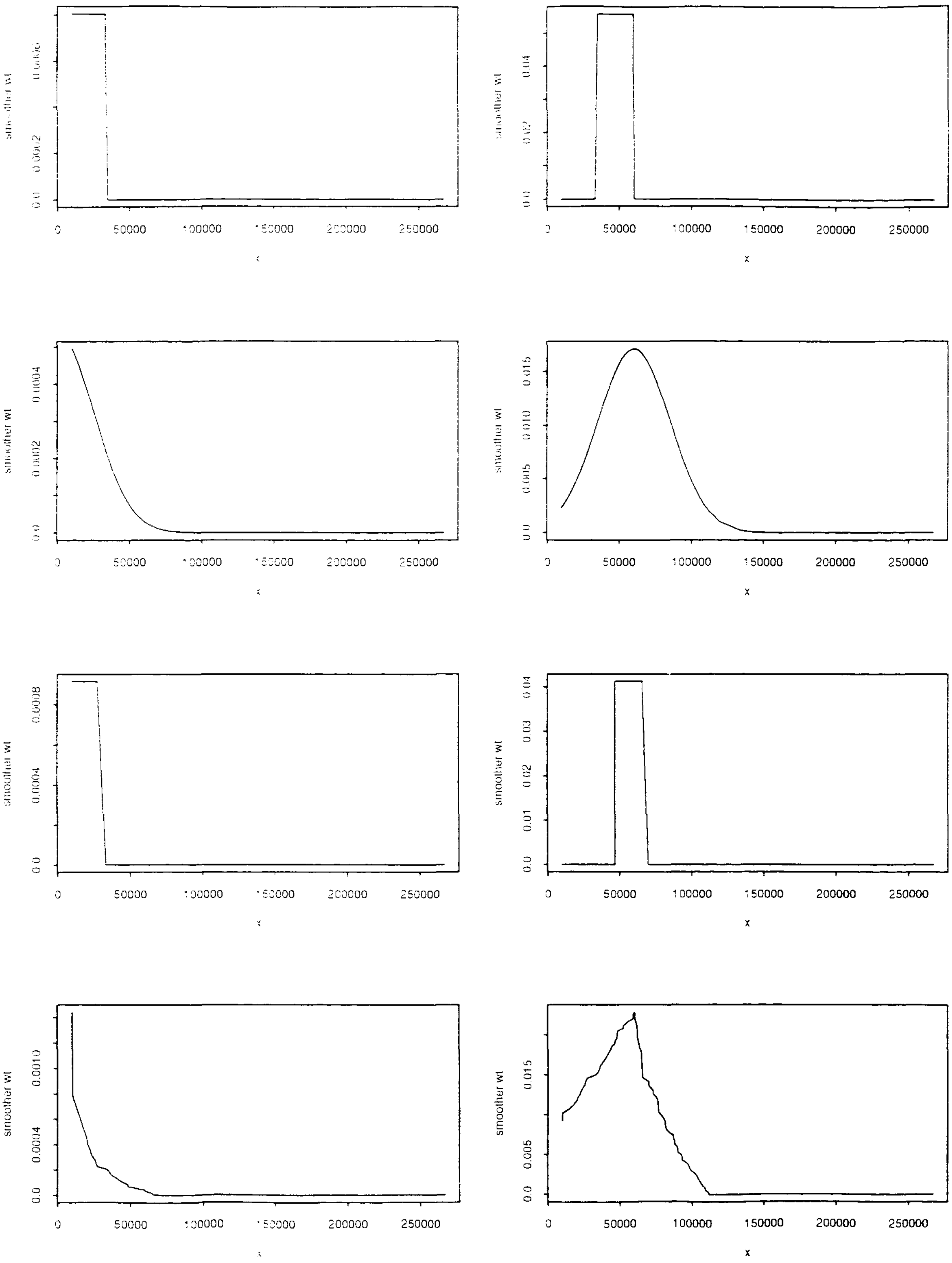


Figure 4.4: Plots of equivalent kernels; separate ratio estimator, LWRE with Gaussian kernel, running ratio estimator and averaged running ratio estimator.

and averaged running ratio estimator. For each type of estimator a target point is chosen near the boundary and at the centre of the data. The running ratio estimator weights drop off abruptly outside the neighbourhood, and this accounts for the jagged appearance of this estimator. In the weighted ratio estimator, the Gaussian kernel weights are smooth and hence the resultant curve is also smooth. These kernel weights, however, have the disadvantage of a fixed bandwidth for all x, unlike the running ratio estimator.

4.7 Discussion and conclusions

In this chapter some alternative nonparametric estimators of the population total were discussed and distinguished as operational or model-based, total-preserving and not total-preserving. Those with the total-preserving property have some good features, especially under repeated sampling, and this is discussed further in Chapter 6.

The nonparametric regression estimators are motivated by assuming a smooth function derived from 'local' quasi-likelihood estimating equations. They can be extended to any class of generalised linear model, or any model with known mean and variance functions. In this thesis, emphasis is mainly on locally weighted ratio or regression estimators. Model-based properties such as bias, variance, and mean squared error of these nonparametric estimators have been investigated. The nonparametric estimators are shown to be more efficient than the parametric methods in certain situations. The separate ratio estimator is an example of a parametric estimator with a varying bandwidth which adapts, to some extent, to the local density of x. It is advantageous to have nonparametric estimators which also take account of this. The running estimators and averaged running estimators described do, to a certain extent, by varying with the design density of the x's. In Chapter 5 a local bandwidth for the Gaussian kernel weight function is discussed, which adapts to the local density and possible other factors. The estimator with a local bandwidth performs considerably better than that with a global bandwidth.

Finally, a new form for the degrees of freedom is derived which can be applied to any prediction setting with heterogenous variance, and is based on the smoother matrix and the assumed variance of y_i. It is a useful measure for calibrating between the separate ratio estimator and the nonparametric estimators in terms of their smoothing parameter and is easily calculated for any of the estimators mentioned.

Chapter 5

Choosing the smoothing parameter

5.1 Introduction

For some practical purposes, it is sufficient to choose the smoothing parameter subjectively, by plotting out a few curves with varying bandwidths, and choosing the one that 'looks best'. In certain contexts choosing the smoothing parameter 'by eye' may not be the best or most accurate method. This is particularly true if the curve is to be used for estimation purposes and some of the local structure is required. This chapter will look at some ways of choosing the smoothing parameter, first generally and then by applying some of the standard methods to the specific context of predicting a population total.

5.2 Automatic methods

5.2.1 Introduction to crossvalidation

The automatic methods described are based on approximating some global error measure which is required to be minimised; probably the most attractive class is *crossvalidation*. This method was first introduced by Stone (1974), and Geisser (1979) and in the nonparametric setting by Clark (1975). The basic idea is to split the dataset into two parts. The first part is used for calculating the estimate and the second part is used for optimising the fit of the estimate by minimising the MSE, PMSE or some other quadratic error measure for the regression curve. In this way we have a method of optimising the window width. These quadratic error forms depend on variance and bias components, which respectively decrease and increase with the smoothing parameter. Therefore, it is desirable to have a smoothing parameter which balances the systematic squared bias with the stochastic nature of the variance.

Suppose $\hat{y}^{(j)}$ is the predicted value from a kernel regression estimator, based on leaving one observation, j, out of the dataset. Then crossvalidation optimises the smoothing parameter, b, by minimising

$$CV(b) = \sum_{j=1}^{n}(y_j - \hat{y}^{(j)})^2, \tag{5.1}$$

where b is in a suitable range of smoothing parameter values. This criterion validates the ability to predict $y_j, \quad j = 1, \ldots, n$ across the subsamples $(x_i, y_i), i \neq j$. The resulting optimal b, denoted by $\hat{b}$, satisfies

$$CV = \inf\{CV(b) | b \in B\}$$

and is used in $\hat{y}$. Crossvalidation, up to a constant term, approximates a quadratic error measure, usually the predictive mean squared error.

Least squares crossvalidation, as described above, was proposed independently by Rudemo (1982) and Bowman (1984) for density estimation. Härdle and Marron (1985a,b) used least squares crossvalidation to choose the bandwidth for kernel smoothers in the regression model. The drawbacks to this type of crossvalidation are:

1. It tends to have several minima, and so minimisation is best performed through a grid search, for a sensible range of values for b.
2. It is subject to a great deal of sample variability. This has been quantified asymptotically by Hall and Marron (1987), for example.
3. It is expensive to compute, especially if a grid search is used.

The crossvalidation criterion, (5.1) above can be written in terms of the smoother matrix, as given by Hastie and Tibshirani (1990), since $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$. Then

$$CV(b) = \frac{1}{n}\sum_{i \in s}\left(\frac{y_i - \hat{y}_i}{1 - S_{ii}}\right)^2,$$

where $S_{ii} = (\mathbf{SS}^T)_{ii}$, the ith diagonal element of $\mathbf{SS}^T$. Also

$$E(CV(b)) \doteq \text{PMSE} + \frac{2}{n} \sum_{i \in s} S_{ii} \text{bias}(x_i)^2.$$

Crossvalidation is easy to compute if the diagonal elements of the smoother matrix can be readily obtained. For smoothing splines this is not such an easy task, and so the related *generalised crossvalidation* proposed by Craven and Wahba (1979) is often used. This replaces the S_{ii} above by its average value $\text{tr}(\mathbf{S})/n$, which is easier to compute.

A weight function can be introduced into crossvalidation to eliminate or reduce boundary effects or to allow for variations in the population density etc. This leads to a weighted crossvalidation criterion as follows:

$$WCV(b) = n^{-1} \sum_{j=1}^{n} w_j (y_j - \hat{y}^{(j)})^2.$$

To make crossvalidation a mathematically justifiable device for selecting the smoothing parameter, it has to be shown that the MSE($\hat{b}$), $\hat{b}$ minimizing the crossvalidation criterion, approximates min[MSE(b)], *i.e.*

$$\frac{\text{MSE}(\hat{b})}{\inf_b \text{MSE}(b)} \overset{a.s}{\rightarrow} 1.$$

A data driven bandwidth, b, satisfying this is said to be *asymptotically optimal.* Härdle and Marron (1985a,b) consider this further, to show that data driven smoothers achieve their optimal rate independently of the smoothness of the underlying regression model. Härdle et al (1988) study how fast this convergence occurs.

A number of crossvalidation and other procedures have been reviewed and used by other authors, including Eubank (1988), Härdle, Hall and Marron (1988) and Härdle (1990a). *Biased crossvalidation* as proposed by Scott and Terrell (1987), is a combination of least-squares crossvalidation and another method, the *plug-in* method, which is considered later. A recent survey of automatic smoothing parameter selection is given by Marron (1988). Park and Marron (1989) look at a comparison of data driven techniques in the related field of density estimation. Sheather and Jones (1991) improve on the crossvalidation method as described by Park and Marron. Titterington (1985) gives a review of several smoothing techniques and work on crossvalidation; in particular Section 7 of his paper is devoted to nonparametric regression and curve fitting. Problems with crossvalidation and other automatic methods are given in Härdle, Hall and Marron (1988).

5.2.2 Penalising functions

If we try to use PMSE(b) as an estimator of the MSE(b), bias occurs, and, as a consequence, the minimizer of PMSE(b) leads to too small a choice of b. In order to compensate for this, PMSE(b) is adjusted by a correction factor, which penalises values of b that are too small.

Rice (1984b) has shown there is a range of bandwidth selectors (based on work by Akaike, Shibata and others) including generalised crossvalidation (GCV) and crossvalidation (CV) which have the following structure:

$$G(b) = n^{-1}\sum_{j=1}^{n}(y_j - \hat{y}_j))^2\Psi(n^{-1}b^{-1}),$$

where $\Psi(n^{-1}b^{-1})$ is a correction or penalising factor, and $G(b)$ is minimised over a range of suitable values of b. Up to a constant term, $G(b)$ is approximately equal to MSE(b).
Simple examples include :

1. Generalised crossvalidation (Craven and Wahba 1979),

$$\Psi_{GCV}(n^{-1}b^{-1}) = (1 - n^{-1}b^{-1}K(0))^{-2}$$

2. Akaike's Information Criterion (Akaike 1974)

$$\Psi_{AIC}(n^{-1}b^{-1}) = \exp(2n^{-1}b^{-1}K(0))$$

3. Shibata's (1981) model selector,

$$\Psi_S(n^{-1}b^{-1}) = (1 + 2n^{-1}b^{-1}K(0)).$$

4. Rice's (1984b) bandwidth selector,

$$\Psi_T(n^{-1}b^{-1}) = (1 - 2n^{-1}b^{-1}K(0))^{-1}.$$

Each of these functions has the same Taylor series expansion . As $nb \to \infty$,

$$\Psi(n^{-1}b^{-1}) = 1 + 2n^{-1}b^{-1}K(0) + O(n^{-2}b^{-2}).$$

All of the above bandwidth selectors are asymptotically optimal, *i.e.* the relative loss of a selected bandwidth $\hat{b}$ to the minimum loss is:

$$\frac{\mathrm{MSE}(\hat{b})}{\min_b \mathrm{MSE}(b)} \overset{a.s}{\to} 1.$$

With increasing observations we assume $\hat{b}$ approximates b, however the convergence rate tends to be quite slow $(n^{1/10})$. Härdle and Marron (1985b) give a theorem to show this for crossvalidation, and Rice (1984b) for the fixed design case, gives a related theorem using penalty functions. Another property of data-driven smoothers is that they achieve their optimal rate, independent of the smoothness of the underlying regression model.

5.2.3 Crossvalidation in predicting totals

Some methods of selecting b, using crossvalidation, are now considered for use with nonparametric regression estimators of a finite population total. The first method is motivated by the prediction mean squared error, which we are trying to estimate; this is the squared difference between the predicted total and the true total with the predicted values based on the leave-one-out estimator. The crossvalidation criterion is given as:

$$CV_1 = \left[\sum_{j=1}^{n}(y_j - \hat{y}^{(j)})\right]^2 .$$

The second method is the usual least squares crossvalidation, based on the sum of the squared discrepancies between the actual y values and the leave-one-out predicted y values and is given by:

$$CV_2 = \sum_{j=1}^{n}\left(y_j - \hat{y}^{(j)}\right)^2 .$$

Note that this crossvalidation criterion is based on sample values only and the measure we are trying to approximate the CV criterion to, the mean squared error say, depends on sample and nonsample values.

A weight may be introduced into the CV criterion, for example, by including the design density (or an estimator of the design density). From some experimentation, including the weight did not appear to improve the method; however, a different weight function might be more appropriate.

The criteria, CV_1 and CV_2 may be simplified by replacing $\hat{y}^{(j)}$ by $S_j'\mathbf{y}$, where S_j' is the jth row of the smoother matrix based on the sample values with the jth element removed, *i.e.*

$$y_j - \hat{y}^{(j)} = (I_j - S_j')\mathbf{y}.$$

For the locally weighted ratio estimator,

$$y_j - \hat{y}^{(j)} = (y_j - \hat{y}_j)\left(1 - \frac{x_j w_{jj}}{\sum_{k\in s} w_{jk}x_k}\right)^{-1}.$$

Hence, CV_1 is written as:

$$\left[\sum_{j=1}^{n}(y_j - \hat{y}^{(j)})\right]^2 = \left[\sum_{j=1}^{n}(y_j - \hat{y}_j)\left(1 - \frac{x_j w_{jj}}{\sum_{k\in s} w_{jk}x_k}\right)^{-1}\right]^2. \quad (5.2)$$

Similarly, for CV_2 we have:

$$\sum_{j=1}^{n}(y_j - \hat{y}^{(j)})^2 = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2\left(1 - \frac{x_j w_{jj}}{\sum_{k\in s} w_{jk}x_k}\right)^{-2}. \quad (5.3)$$

The first crossvalidation criterion may be written in terms of the weighted discrepancy between y and its predicted value, weighted according to smoother matrix weights. In the case of the second criterion above, we have a weighted version of the prediction error which is of a similar form to the penalising functions described in Section 5.2.2. The penalising functions described in Section 5.2.2 are for mean functions. The penalising function in our case could be taken as $(1 - S_{ii})^{-2}$. This factor could be changed to take account of other penalising factors such as $(1 - \mathrm{tr}(S)/n)^{-2}$ in the generalised crossvalidation case or any of the other examples. Recall that

$$G(b) = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2\Psi(n^{-1}b^{-1}).$$

If we replace $\Psi(n^{-1}b^{-1})$ by $(1 - S_{jj})^{-2}$ then

$$G(b) \approx \sum_{j=1}^{n}(y_j - \hat{y}_j)^2(1 + 2S_{jj}),$$

using a Taylor series approximation. Then

$$G(b) \approx \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 + 2\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 S_{jj}$$

and taking expectations we arrive at

$$E(G(b)) \approx PSE(b) + 2\sum_{j=1}^{n} S_{jj}\text{bias}^2(x_j).$$

The first term is based on the sample and the last term is the error due to the bias contributions. A similar expectation for CV_1 gives:

$$E(CV_1) = MSE(b) + \sum_{j=1}^{n}\text{bias}(x_j)\sum_{j=1}^{n} S_{jj}\text{bias}(x_j) + (\sum_{j=1}^{n} S_{jj}\text{bias}(x_j))^2.$$

It was found that the first method, CV_1 was similar to the sample squared error, in the examples considered. This crossvalidation criterion tends to be minimised at small and large values of b but not in between, for all samples. This is incorrect if the optimal b lies somewhere in between the two extremes. Returning to equation (5.2), we can see why a large b is often chosen. The weight applied to $(y_j - \hat{y}_j)$ is

$$\left(1 - \frac{x_j w_{jj}}{\sum_{k\in s} w_{jk}x_k}\right)^{-1}.$$

This is small when $\frac{x_j w_{jj}}{\sum_{k\in s} w_{jk}x_k}$, ith diagonal element of $\mathbf{S}$, is as small as possible, i.e. as b gets larger.

The second type of pointwise crossvalidation, CV_2, tends to chose small values of b as the minimising values; it underestimates the 'true' value of b, sometimes considerably. Again, this is possibly because the squared discrepancy in (5.3) is minimised when small b is used.

The underlying problem with both of these methods is that they depend on sample values only to choose the smoothing parameter, for the purpose of prediction of the population values.

Modified crossvalidation

The modified crossvalidation approach considered is similar to ν-fold crossvalidation discussed by Burman (1989). Instead of removing one observation at a time from the prediction of y_i in crossvalidation, a proportion, $1 - p$, of the sample observations are removed each time and the remaining np observations are used to predict. This crossvalidation criterion is computed for several subsamples, size np, from the original sample and an average over the subsamples is taken. The criterion, for the locally weighted ratio estimator, is

$$CV_3 = \frac{1}{100}\sum_{r=1}^{100}\left(\sum_{j\in s} y_j - x_j\frac{\sum_{l\in sub_r} W_b(x_j,x_l)y_l}{\sum_{l\in sub_r} W_b(x_j,x_l)x_l}\right)^2,$$

where sub_r is the subsample, size np drawn from the sample, for varying proportions ($p = 0.5, 0.75, 0.9, 0.95$). Examples of these crossvalidatory methods are presented in Section 5.5 with the hospital dataset of Section 1.4.

This criterion is extended further to take account of the estimation of the unknown non-sample part of the population only. This leads to a crossvalidation criterion similar to CV_1 :

$$CV_4 = \left(\frac{1}{100}\sum_{r=1}^{100}\left(\sum_{j\notin sub_r} y_j - x_j\frac{\sum_{l\in sub_r} W_b(x_j,x_l)y_l}{\sum_{l\in sub_r} W_b(x_j,x_l)x_l}\right)\right)^2. \tag{5.4}$$

Similarly for a criterion similar to CV_2 we have :

$$CV_5 = \frac{1}{100}\sum_{r=1}^{100}\left(\sum_{j\notin sub_r} y_j - x_j\frac{\sum_{l\in sub_r} W_b(x_j,x_l)y_l}{\sum_{l\in sub_r} W_b(x_j,x_l)x_l}\right)^2. \tag{5.5}$$

When $p = (n-1)/n$ and sufficient subsamples are drawn, (5.4) and (5.5) above reduce to crossvalidation methods (5.2) and (5.3) respectively. Weighted versions of (5.4) and (5.5) above were also considered, with weights based on the underlying design density of the x_j not in the subsample. The weighted version did not improve dramatically on the unweighted methods, however a different weight function could have been considered.

An approximation for the modified crossvalidation criterion

The above forms of modified crossvalidation are a type of averaged crossvalidation. Suppose we consider all possible subsamples from the original sample, calculate the non-subsample squared error for each, and then average these non-subsample squared errors over all possible subsamples that could have been selected. This can be written as

$$CV_4 = \left(E_\pi\left(\sum_{j\notin sub} y_j - x_j\frac{\sum_{l\in sub} W_b(x_j,x_l)y_l}{\sum_{l\in sub} W_b(x_j,x_l)x_l}\right)\right)^2,$$

where E_π is the expectation taken with respect to simple random sampling n_s units from n and where n_s is the number of units in the subsample. The averaged crossvalidation criterion, CV_4 above, for the LWRE can be written as

$$CV_4 = \left[\sum_{i=1}^{n} y_i \left(1 - \frac{n_s}{n}\right) - \sum_{i=1}^{n} \frac{x_i \sum_{j \in s} y_j \omega_{ij}}{\sum_{j \in s} x_j \omega_{ij}} \left(1 - \frac{n_s - 1}{n - 1}\right) + \frac{n - n_s}{n - 1} \sum_{i=1}^{n} \frac{x_i y_i \omega_{ii}}{\sum_{j \in s} x_j \omega_{ij}} \right]^2 ,$$

where the ω_{ij} are the usual kernel weights based on the sample values. The following points should be noted:

1. The first and second terms in the expression above are similar to the sample squared error (multiplied by a constant). When this type of crossvalidation is considered for various bandwidths, the shape of the function is similar to that of the sample squared error.

2. The approximation works best as $n_s \to n$, i.e.as the subsample size approaches the sample size. From the numerical work carried out on the hospital dataset , it appears that the shape of this crossvalidation function is similar for the varying proportions.

An illustration of this modified approach is given in Section 5.5 using the hospital dataset of Section 1.4. It appears that the modified crossvalidation has the same problems of smoothing parameter selection as the earlier described crossvalidation methods, for possibly the same reasons.

5.3 Selector methods based on asymptotics

5.3.1 Introduction

The basis for these methods is the asymptotic expansion of bias, variance and MSE terms, to arrive at an asymptotically optimal bandwidth, i.e. as $n \to \infty$, $b \to 0$ and $nb \to \infty$. In this section the asymptotic properties of nonparametric estimators such as the Nadaraya-Watson and Gasser-Müller estimators are outlined. The asymptotic properties of the locally weighted ratio estimator are also given. First some assumptions and definitions are given.

The kernel is chosen to be a continuous, bounded, symmetric real function (having compact support in an interval) and integrating to one. It is said that a kernel is of order κ if it satisfies the following moment conditions:

$$\int_{-1}^{1} K(u)u^j \, du = 0, \; (j = 1, \ldots, \kappa - 1)$$

$$\int_{-1}^{1} K(u)u^\kappa \, du = \alpha \neq 0.$$

The use of higher order kernels provides one generalisaton of kernel estimators; Gasser and Müller (1979) consider these in more detail.

Recall, in Section 3.4, that the form of the weight function proposed by Nadaraya (1964) and Watson (1964) was

$$W_b(x, x_i) = \frac{K(\frac{x-x_i}{b})}{\sum_{i=1}^{n} K(\frac{x-x_i}{b})},$$

where K is the kernel function and b the bandwidth. Gasser and Müller (1979) define a related weight

$$W_b(x, x_i) = n \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{b}\right) du$$

where $s_{i-1} \leq x_i \leq s_i$ and x_i are the ordered values of the explanatory variable, but any sequence $\{s_i\}$ including the x_i could be used.

There is much literature on the Gasser-Müller estimator, which, in some respects, is to be preferred. Properties of the convolution type estimators, such as minimax optimality, are studied by Gasser and Engel (1990). All of the kernel estimators described so far have similar properties for uniformly spaced x_i, but these properties differ in the random design case.

5.3.2 Gasser-Müller estimator

Below we consider the asymptotic bias and variance of this estimator, for fixed and random design densities, but first some assumptions are required (Jennen-Steinmetz and Gasser, 1988):

1. $m(.)$ is κ times continuously differentiable for some $\kappa \geq 2$.

2. $K(u) = 0$ for $|u| > 1$, and $\int K(u)u^j \, du = 1$ for $j = 0, = 0$ for $j = 1, \ldots, \kappa - 1$, and $\neq 0$ for $j = \kappa$.

3. $K(u)$ is Lipschitz continuous.

4. $f(u) > 0$ for $u \in [0, 1]$.

5. $f(u)$ is Lipschitz continuous.

6. $b \to 0$ and $nb \to \infty$ as $n \to \infty$.

Such assumptions are usual in this context. Conditions 1, 3 and 5 above are for smoothness. Assumption 4 guarantees that there are no gaps in the function, and moment conditions in 2 define a kernel of order κ. The last assumption says the bandwidth must shrink with increasing sample size, as the number of points in the smoothing interval increases. The restriction is to univariate x_i, but this may be removed.

The asymptotic mean and variance for the fixed design case are given as:

$$E_\xi(\hat{m}(x)) = \frac{1}{b}\int_0^1 K\left(\frac{x-s}{b}\right) m(s)\, ds + O\left(\frac{1}{n}\right)$$

and

$$\mathrm{var}_\xi(\hat{m}(x)) = \frac{\sigma^2(x)}{nb}\int_{-\tau}^{\tau} K^2(x)\, dx + O\left(\frac{1}{(nb)^2}\right).$$

From these, an expression for the mean squared error (or mean integrated squared error MISE) can be derived; see Gasser and Müller (1979) for proof of above. This is given by

$$E_\xi(\hat{m}(x) - m(x))^2 \doteq \frac{\sigma^2(x)}{nb}\int K(u)^2\, du + \frac{(m''(x))^2 b^4}{4}\left(\int K(u)u^2\, du\right)^2.$$

The asymptotically optimal smoothing parameter b^*, say, with respect to the MSE is obtained by taking the first derivative of MSE with respect to b :

$$b^*_{\mathrm{GM}} = \left(\frac{\sigma^2(x)\int K^2(u)\, du}{(\int K(u)u^2\, du)^2 m''(x_i)^2}\right)^{1/5} n^{-1/5}. \qquad (5.6)$$

Substituting this back into the MSE gives the MSE at the optimal bandwidth value.

For the random design case, the variance is calculated as above with an extra $f(x)$ in the denominator and is multiplied by 2 (Jennen-Steinmetz and Gasser(1988)).

5.3.3 Nadaraya-Watson Estimator

This is motivated as an estimator of a conditional expectation, suggesting its use when the explanatory variable is random. Bierens (1987) reviews the asymptotic properties of this estimator and gives conditions for asymptotic normality. Collomb (1977) also gives an asymptotic evaluation of bias, variance and distribution; under similar assumptions as above for random x_i, the bias and variance are:

$$E_\xi(\hat{m}(x) - m(x)) = \frac{b^2}{2}(m''(x_i) + 2\frac{m'(x_i)f_s'(x_i)}{f(x)}) \int u^2 K(u)\,du + o(b^2) + O\left(\frac{1}{nb}\right)^2$$

and

$$\mathrm{var}_\xi(\hat{m}(x)) = \frac{\sigma^2(x)}{nbf(x)} \int K^2(u)du + o\left(\frac{1}{nb}\right).$$

From these, an expression for the asymptotic MSE is derived and hence the value of the bandwidth which optimises the MSE. This is shown to be proportional to $n^{-1/5}$. The MSE is given by:

$$\frac{\sigma^2(x)}{nbf(x)} \int K^2(u)du + \frac{b^4}{4}\left(m''(x_i) + 2\frac{m'(x_i)f_s'(x_i)}{f(x)}\right)^2 \left(\int K(u)u^2\,du\right)^2$$

for fixed kernel K. The optimal bandwidth value, b^*, from this is:

$$b^*_{\mathrm{NW}} = \left(\frac{\sigma^2(x)\int K(u)^2\,du)}{4f(x)(m''(x_i) + 2\frac{m'(x_i)f_s'(x_i)}{f(x)})^2(\int K(u)u^2\,du)^2}\right)^{1/5} n^{-1/5}. \qquad (5.7)$$

Note the forms of b^*_{GM} and b^*_{NW} differ because of their differing bias and variance terms. This asymptotically optimal value of b^* is used in the *plug-in* method, whereby estimates of the unknown functions $\sigma^2(x), m''(x_i)$ *etc.*, are substituted back into the formula above. The estimates may be derived by further smoothing, leading to a second order bandwidth selection problem. This method achieves the same efficiency as the crossvalidation methods but it has some disadvantages; these include the choice of the preliminary bandwidth values and requiring the underlying regression function to be twice differentiable. For these reasons it is not used as frequently as crossvalidation or the other automatic selection techniques. The automatic methods are easier to use for those unfamiliar with asymptotics.

Comparison of these estimators was given in Chapter 3, in terms of their asymptotic properties and is also discussed by Chu and Marron (1991).

5.3.4 Asymptotic behaviour of the LWRE

In this section, the asymptotic properties of the locally weighted ratio estimator (LWRE) are considered, where n and N increase together such that $n/N \to f$, the sampling fraction, $0 < f \leq 1$. Below we give a theorem for the asymptotic bias and variance, but first some terminology and notation are defined.

Suppose sample, nonsample and population densities are generated by $f_s(x)$, $f_{P-s}(x)$ and $f_P(x)$ respectively, defined by

$$n^{-1}\sum_{j=1}^{n} I(x_j \leq x) \to \int_{-\infty}^{x} f_s(u)\, du$$

and

$$N^{-1}\sum_{i=1}^{N} I(x_i \leq x) \to \int_{-\infty}^{x} f_P(u)\, du,$$

etc. It follows that

$$n^{-1} = f_s(x_i)(x_{i+1} - x_i) + o(n^{-1})$$

and if $u_i = (x_i - x_j)/b$, then

$$\sum_{i=1}^{N} H(u_i)(u_{i+1} - u_i) = \int H(u)\, du + O(N^{-1}b^{-1})$$

for any function H with bounded derivative. We denote

$$\kappa_1 = \int_{-\infty}^{\infty} u^2 K(u)\, du, \qquad \kappa_2 = \int_{-\infty}^{\infty} K^2(u)\, du.$$

Theorem 5.1 *Let $K(u)$ be a symmetric density function with $\int uK(u)\, du = 0$; assume sample and population x are in the interval $[c,d]$ and are generated by the densities f_s and f_P respectively, both bounded away from zero on $[c,d]$, and assumed to have continuous first derivatives. Let $\hat{T}$ be defined as*

$$\hat{T}_{LWRE} = \sum_{i=1}^{N} x_i \frac{\sum_{j=1}^{n} K\left(\frac{(x_i - x_j)}{b}\right) y_j}{\sum_{j=1}^{n} K\left(\frac{(x_i - x_j)}{b}\right) x_j}$$

and suppose $m(x_i)$ has a continuous second derivative; then

$$\begin{aligned} E_\xi(\hat{T}_{LWRE}) - T &= b^2 N\left(\frac{\kappa_1}{2}\right) \int f_P(x)\left[2\frac{f_s'(x)}{f_s(x)}\left(m'(x) - \frac{m(x)}{x}\right) + m''(x)\right] dx \\ &\quad + o(Nb^2 + b^{-1}n^{-1}N) \end{aligned}$$

and

$$var_\xi(\hat{T}_{LWRE}) = N^2 n^{-1} \int \sigma^2(x) f_s(x)^{-1} (f_P(x))^2 \, dx \; + o(n).$$

Proof of Theorem

We assume the conditions for the Gasser-Müller and Nadaraya-Watson estimators, given in Section 5.3.2, hold. Now

$$E_\xi(\hat{T}) = \sum_{i=1}^{N} x_i E_\xi \left(\frac{1/nb \sum_{j=1}^{n} K((x_i - x_j)/b) y_j}{1/nb \sum_{j=1}^{n} K((x_i - x_j)/b) x_j} \right).$$

Suppose the numerator and denominator are written as:

$$\hat{r}(x_i) = \frac{1}{nb} \sum_{j=1}^{n} K((x_i - x_j)/b) y_j$$

and

$$\hat{s}(x_i) = \frac{1}{nb} \sum_{j=1}^{n} K((x_i - x_j)/b) x_j.$$

Then using the result for any function of two variables, (see Kendall and Stuart (1972))

$$E_\xi \left(\frac{\hat{r}(x_i)}{\hat{s}(x_i)} \right) \doteq \frac{E_\xi(\hat{r}(x_i))}{E_\xi(\hat{s}(x_i))} + \frac{E_\xi(\hat{r}(x_i))}{E_\xi(\hat{s}(x_i))} \left[\frac{\mathrm{var}_\xi(\hat{s}(x_i))}{E_\xi(\hat{s}(x_i))^2} - 2 \frac{\mathrm{cov}_\xi(\hat{r}(x_i), \hat{s}(x_i))}{E_\xi(\hat{r}(x_i)) E_\xi(\hat{s}(x_i))} \right], \quad (5.8)$$

where

$$E_\xi(\hat{r}(x_i)) = \frac{1}{b} \left[\int_{-\infty}^{\infty} K((x_i - u)/b) m(u) f_s(u) \, du + o(n^{-1}) + O(n^{-1} b^{-1}) \right]$$

and

$$E_\xi(\hat{s}(x_i)) = \frac{1}{b} \left[\int_{-\infty}^{\infty} K((x_i - u)/b) u f_s(u) \, du + o(n^{-1}) + O(n^{-1} b^{-1}) \right].$$

Then considering the leading term in (5.8) and using a change of variable

$$E_\xi \left(\frac{\hat{r}(x_i)}{\hat{s}(x_i)} \right) \approx \frac{\int K(u) m(x_i - ub) f_s(x_i - ub) \, du + o(n^{-1} b^{-1})}{\int K(u)(x_i - ub) f_s(x_i - ub) \, du + o(n^{-1} b^{-1})}. \quad (5.9)$$

But

$$\int K(u) m(x_i - ub) f_s(x_i - ub) \, du$$

$$\begin{aligned}
&= \int K(u)m(x_i)f_s(x_i)\,du \\
&\quad +b^2 \int u^2 K(u)\,du \left(f_s'(x_i)m'(x_i) + \frac{m''(x_i)f_s'(x_i)}{2} + \frac{f_s''(x_i)m(x_i)}{2} \right) \\
&\quad +o(n^{-1}b^{-1}) + o(b^2),
\end{aligned}$$

by Taylor series expansion for a twice differentiable function. A similar expansion can be obtained for

$$\int K(u)(x_i - ub)f_s(x_i - ub)\,du.$$

Then from (5.9),

$$\begin{aligned}
E_\xi\left(\frac{\hat{r}(x_i)}{\hat{s}(x_i)}\right) &= \left\{ \frac{m(x_i)}{x_i} + \frac{b^2\kappa_1}{f_s(x_i)x_i}\left(f_s'(x_i)m'(x_i) + m''(x_i)\frac{f_s(x_i)}{2} + \frac{m(x_i)f_s''(x_i)}{2} \right) \right\} \times \\
&\quad \left\{ 1 - \frac{b^2\kappa_1}{x_i f_s(x_i)}\left(f_s'(x_i) + f_s''(x_i)\frac{x_i}{2} \right) \right\} + o(n^{-1}b^{-1}) + o(b^2) \\
&= \frac{m(x_i)}{x_i} + \frac{b^2\kappa_1}{f_s(x_i)x_i}\left\{ f_s'(x_i)\left(m'(x_i) - \frac{m(x_i)}{x_i} \right) + m''(x_i)\frac{f_s(x_i)}{2} \right\} \\
&\quad +o(b^2) + o(n^{-1}b^{-1}). \qquad (5.10)
\end{aligned}$$

So an expression for the asymptotic expectation is:

$$\begin{aligned}
E_\xi(\hat{T}) &= \sum_{i=1}^{N} m(x_i) + b^2 \sum_{i=1}^{N} \frac{\kappa_1}{f_s(x_i)}\left\{ f_s'(x_i)\left(m'(x_i) - \frac{m(x_i)}{x_i} \right) + \frac{m''(x_i)f_s(x_i)}{2} \right\} \\
&\quad +o(b^2 N) + o(n^{-1}b^{-1}N). \qquad (5.11)
\end{aligned}$$

The term for the asymptotic bias is obtained by taking the first term on the right hand side above to the left hand side.

The proof for the variance follows a similar line to that given in Dorfman (1993). We write

$$\begin{aligned}
\text{var}_\xi(\hat{T}) &= \text{var}_\xi\left(\sum_{j=1}^{n} y_j \left(\sum_{i=1}^{N} \frac{x_i K((x_i - x_j)/b)}{\sum_{k\in s} x_k K((x_i - x_k)/b)} \right) \right) \\
&= \text{var}_\xi\Big(\sum_{j=1}^{n} y_j W_j\Big)
\end{aligned}$$

where

$$W_j = \sum_{i=1}^{N} \frac{x_i K((x_i - x_j)/b)}{\sum_{k\in s} x_k K((x_i - x_k)/b)}. \qquad (5.12)$$

Then the variance

$$\begin{aligned}\mathrm{var}_\xi(\sum_{j=1}^{n} y_j W_j^2) &= \sum_{j=1}^{n} W_j^2\,\mathrm{var}_\xi(y_j)\\ &= \sum_{j=1}^{n} W_j^2\sigma^2(x_j). \qquad (5.13)\end{aligned}$$

An asymptotic expression for W_j can be extended to one for W_j^2 and hence we can write down the asymptotic variance. With a change of variable, $u_k = (x_i - x_k)/b$,

$$\begin{aligned}&\textstyle\sum_{k\in s} x_k K((x_i - x_k)/b) = \sum_{k\in s} K(u_k)(x_i - u_k b)\\ &= n\textstyle\sum_{k\in s} K(u_k)(x_i - u_k b)\,(f_s(x_k)(x_{k+1} - x_k) + o(n^{-1}))\\ &= nb\left[\textstyle\sum_{k\in s} K(u_k)(u_{k+1} - u_k)(x_i - u_k b)[f_s(x_i) - u_k h f_s'(x_i) + \ldots] + o(1)\right]\\ &= nb\,[f_s(x_i)x_i + o(1)]\end{aligned}$$

So

$$\begin{aligned}W_j &= \sum_{i=1}^{N} \frac{x_i K((x_i - x_j)/b)}{nb\,(f_s(x_i)x_i + o(1))}\\ &= \sum_{i=1}^{N} \frac{K((x_i - x_j)/b)}{nbf_s(x_i)}\,(1 + o(1))\\ &= \frac{N}{n}\left[\int \frac{K((u - x_j)/b)f_P(u)}{bf_s(u)}\,du + o(1)\right]. \qquad (5.14)\end{aligned}$$

Substituting v for $(u - x_j)/b$ in the above and taking Taylor series expansions we arrive at, as $b \to 0$,

$$W_j = \frac{Nf_P(x_j)}{nf_s(x_j)} + o(1).$$

Substituting this back into formula (5.13), for the variance, we have

$$\begin{aligned}\mathrm{var}_\xi(\hat{T}) &= \sum_{j=1}^{n} \sigma^2(x_j)\left(\frac{Nf_P(x_j)}{nf_s(x_j)}\right)^2 + o(n)\\ &= N^2 n^{-1}\int \sigma^2(x)f_s(x)^{-1}[f_P(x)]^2\,dx + o(n). \qquad (5.15)\end{aligned}$$

The bias term above depends not only on the first and second derivatives of the underlying regression function but also on $f_s(x)$ and its first derivative. There is a dependence on the slope of $m(x)$, in $m'(x_i)$, and its curvature, in $m''(x_i)$, in this bias expression. When $m(x_i) = \beta x_i$, the linear regression model with no intercept term, the bias term is zero, which is what we expect.

The first term in the bias consists of a b^2 term, and the first term in the variance above is constant. For calculating the mean squared error of this estimator, which involves variance and squared bias, it would be more accurate to take the variance up to terms of order b^4 making it comparable with the squared bias. This can be done but leads to a complicated expression for the asymptotic variance, which is of no practical use.

A globally optimal bandwidth, b_{opt}, is normally obtained by minimising the asymptotic mean squared error. This involves unknown functions of the design densities $f_s(x)$, $f_P(x)$, the underlying superpopulation model $m(x)$, the variance function $\sigma^2(x)$ and their derivatives. A 'plug-in' approach does not seem feasible here, since the first term in the variance does not involve b.

5.3.5 Local or variable bandwidth

Recently there has been some attention in the literature on bandwidths in kernel estimation which vary locally with the x values; the running line estimator is an example of an estimator with a varying bandwidth. The paper by Jones (1990) makes the distinction between two 'variable' methods that could be employed in kernel density estimates and which also apply to the regression setting. One method, specifically referred to as the 'variable' bandwidth approach, reflects a variable amount of smoothing at each of the sample data points used in the estimator, e.g. for the Nadaraya-Watson estimator

$$m_V(x) = \frac{\sum_{j=1}^{n} K\left(\frac{(x-x_j)}{b(x_j)}\right) y_j}{\sum_{j=1}^{n} K\left(\frac{(x-x_j)}{b(x_j)}\right)}.$$

This type of 'variable' bandwidth is discussed by Abramson (1982), in density estimation, where it is shown that $b(x_j) \propto f(x_j)^{-\frac{1}{2}}$ gives good results. Also Hall (1990) discusses variable bandwidths in reducing bias in kernel regression and Fan and Gijbels (1992) in their local linear smoother.

The other, 'local' bandwidth method, relates to the target value itself. In this case the bandwidth is allowed to vary with the local density at the x value, e.g, for the Nadaraya-Watson estimator

$$m_L(x) = \frac{\sum_{j=1}^{n} K\left(\frac{(x-x_j)}{b(x)}\right) y_j}{\sum_{j=1}^{n} K\left(\frac{(x-x_j)}{b(x)}\right)}.$$

Müller and Stadtmüller (1987) consider local bandwidth kernel regression estimators for fixed designs, and show how these are superior to global bandwidth estimators, in terms of asymptotic mean squared error, for optimally chosen bandwidths. Local bandwidth selection for kernel regression estimators has also been addressed by Staniswallis (1989).

An intuitive form for the local bandwidth is

$$b(x) \propto \left[\frac{\sigma^2(x)}{f(x)}\right]^\alpha ,$$

where $0 \leq \alpha < 1$, implying use of a large bandwidth in areas of low density and large conditional variance. The optimal choice of the local or variable bandwidth is obtained by minimising the conditional asymptotic mean squared error of the estimator. A local bandwidth is considered in an example in Section 5.5.

5.4 Other methods of selection

5.4.1 Equating degrees of freedom

This is a simple way of using the sample data to choose a value of the smoothing parameter. Degrees of freedom forms have already been discussed in Sections 3.6 and 4.4. The formulae for the degrees of freedom depend on the smoother matrix and, in the case of the alternative form for the degrees of freedom derived, the diagonal matrix with variances of y_i as the diagonal entries.

If we fix the degrees of freedom, say, to the number of strata used in stratification, then for any of the linear smoothers we can calibrate the smoothing parameter to a particular number of degrees of freedom. This makes it possible, for example, to compare nonparametric regression estimators with each other and with the separate ratio estimator. It is necessary to try a range of values of the smoothing parameter in the degrees of freedom formula in order to arrive at the one that gives the required degrees of freedom, i.e. to employ a numerical search.

If the purpose of any experiment is to compare between estimators, such as those described, this is possibly the easiest and most direct method of calibrating the smoothing parameter.

5.4.2 Methods specific to the locally weighted ratio estimator

A large b approximation

In this section other methods of smoothing parameter selection are briefly described. These are specific to the LWRE with a Gaussian weight function. The first involves a large b approximation of the model mean squared error.

The Gaussian weight function can be expanded. For example, if $G_{ij} = -k(x_i - x_j)^2/2$ say, where $k = 1/b^2$, then

$$W_b(x_i, x_j) = \exp(G_{ij}) = 1 + G_{ij} + \frac{G_{ij}^2}{2!} + \cdots .$$

Using this expansion, a large b (or small k) approximation to the mean squared error can be derived. This approximation is considered for terms up to and including k^2 only. The mean squared error is given as: $a_1 + a_2k + a_3k^2$, where a_1, a_2 and a_3 are coefficients obtained from the expansion. These coefficients are given in Appendix A. Minimising the approximate mean squared error with respect to b (or k) gives:

$$b_{opt}^2 = \frac{-2a_3}{a_2}, \tag{5.16}$$

An approximate method for selecting a bandwidth value in this particular case requires us to calculate the coefficients a_2 and a_3.

The denominator term, a_2 above, equals zero when either of the following is true:

(a)

$$\frac{\sum_{j\in s} m(x_j)}{\sum_{j\in s} x_j} = \frac{\sum_{i=1}^N m(x_i)}{\sum_{i=1}^N x_i} \qquad \text{or}$$

(b)

$$\sum_{i=1}^N x_i \sum_{j\in s}(x_i - x_j)^2 m(x_j) \sum_{k\in s} x_k = \sum_{i=1}^N x_i \sum_{j\in s}(x_i - x_j)^2 x_j \sum_{k\in s} m(x_k).$$

When this is the case, the optimal bandwidth, $b_{opt} = \infty$. The first condition, (a), is true if the sample chosen is balanced or $m(x_j) = \beta x_j$, the linear model through the origin. The second condition, (b), is true for $m(x_j) = \beta x_j$ also.

These conditions indicate that the ratio estimator is the optimal estimator of the population total when we have either a balanced sample or an underlying superpopulation which is linear through the origin. We cannot expect to improve on the ratio estimator in these cases, but when we move away from balanced sampling and the linear superpopulation model, the optimal b starts to move away from $b = \infty$ to a smaller value. This is when the smoothing approach is of greatest benefit.

The form of b_{opt}, given in (5.16), is a complicated expression involving the sample and population x_i; it might, however, prove useful in finding an optimal value of b for any sample at hand. Other kernel weight functions could be used and expanded as above, leading to a slightly different bandwidth formula.

Approximating the predicted value by a quadratic.

Here the predicted value, $\hat{y}_i$, appearing in the estimator is approximated by a quadratic in x_i. This is made possible by the expansion of the Gaussian weight function, as described above, and then by the expansion of the LWRE as a function in x_i. A better approximation is obtained using the square root of the predicted value instead; this ensures the x_i terms appearing in the expansion are orthogonal to one another. For example,

$$\begin{aligned}\sqrt{\hat{y}_i} \doteq x_i^{\frac{1}{2}} \quad & \frac{(\sum y_j)^{\frac{1}{2}}}{(\sum x_j)^{\frac{1}{2}}} \left\{1 + \frac{k}{4}\left(\frac{\sum x_j^3}{\sum x_j} - \frac{\sum x_j^2 y_j}{\sum y_j}\right)\right\} \\ & -x_i^{\frac{3}{2}} \frac{(\sum y_j)^{\frac{1}{2}}}{(\sum x_j)^{\frac{1}{2}}} \frac{k}{2} \left\{\frac{\sum x_j^2}{\sum x_j} - \frac{\sum x_j y_j}{\sum y_j}\right\},\end{aligned} \tag{5.17}$$

where all sums are over $j \in s$ and $k = 1/b^2$. An appropriate model is fitted, using the sample data, to obtain the coefficients of $x_i^{1/2}$ and $x_i^{3/2}$ appearing in (5.17) above. The coefficients from the fitted model and (5.17) are then equated to give an approximate bandwidth value. This can be a useful starting point for the choice of the smoothing parameter.

5.5 Examples

In this section we describe some examples used to illustrate smoothing parameter selection. The first example is related to the crossvalidation methods we have described, in particular the *total* and *pointwise* crossvalidation criteria, CV_1 and CV_2, above and the methods based on modified crossvalidation. The second example is used to compare some of the nonparametric regression estimators already described, and in particular to introduce a local bandwidth into the kernel estimator.

5.5.1 Example 1

The following example is based on the hospitals dataset of Section 1.4. We selected 100 samples of size 200 (using simple random sampling without replacement) from the population. For each sample the following quantities were calculated, using the LWRE with Gaussian weight for a range of bandwidth values: the squared error for the population total, the squared error for non-sample total, and the crossvalidation criteria CV_1 and CV_2 above. These quantities were then averaged over the 100 random samples and a plot of these average values can be found in Figure 5.1. The squared error functions can be very different for different samples drawn from the population, so by averaging we lose some of the variability between samples.

Figures 5.2-5.4 give three plots for three separate samples. For each, the nonsample squared error, crossvalidation criteria CV_1 and CV_2 and the modified crossvalidation criterion CV_4 for varying proportions are plotted against bandwidth. We note that the modified crossvalidation appears to behave the same way for all samples even though the squared error plots suggest otherwise. The modified crossvalidation methods based on the crossvalidation method, CV_1, all tend to increase then decrease again with bandwidth, as the sample squared error would do. The modified crossvalidation method based on CV_2 behaves differently, being minimised at a small value of the bandwidth.

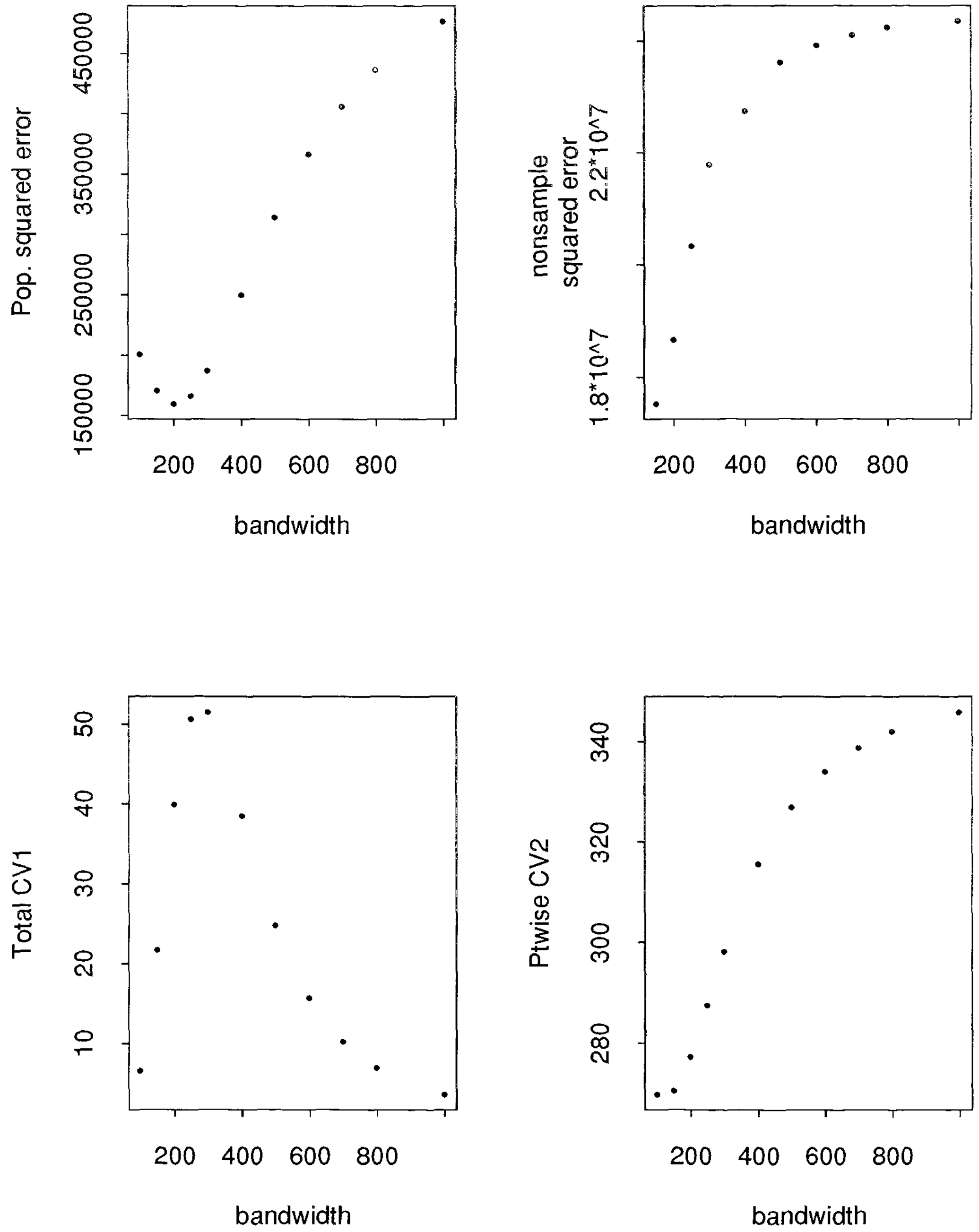


Figure 5.1: Plots of squared error and crossvalidation criterion 1 and 2, averaged over 100 random samples

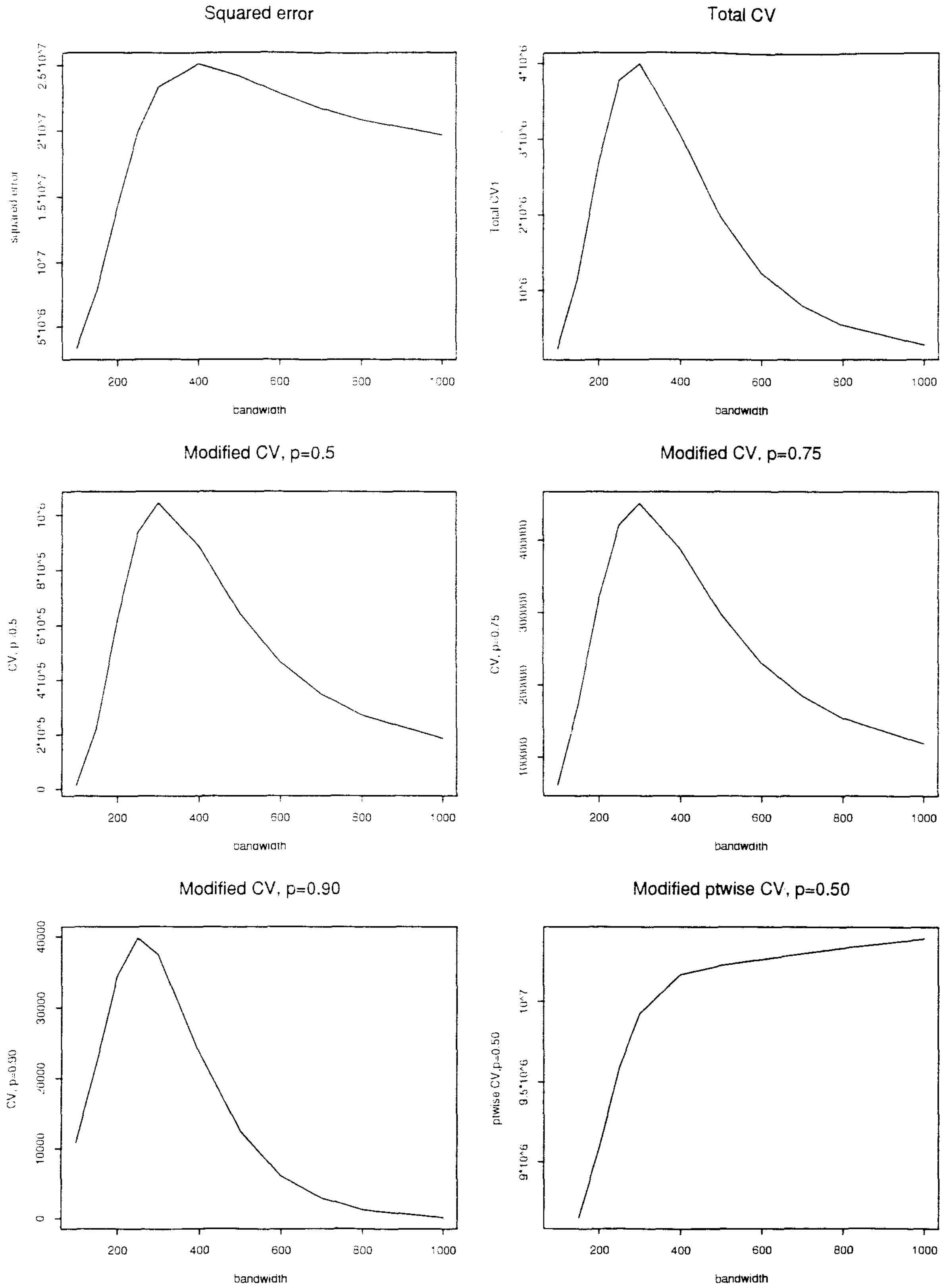


Figure 5.2: Plots of squared error, crossvalidation criterion 1 and modified crossvalidation for sample 1

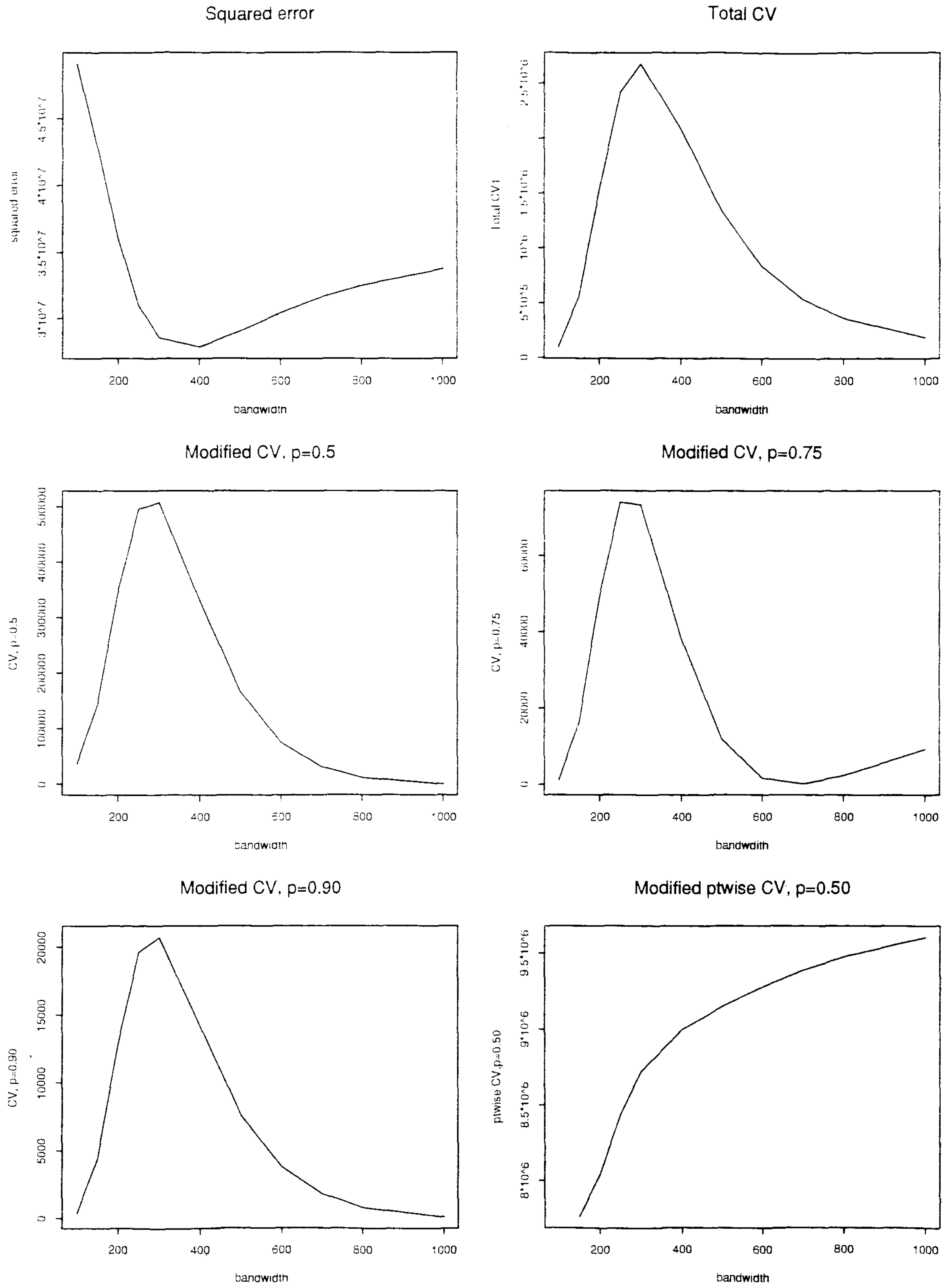


Figure 5.3: Plots of squared error, crossvalidation criterion 1 and modified crossvalidation for sample 2

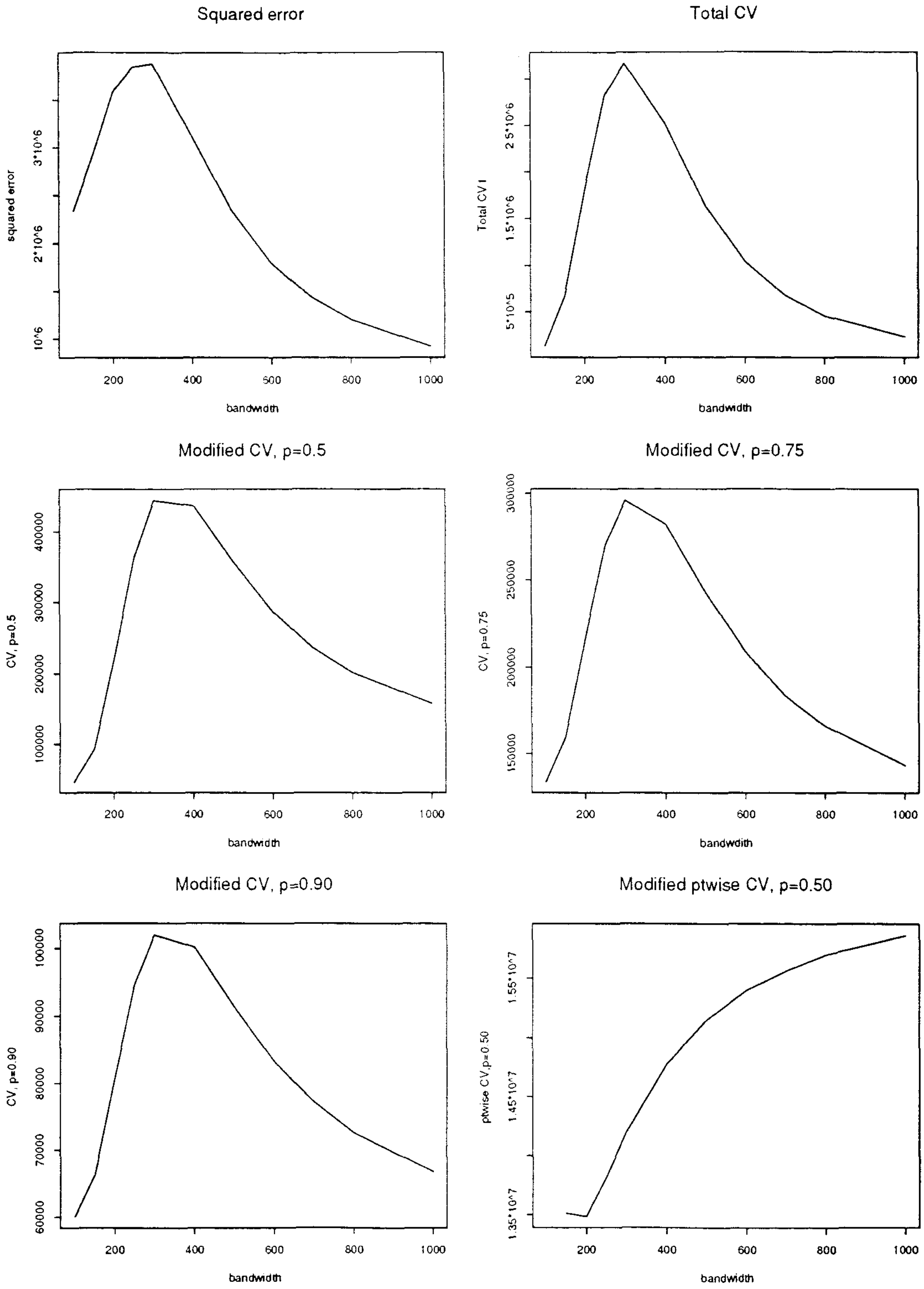


Figure 5.4: Plots of squared error, crossvalidation criterion 1 and modified crossvalidation for sample 3

5.5.2 Example 2

Background

In order to investigate the squared error properties of the estimators described so far we conduct experiments for two finite populations. It is shown that the nonparametric regression estimators provide improvements on the ratio and separate ratio estimator for the two finite populations. A local and global bandwidth are included for the locally weighted ratio estimator to compare their performance. The experiment is outlined below.

(i). For each population, 100 random samples of size n were selected from a population of size N following an optimal allocation rule for stratified random sampling.

(ii). For each random sample drawn we predicted a non-sample total using the following estimators:

1. a ratio estimator,
2. a separate ratio estimator,
3. a locally weighted ratio estimator (LWRE) with a Gaussian weight function having global and local bandwidths,
4. a running ratio estimator (RRE) for a range of span values, and
5. an averaged running ratio estimator (ARRE) for a range of spans.

For each the corresponding squared error associated with predicting the non-sample total was also calculated.

The local bandwidth, in (3) above, was chosen motivated by work on asymptotics. If $f_{P-s}(x_i)$ denotes the design density of non-sample values and $\mathrm{var}(y_i|x_i) \propto x_i$ then one might try

$$b(x_i) \propto \left(\frac{x_i}{f_{P-s}(x_i)}\right)^{(1/5)}.$$

This suggests a bandwidth which increases with increasing x_i (variance) and in regions where the design density is low, and decreases in regions where design density is high. This is an example of a simple version of a local bandwidth. This simple local bandwidth was found to perform satisfactorily.

(iii). The average squared error over all 100 samples (and the corresponding standard error of the mean) was calculated for the various estimators, at the various smoothing parameter values.

Below are the two examples: the first, an actual dataset of observations from the hospital survey referred to in Section 1.4, the other based on a simulation. In each case the smoothing approach improves on the ratio estimator and separate ratio estimator, for certain values of the smoothing parameter, when averaging over 100 samples.

The program to perform the above was written in Sun Fortran.

Hospital example

Data from the hospital survey, described in Section 1.4, was used. From the population of $N = 393$ observations samples of size 200 were selected, and 6 strata used. The plots of the average squared error over 100 samples are given in Figure 5.6, for each of the estimators: LWRE with Gaussian kernel weight using a global and local bandwidth, running ratio estimator, and averaged running ratio estimator. Also included on each plot is a line indicating the position of the average squared error for the separate ratio estimator.

The average squared error for ratio estimator is 6.55×10^7 (standard error 3.5×10^6), and for the separate ratio estimator 1.12×10^7 (standard error 1.1×10^6).

The density estimate in the local bandwidth was based on a kernel density estimate with Gaussian weight using a bandwidth of 200.0. The results are also given in Table 5.1. The values in the brackets are the standard errors associated with the average squared error. Also given is the ratio of averaged squared error of the estimator to the averaged squared error of the separate ratio estimator.

estimator(with bandwidth)	Ave squared error/10^7	ASE($\hat{T}$)/ASE($\hat{T}_{SRE}$)
Separate ratio estimator	1.12(0.11)	1.00
Ratio estimator	6.55(0.35)	5.87
ARRE		
1	1.61(0.18)	1.44
10	1.10(0.11)	0.98
30	1.04(0.11)	0.93
40	1.04(0.11)	0.93
50	1.06(0.11)	0.95
70	1.11(0.12)	0.99
90	1.20(0.13)	1.08
150	1.67(0.17)	1.59
Running ratio est.		
1	1.61(0.18)	1.44
5	1.36(0.13)	1.22
10	1.14(0.11)	1.02
20	1.05(0.10)	0.94
30	1.04(0.11)	0.93
50	1.02(0.11)	0.91
70	1.06(0.11)	0.95
100	1.11(0.12)	1.00
200	2.38(0.20)	2.13
LWRE Gaussian wt:		
global bandwidth		
20	1.11(0.11)	1.00
30	1.08(0.11)	0.97
50	1.06(0.11)	0.95
60	1.06(0.11)	0.95
75	1.07(0.11)	0.96
100	1.12(0.12)	1.00
150	1.32(0.13)	1.18
200	1.64(0.16)	1.46
300	2.51(0.21)	2.25
Local bandwidth		
3.0	1.18(0.11)	1.06
5.0	1.12(0.11)	1.00
10.0	1.06(0.11)	0.95
20.0	1.04(0.11)	0.93
30.0	1.04(0.11)	0.93
40.0	1.08(0.11)	0.96
50.0	1.17(0.12)	1.05
100.0	2.36(0.20)	2.12

Table 5.1: Averaged squared error results for the hospital dataset.

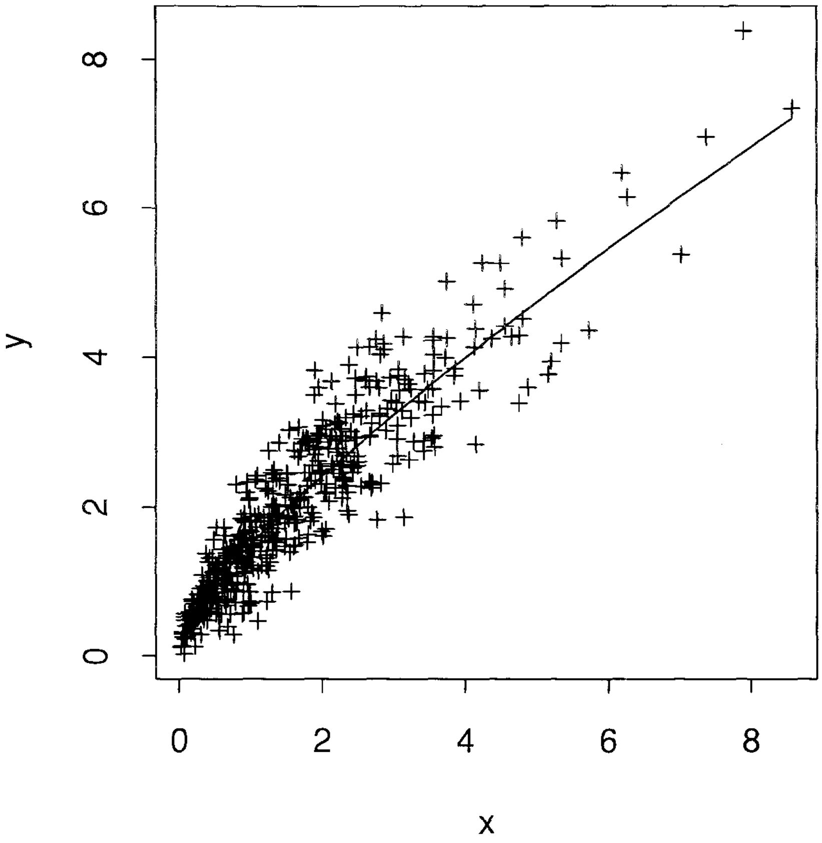


Figure 5.5: A plot of the simulated dataset including the true underlying model

Simulated dataset

The second example is a simulated dataset of 500 observations based on the following model:

$$y_i = m(x_i) + \epsilon_i$$

where $x_i \sim \text{Gamma}(1.6)$,

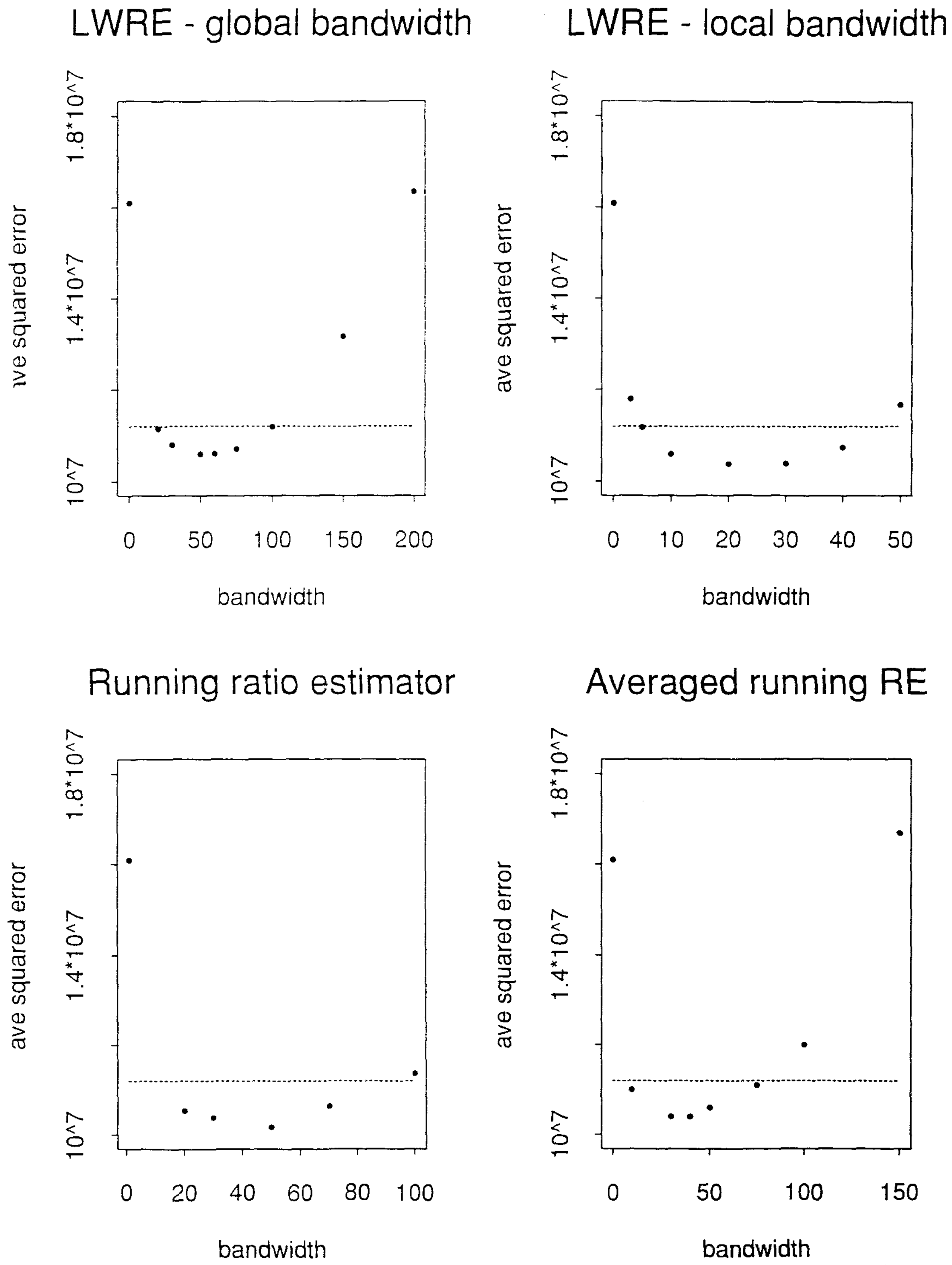
$$m(x_i) = \sqrt{x_i} + x_i/2$$

and

$$\epsilon_i \sim N(0, 0.4\sqrt{x_i}).$$

Samples of size 200 were selected based on 5 strata. A plot of these data are given in Figure 5.5. Plots of the averaged squared error against smoothing parameter are given for the various estimators in Figure 5.7.

The averaged squared error for the ratio estimator is 3523.75 (standard error 100.5), and for the separate ratio estimator 186.00 (standard error 16.4).


LWRE - global bandwidth

LWRE - local bandwidth

Running ratio estimator

Averaged running RE

Figure 5.6: Average squared error plots for the hospital dataset

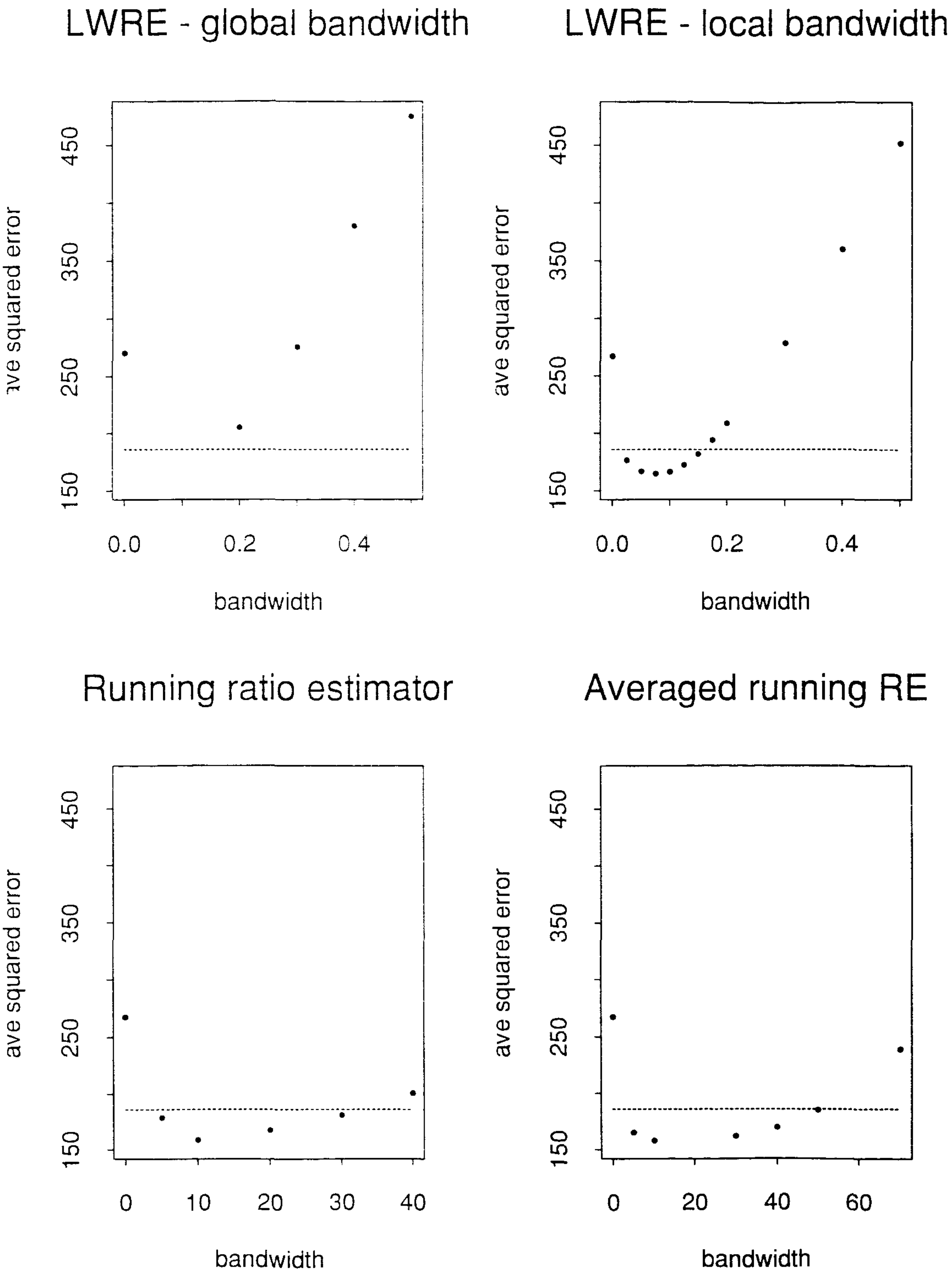


Figure 5.7: Average squared error plots for the simulated dataset

Summary

The plots in Figures 5.6 and 5.7 show that the various nonparametric estimators are, for some values of the smoothing parameter, more efficient than the ratio and separate ratio estimators over repeated sampling. The horizontal line on the plots indicates the averaged squared error of the separate ratio estimator. The averaged running ratio estimator (ARRE) performs particularly well as does the running ratio estimator (RRE) and LWRE with Gaussian weight based on a local bandwidth. This is possibly because of the varying bandwidth used with these estimators, which takes account of underlying changes in the x variable. The LWRE, based on a global bandwidth, does not perform as well when compared with the separate ratio estimator, because of bias problems. This is due to its lack of 'total-preservation', a property which the ratio estimators and the averaged running ratio estimator all have (see Section 4.3).

Table 5.1 gives the empirical results of the simulation based on the Hospital dataset. The second column gives the measure of averaged squared error over 100 randomly drawn samples from the population, and the third column the ratio of this measure for each estimator with that of the separate ratio estimator. As figures 5.6 and 5.7 illustrate the nonparametric regression estimators are always more efficient than the ratio estimator and, for some values of the smoothing parameter, are more efficient than the separate ratio estimator. The standard errors associated with these measures for the nonparametric regression estimators are comparable with the separate ratio estimator but the ratio estimator stands out as being the worst estimator in this example. There is certainly much to be gained, in terms of efficiency from using nonparametric regression estimators in preference to both the ratio and separate ratio estimators.

Introducing a π weight into the estimators to ensure design unbiasedness is discussed in Chapter 6.

5.6 Discussion

In this chapter some possible methods of choosing the smoothing parameter have been discussed. In some applications of nonparametric regression, the subjective choice of comparing curves with varying smoothing parameters and choosing the one that looks the best fit to the data, is sufficient. The problem of predicting a finite population total may require a more objective choice of the smoothing parameter. Many of the methods described in the literature are based on a choice for the regression function and not on a function of the response such as a total or a mean. The standard automatic method, of crossvalidation, has not performed particularly well in this prediction setting. It tends to choose a smoothing parameter that is either too small or too large, mimicking the behaviour of the sample squared error. This is not particularly satisfactory since the optimal choice, in terms of minimising mean squared error, lies somewhere between the two extremes. Modifications of crossvalidation, by removing a proportion of the observations and using the remaining to predict, do not dramatically improve on the standard crossvalidation method either. This is one possible area that requires further research.

The asymptotic bias and variance results, however, show more promise. A local bandwidth is useful since it allows the window width to vary with the underlying design density and possibly other factors such as variance. A varying bandwidth is a feature implicit in the running estimators described and in smoothing splines (Silverman, 1985). We note that, although the asymptotic results can lead to some useful bandwidth selectors, in practice these are difficult to achieve because of the unknown functions of the underlying regression model.

Finally in the last section, some other possible methods of smoothing parameter selection were discussed. These included equating degrees of freedom, particularly useful in calibrating between estimators when comparisons are being made. Also methods based on a large b approximation of the model mean squared error and approximating the estimator $\hat{y}_i$ to a quadratic in x_i were considered; these are specific methods related to the locally weighted ratio estimator with Gaussian weight.

In conclusion, there does not appear to be one method of smoothing parameter selection that stands out as being the best method to use in this problem, except perhaps the degrees of freedom method for its ease of computation. Crossvalidation does not work particularly well. Asymptotics tend to rely heavily on the knowledge of certain functions of the data and their derivatives, in particular the unknown regression model. These functions can be estimated, but then we have an additional estimation problem. The degrees of freedom method is possibly the easiest to use and relates to the parametric regression setting well. The methods described so far have not been exhaustive and from this work we see much scope for further research into the important question of smoothing parameter selection.

Chapter 6

Properties under repeated sampling

6.1 Introduction

In Chapter 2, the notion of a sampling design was introduced. The probability function involved plays a central role in determining properties such as the mean and the variance of random quantities calculated from a sample. Common examples of designs used are simple random sampling (SRS) and stratified simple random sampling (SSRS).

Design-based properties, such as the bias and variance, can be calculated for each of the estimators mentioned so far. Some of the estimators are unbiased or approximately unbiased under repeated sampling. If the nonparametric estimators could be modified in some way, to remove or reduce the bias, then this would lead to a useful class of estimators of the population total. It is possible to do this, by introducing a weight associated with each sample observation; these weights are the reciprocals of the sample inclusion probabilities.

6.2 π-weighted estimators

The notion of a π-weighted estimator is briefly described in Section 2.4. It is derived by weighting each sample observation appearing in the estimator by the inverse of its inclusion probability, i.e. $1/\pi_j$, $j = 1, \ldots, n$. The simplest example is the Horvitz-Thompson estimator, already mentioned in Section 2.4.

Introducing inclusion probabilities as weights ensures approximate unbiasedness under repeated sampling. If the estimator we have chosen to weight in this way is total-preserving then it will be shown, in Section 6.3, that such π-weighted estimators are approximately design unbiased. The π-weighting increases the importance of elements in the sample, so that each sample value represents $1/\pi_j$ of the population values.

The inclusion probabilities described depend on the underlying design that has been adopted. Stratified simple random sampling with optimal allocation is one design we are interested in. The weights associated with a sample observation differ between strata, but are constant within strata.

Estimators which are unbiased under repeated sampling, i.e.

$$E_{\pi}(\hat{T}) - T = \sum_{s \in \mathcal{S}} \pi(s)\hat{T} - T = 0,$$

or have small design mean squared error,

$$E_{\pi}(\hat{T} - T)^2 = \sum_{s \in \mathcal{S}} \pi(s)(\hat{T} - T)^2$$

are particularly useful. Under simple random sampling (SRS), the ratio estimator has bias of order $1/n$ and under stratified SRS it remains biased (also order $1/n$, see Cochran, 1977). Introducing inclusion probabilities as weights in the ratio estimator, for designs other than simple random sampling where the π_is are equal, ensures approximate design unbiasedness.

The approximate bias and variance of some π-weighted estimators are now found, under any design. The estimators considered include the locally weighted ratio estimator, weighted according to a kernel function or Uniform k-nearest neighbour. Similar results for the ratio estimator are given by Cochran (1977) and by Särndal, Swensson and Wretman (1992). A similar line of argument is followed to that given in Särndal et al. (1992), in order to derive expressions for approximate bias and variance of the π-weighted estimators under any design we choose.

In the first instance, we concentrate on a locally weighted generalised linear model estimator, and then obtain results for the ratio and regression estimators as special cases of this. One or more explanatory variables are considered, by introducing matrix notation into the calculations.

6.3 A π-weighted local GLM estimator.

6.3.1 Introduction

Särndal et al. (1992) consider the following for the regression estimator of the population total:

$$\hat{T}_{LR} = \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})'\hat{\mathbf{B}}, \tag{6.1}$$

where the $\hat{t}_{y\pi}$, $\hat{\mathbf{t}}_{x\pi} = (t_{x_1\pi}, \ldots, t_{x_p\pi})$ denote the π-weighted sample totals of y and the p regressor or explanatory variables respectively. This is an example of a *difference estimator.* Also

$$\hat{\mathbf{B}} = \hat{\mathbf{C}}^{-1}\hat{\mathbf{t}}_0,$$

where

$$\hat{\mathbf{C}} = \sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k \pi_k}$$

and

$$\hat{\mathbf{t}}_0 = \sum_{k \in s} \frac{\mathbf{x}_k y_k}{\sigma_k^2 \pi_k}. \tag{6.2}$$

In order to include multiple explanatory variables, the above has been written in matrix notation as in Särndal et al. (1992). For instance, $\mathbf{x}_k = (1, x_k)'$, and $\mathbf{B}$=$(B_1, B_2)'$ corresponds to a linear regression model with an intercept term.

The above can be extended to a generalised linear model estimator. The estimating equations for estimating the parameters in a generalised linear model can be written as:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ir}}{V(\mu_i)g'(\mu_i)} = 0, \qquad (r = 1, \ldots, p)$$

where $g(\mu_i) = \eta_i$, the linear predictor, and $V(\mu_i)$, is the variance function. This variance function might, for example, be proportional to the mean.

This is an example of when survey sampling becomes model assisted, and inferences based on surveys may also depend on the underlying superpopulation model.

It should be noted here that in order to keep the estimation of parameters non-iterative, the generalised linear model estimators discussed so far are restricted to a certain class of models. Models with Normal errors, or Poisson errors provided an offset $\log(x_i)$ and intercept term are included in the linear predictor, are considered since they ensure non-iterative estimation with an appropriate, usually the canonical, link function.

Some examples of these models are outlined below.

1. For the ratio estimator; we assume an identity link, $g(\mu_i) = \mu_i = \beta x_i$ and $\text{var}_\xi(\mu_i) \propto x_i$.

2. For linear regression (with an intercept term); we assume an identity link again, $g(\mu_i) = \mu_i = \beta_0 + \beta_1 x_i$, and $\text{var}_\xi(\mu_i) = \sigma^2$, a constant.

So far a global estimator of the population total has been considered. Attention is now focused on the case of a locally weighted generalised linear model estimator, which is an extension of the global case. A locally weighted generalised linear model estimator may be written as:

$$g(\mu_i) = \mathbf{x}_i^{'}\hat{\mathbf{B}}_i. \tag{6.3}$$

The $\hat{\mathbf{B}}_i$ is the locally weighted version of the $\hat{\mathbf{B}}$ described above in (6.2), for the global estimator. Each $\hat{B}_{pi}$ corresponds to a locally weighted parameter estimate associated with the pth explanatory variable, including constants. In this case

$$\hat{\mathbf{B}}_i = \hat{\mathbf{C}}_i^{-1}\hat{\mathbf{t}}_{0i}$$

and the corresponding elements of the matrices $\hat{\mathbf{C}}_i$ and $\hat{\mathbf{t}}_{0i}$ above are

$$\hat{c}_{rr'i} = \sum\nolimits_{k\in s} \frac{W_b(x_i,x_k)x_{rk}x_{r'k}}{\sigma_k\pi_k g'(\mu_k)}, \qquad r \neq r' = 1,\ldots,p \tag{6.4}$$

$$\hat{t}_{r0i} = \sum\nolimits_{k\in s} \frac{W_b(x_i,x_k)x_{rk}y_k}{\sigma_k\pi_k g'(\mu_k)}, \qquad r = 1,\ldots,p. \tag{6.5}$$

If we assume case 1 above, then $\hat{\mathbf{B}}_i = \hat{B}_i$, and the parameter estimate for the locally weighted ratio estimator becomes

$$\hat{B}_i = \frac{\hat{t}_{iy\pi}}{\hat{t}_{ix\pi}} = \frac{\sum_{k\in s} W_b(x_i, x_k)y_k/\pi_k}{\sum_{k\in s} W_b(x_i, x_k)x_k/\pi_k}.$$

6.3.2 Approximate bias under repeated sampling

The approximate bias, under repeated sampling, of the locally weighted generalised linear model estimators, (6.3) above, is now considered.

Theorem 6.1 *The approximate bias of the locally weighted GLM estimator is*

$$E_{\pi}(\sum_{i=1}^{N} g(\hat{y}_i)) - \sum_{i=1}^{N} g(y_i) \doteq \sum_{i=1}^{N}\sum_{r=1}^{p} B_{ri}x_{ri} - \sum_{i=1}^{N} g(y_i).$$

Proof of theorem

By Taylor series linearisation, $\hat{\mathbf{B}}_i$ in (6.3) is approximated by

$$\hat{\mathbf{B}}_i \doteq \mathbf{B}_i + \mathbf{C}_i^{-1}\left(\hat{\mathbf{t}}_{0i} - \hat{\mathbf{C}}_i\mathbf{B}_{\mathrm{i}}\right). \tag{6.6}$$

The proof of this is quite easy to derive, and is outlined below; the global case is given by Särndal et al. (1992).

The following proof is another application of the general Taylor series linearisation technique. The estimator

$$\hat{\mathbf{B}}_i = \hat{\mathbf{C}}_i^{-1}\hat{\mathbf{t}}_{oi}$$

where $\hat{\mathbf{C}}_i$ is a $p \times p$ symmetric matrix with elements given by $\hat{c}_{rr'i}$ in equation (6.4), and $\hat{\mathbf{t}}_{0i}$ is a $p \times 1$ vector of the elements $\hat{t}_{p0i}$ described in equation (6.5).

Taylor series linearisation amounts to finding a linear approximation

$$\hat{\mathbf{B}}_i \doteq \hat{\mathbf{B}}_{0i} = \mathbf{B}_i + \sum_{r=1}^{p}\sum_{r\leq r'} \mathbf{a}_{rr'i}(\hat{c}_{rr'i} - c_{rr'i}) + \sum_{r=1}^{p} \mathbf{a}_{r0i}(\hat{t}_{r0i} - t_{r0i}) \tag{6.7}$$

where $\mathbf{a}_{rr'i}$ and $\mathbf{a}_{r0i}$ are vectors defined as

$$\begin{aligned} \mathbf{a}_{rr'i} &= \left.\frac{\partial\hat{\mathbf{B}}_i}{\partial\hat{c}_{rr'i}}\right| \\ \mathbf{a}_{r0i} &= \left.\frac{\partial\hat{\mathbf{B}}_i}{\partial\hat{t}_{r0i}}\right| \end{aligned}$$

evaluated at the true values, $\hat{\mathbf{C}}_i = \mathbf{C}_i$ and $\hat{\mathbf{t}}_{0i}= \mathbf{t}_{0i}$. We obtain the following

$$\begin{aligned} \mathbf{a}_{rr'i} &= -(\mathbf{C}_i^{-1}\mathbf{\Delta}_{rr'i}\mathbf{C}_i^{-1})\mathbf{t}_{0i} \\ \mathbf{a}_{r0i} &= \mathbf{C}_i^{-1}\lambda_{ri} \end{aligned}$$

where $\mathbf{\Delta}_{rr'i}$ is a $p \times p$ matrix with ones in positions (r, r') and (r', r) and zeros elsewhere, and λ_{ri} is a $p \times 1$ vector with the rth component 1 and zeros elsewhere.

Evaluating the derivatives at ($\mathbf{C}_i$, $\mathbf{t}_{0i}$) and inserting into (6.7) gives

$$\begin{aligned} \hat{\mathbf{B}}_{0i} = {} & \mathbf{B}_i - \textstyle\sum_{r=1}^{p}\sum_{r\leq r'} \mathbf{C}_i^{-1}\mathbf{\Delta}_{rr'i}\mathbf{B}_i(\hat{c}_{rr'i} - c_{rr'i}) + \sum_{r=1}^{p} \mathbf{C}_i^{-1}\lambda_{ri}(\hat{t}_{r0i} - t_{r0i}) \\ = {} & \mathbf{B}_i + \mathbf{C}_i^{-1}\left(\hat{\mathbf{t}}_{0i} - \hat{\mathbf{C}}_i\mathbf{B}_i\right) \end{aligned}$$

the linearised form of $\hat{\mathbf{B}}_i$.

Substituting the approximation (6.6) above into (6.3), an approximation for $g(\hat{y}_i)$ can be found and hence an approximation for the bias term. It follows that

$$g(\hat{y}_i) \doteq \mathbf{x}_i'\mathbf{B}_i + \mathbf{x}_i'\mathbf{C}_i^{-1}(\hat{\mathbf{t}}_i - \hat{\mathbf{C}}_i\mathbf{B}_i).$$

Taking the design expectation we arrive at

$$E_\pi(g(\hat{y}_i)) \doteq \mathbf{x}_i'\mathbf{B}_i = \sum_{r=1}^{p} B_{ri}x_{ri}$$

and hence the approximate bias results.□

The bias term is zero when the estimator has a smoother matrix which is *total-preserving*. A predictor is represented by $\mathbf{1}^T\mathbf{S}\mathbf{y}$, for any linear smoother matrix $\mathbf{S}$, or $\mathbf{1}^T\mathbf{x}^T\mathbf{B}$. If the linear smoother matrix is total preserving then $\mathbf{1}^T\mathbf{S}\mathbf{y}$ must give the total exactly and hence so must $\mathbf{1}^T\mathbf{x}^T\mathbf{B}$, since they are equivalent. Therefore, the locally weighted GLM estimator is approximately unbiased under repeated sampling, provided it has a smoother matrix which is total-preserving.

It should be noted that to estimate the population total we need to invert the transformation applied to μ (the link function) and apply it to the right hand side of the equation above. There are various ways of doing this, some described in Carroll and Ruppert (1988), but we do not consider them further here.

6.3.3 A π-weighted local regression estimator

Approximate bias

The results obtained from Section 6.3 are used to find the approximate bias under repeated sampling of the locally weighted regression estimator.

The components of the linear model are the identity link $g(\mu_i) = \mu_i = \beta_0 + \beta_1 x_i$ and a constant variance σ^2. The approximate bias using the result from above is

$$E_\pi(\sum_{i=1}^{N} \hat{y}_i - y_i) \doteq \sum_{i=1}^{N}\sum_{r=1}^{p} B_{ri}x_{ri} - \sum_{i=1}^{N} y_i \qquad (6.8)$$

where $\mathbf{B}_i = (B_1, B_2)'$ and

$$\mathbf{B}_i = \begin{pmatrix} \sum_{j=1}^{N} w_{ij}/\pi_j & \sum_{j=1}^{N} w_{ij}x_j/\pi_j \\ \sum_{j=1}^{N} w_{ij}x_j/\pi_j & \sum_{j=1}^{N} w_{ij}x_j^2/\pi_j \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^{N} w_{ij}y_j/\pi_j \\ \sum_{j=1}^{N} w_{ij}y_jx_j/\pi_j \end{pmatrix}, \qquad (6.9)$$

where $w_{ij} = W_b(x_i, x_j)$.

Solving (6.9) above, and substituting into the approximate bias formula (6.8), it is shown that the bias is approximately zero, for the same reason described in Section 6.3.2 above.

6.3.4 A π-weighted local ratio estimator

Approximate bias

This estimator has received the most attention throughout this thesis. It is an example of a local GLM estimator with $\text{var}_\xi(\mu_i) = x_i\sigma^2$, and $g(\mu_i) = \mu_i = \beta x_i$, written as:

$$\hat{T}_{LWRE} = \sum_{i=1}^{N} x_i \frac{\sum_{j\in s} W_b(x_i, x_j)y_j/\pi_j}{\sum_{j\in s} W_b(x_i, x_j)x_j/\pi_j} = \sum_{i=1}^{N} x_i \frac{\hat{t}_{iy\pi}}{\hat{t}_{ix\pi}} = \sum_{i=1}^{N} x_i\hat{R}_i, \tag{6.10}$$

using similar notation to Särndal et al. (1992), Remark 5.6.1, but including an extra subscript in the variables to denote that they depend on the target x_i. Then

$$\begin{aligned} E_\pi(\hat{t}_{iy\pi}) &= E_\pi\left(\textstyle\sum_{j\in s} W_b(x_i, x_j)y_j/\pi_j\right) \\ &= \textstyle\sum_{j=1}^{N} W_b(x_i, x_j)y_j \\ &= t_{iy}, \text{ etc.} \end{aligned}$$

By applying the results from Section 6.3 to (6.10),

$$\sum_{i=1}^{N} x_i\hat{R}_i \doteq \sum_{i=1}^{N} x_i R_i + \sum_{i=1}^{N} \frac{x_i}{t_{ix}}(\hat{t}_{iy\pi} - R_i\hat{t}_{ix\pi}),$$

approximating to the first term only in the Taylor series expansion. Taking expectations of this under the design employed:

$$E_\pi(\hat{T}_{LWRE}) \doteq \sum_{i=1}^{N} x_i\frac{t_{iy}}{t_{ix}} + \sum_{i=1}^{N} \frac{x_i}{t_{ix}}(t_{iy} - \frac{t_{iy}}{t_{ix}}t_{ix}) = \sum_{i=1}^{N} x_i\frac{t_{iy}}{t_{ix}},$$

and therefore

$$E_\pi(\hat{T}_{LWRE} - T) \doteq \sum_{i=1}^{N} x_i\frac{t_{iy}}{t_{ix}} - T.$$

Hence, an expression for the approximate bias of the locally π-weighted ratio estimator under any design has been derived. The estimator, $\hat{T}_{LWRE}$, is approximately unbiased when

$$\sum_{i=1}^{N} x_i\frac{t_{iy}}{t_{ix}} = \sum_{i=1}^{N} y_i,$$

i.e. when the estimator has a smoother matrix that is total-preserving (see Section 4.3).

Error associated with the approximate bias

The approximation above is only rough, and does not work particularly well in samples that are small relative to the population size. In order to make the approximation more accurate, the next term in the Taylor series linearisation is included. This gives the order of magnitude of the error associated with the approximate bias.

The first term in the Taylor series expansion for the bias has already been given. The next, quadratic term, gives us some indication of the size of error associated with the approximation.

Recall that

$$\hat{R}_i = \frac{\hat{t}_{iy\pi}}{\hat{t}_{ix\pi}} = f(\hat{t}_{iy\pi}, \hat{t}_{ix\pi}), \tag{6.11}$$

and Särndal et al. (1992), in their derivations, neglected all terms after the first. In order to find the next term in the Taylor series expansion we must consider the second derivatives of the function (6.11), evaluated at the true values, t_{ix}, t_{iy}. The following coefficients are obtained from the second order derivatives:

$$a_1 = 0, \quad a_2 = \frac{2t_{iy}}{t_{ix}^3}, \quad a_3 = \frac{-1}{t_{ix}^2},$$

for the $\hat{t}_{iy\pi}^2/2!$, $\hat{t}_{ix\pi}^2/2!$ and $\hat{t}_{iy\pi}\hat{t}_{ix\pi}$ terms in the expansion respectively. Then

$$\hat{R}_i \doteq R_i + \frac{1}{t_{ix}}(\hat{t}_{iy\pi} - R_i\hat{t}_{ix\pi}) + \frac{1}{2!t_{ix}^2}(2R_i\hat{t}_{ix\pi}^2 - 2\hat{t}_{iy\pi}\hat{t}_{ix\pi}).$$

Taking design-expectations leads to

$$E_\pi(\hat{T}_{LWRE}) = E_\pi(\sum_{i=1}^{N} x_i R_i) + \sum_{i=1}^{N} \frac{x_i}{2!t_{ix}^2} E_\pi\left(2R_i\hat{t}_{ix\pi}^2 - 2\hat{t}_{ix\pi}\hat{t}_{iy\pi}\right).$$

The first term is just the $O(1)$ term that is fixed for any sample size and, for the total-preserving estimators, equals the population total. The second term above is the error term we are interested in and consider in more detail. Now

$$\begin{aligned} E_\pi(\hat{t}_{ix\pi}^2) &= \text{var}_\pi(\hat{t}_{ix\pi}) + \left(E_\pi(\hat{t}_{ix\pi})\right)^2 \\ &= \sum\sum\nolimits_U \tfrac{\pi_{kl}}{\pi_k\pi_l} W_b(x_i, x_k) W_b(x_i, x_l) x_l x_k \end{aligned}$$

and

$$\begin{aligned} E_\pi(\hat{t}_{ix\pi}\hat{t}_{iy\pi}) &= \mathrm{cov}_\pi(\hat{t}_{iy\pi}, \hat{t}_{ix\pi}) + E_\pi(\hat{t}_{iy\pi})E_\pi(\hat{t}_{ix\pi}) \\ &= \sum\sum_U \tfrac{\pi_{kl}}{\pi_k\pi_l} W_b(x_i, x_k)W_b(x_i, x_l)x_k y_l. \end{aligned}$$

For example, under stratified simple random sampling

$$E_\pi(\hat{t}^2_{ix\pi}) = \left(\sum_{i=1}^{N} W_b(x_i, x_j)x_j\right)^2 + \sum_{h=1}^{H} \frac{(N_h - n_h)N_h}{n_h} S^2_{hix}$$

and

$$E_\pi(\hat{t}_{ix\pi}, \hat{t}_{iy\pi}) = \left(\sum_{j=1}^{N} W_b(x_i, x_j)x_j\right)\left(\sum_{j=1}^{N} W_b(x_i, x_j)y_j\right) + \sum_{h=1}^{H} \frac{(N_h - n_h)N_h}{n_h} S_{hixy},$$

where the S^2_{hix} and S_{hixy} are the individual stratum population variances and covariances, weighted according to the target x_i.

Hence, the error term for the stratified simple random sample case is

$$\sum_{i=1}^{N} \frac{x_i}{2!t^2_{ix}} E_\pi\left(2R_i\hat{t}^2_{ix\pi} - 2\hat{t}_{ix\pi}\hat{t}_{iy\pi}\right)$$

$$= \sum_{i=1}^{N} \frac{x_i}{t^2_{ix}} \left[R_i \left(\sum_{i=1}^{N} W_b(x_i, x_j)x_j\right)^2 + R_i \sum_{h=1}^{H} \frac{(N_h - n_h)N_h}{n_h} S^2_{hix} \right.$$

$$\left. - \left(\sum_{j=1}^{N} W_b(x_i, x_j)x_j\right)\left(\sum_{j=1}^{N} W_b(x_i, x_j)y_j\right) - \sum_{h=1}^{H} \frac{(N_h - n_h)N_h}{n_h} S_{hixy}. \right]$$

The first and third term above are fixed, for a suitable b, and the remaining error term is of order

$$\frac{N_h(N_h - n_h)}{n_h(N_h - 1)} \doteq \left(\frac{N_h}{n_h} - 1\right).$$

This indicates that the size of the error term depends on n_h/N_h, the strata sampling fraction. If n_h is small relative to N_h then the error term on the approximate bias is large, otherwise if n_h/N_h is close to 1, i.e. $n_h \to N_h$, the approximation does well.

Approximate variance under repeated sampling

Again, the methods described below are based on the proofs given in Särndal et al. (1992) for finding the approximate variance of the π-weighted ratio estimator under repeated sampling. The π_k, π_{kl} are the first and second order inclusion probabilities and $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

Theorem 6.2 *The approximate variance of the π-weighted local ratio estimator is*

$$AV(\hat{T}_{LWRE}) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{x_i x_{i'}}{t_{ix} t_{i'x}} \left[cov_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'y\pi}) - R_i \, cov_\pi(\hat{t}_{ix\pi}, \hat{t}_{i'y\pi}) \right. \qquad (6.12)$$
$$\left. - R_{i'} \, cov_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'x\pi}) + R_i R_{i'} \, cov_\pi(\hat{t}_{ix\pi}, \hat{t}_{i'x\pi}) \right]$$

Proof of theorem

To find the approximate variance of the π-weighted local ratio estimator, $\hat{T}_{LWRE}$, we note that the ratios R_i, appearing in the estimator, depend upon the i. Recall that

$$\hat{T}_{LWRE} = \sum_{i=1}^{N} x_i \hat{R}_i,$$

where

$$\hat{R}_i = \frac{\hat{t}_{iy\pi}}{\hat{t}_{ix\pi}}.$$

The approximate variance is

$$\begin{aligned} AV(\hat{T}_{LWRE}) &= AV(\sum_{i=1}^{N} x_i \hat{R}_i) \\ &= \sum_{i=1}^{N} AV(x_i \hat{R}_i) + \sum \sum_{i \neq i'} ACOV(x_i \hat{R}_i, x_{i'} \hat{R}_{i'}). \qquad (6.13) \end{aligned}$$

The first term on the right hand side can be written as

$$\begin{aligned} \sum_{i=1}^{N} AV(x_i \hat{R}_i) &= \sum_{i=1}^{N} x_i^2 AV(\hat{R}_i) \\ &= \sum_{i=1}^{N} \frac{x_i^2}{t_{ix}^2} \sum \sum_{U} \Delta_{kl} \left(\frac{y_k - R_i x_k}{\pi_k} \right) \left(\frac{y_l - R_i x_l}{\pi_l} \right) \\ &= \sum_{i=1}^{N} \frac{x_i^2}{t_{ix}^2} \left(\text{var}(\hat{t}_{iy\pi}) + R_i^2 \, \text{var}(\hat{t}_{ix\pi}) - 2R_i \, \text{cov}(\hat{t}_{iy\pi}, \hat{t}_{ix\pi}) \right) (6.14) \end{aligned}$$

The individual $\mathrm{var}_\pi(\hat{t}_{iy\pi})$, etc. are calculated for the particular design considered; we might, for example, be interested in the approximate variance of this estimator under stratified simple random sampling, in which case we would consider the individual variances and covariances of the $\hat{t}_{iy\pi}$ and $\hat{t}_{ix\pi}$ under stratified simple random sampling. This is considered some more later, but next we turn to the second term in equation (6.13) above. Now,

$$\mathrm{cov}_\pi(x_i\hat{R}_i, x_{i'}\hat{R}_{i'}) = x_i x_{i'}\,\mathrm{cov}_\pi(\hat{R}_i, \hat{R}_{i'})$$

where

$$\mathrm{cov}_\pi(\hat{R}_i, \hat{R}_{i'}) = E_\pi(\hat{R}_i\hat{R}_{i'}) - E_\pi(\hat{R}_i)E_\pi(\hat{R}_{i'})$$

and

$$E_\pi(\hat{R}_i\hat{R}_{i'}) = E_\pi\left(\frac{t_{iy\pi}t_{i'y\pi}}{t_{ix\pi}t_{i'x\pi}}\right). \tag{6.15}$$

Using the Taylor series approximation to the first term only, we have

$$\hat{R}_i \doteq R_i + \frac{1}{t_{ix}}\left(\hat{t}_{iy\pi} - R_i\hat{t}_{ix\pi}\right).$$

Then (6.15) above becomes

$$\begin{aligned}
&E_\pi\left(R_i + \frac{1}{t_{ix}}(\hat{t}_{iy\pi} - R_i\hat{t}_{ix\pi})\right)\left(R_{i'} + \frac{1}{t_{i'x}}(\hat{t}_{i'y\pi} - R_{i'}\hat{t}_{i'y\pi})\right)\\
= \;& E_\pi\left[R_iR_{i'} + \frac{R_{i'}}{t_{ix}}(\hat{t}_{iy\pi} - R_i\hat{t}_{ix\pi}) + \frac{R_i}{t_{i'x}}(\hat{t}_{i'y\pi} - R_i\hat{t}_{i'x\pi})\right.\\
&\left.+\frac{(\hat{t}_{iy\pi} - R_i\hat{t}_{ix\pi})(\hat{t}_{i'y\pi} - R_{i'}\hat{t}_{i'x\pi})}{t_{ix}t_{i'x}}\right]\\
= \;& R_iR_{i'} + E_\pi\left(\frac{(\hat{t}_{iy\pi} - R_i\hat{t}_{ix\pi})(\hat{t}_{i'y\pi} - R_{i'}\hat{t}_{i'x\pi})}{t_{ix}t_{i'x}}\right).
\end{aligned}$$

So the covariance term is

$$\begin{aligned}
\mathrm{cov}_\pi(\hat{R}_i, \hat{R}_{i'}) &= E_\pi\left(\frac{(\hat{t}_{iy\pi} - R_i\hat{t}_{ix\pi})(\hat{t}_{i'y\pi} - R_{i'}\hat{t}_{i'x\pi})}{t_{ix}t_{i'x}}\right)\\
&= \frac{1}{t_{ix}t_{i'x}}\left[E_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'y\pi}) - R_iE_\pi(\hat{t}_{ix\pi}, \hat{t}_{i'y\pi}) - R_{i'}E_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'x\pi})\right.\\
&\quad\left.+R_iR_{i'}E_\pi(\hat{t}_{ix\pi}, \hat{t}_{i'x\pi})\right]\\
&= \frac{1}{t_{ix}t_{i'x}}\left[\mathrm{cov}_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'y\pi}) - R_i\,\mathrm{cov}_\pi(\hat{t}_{ix\pi}, \hat{t}_{i'y\pi})\right.\\
&\quad\left.-R_{i'}\,\mathrm{cov}_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'x\pi}) + R_iR_{i'}\,\mathrm{cov}_\pi(\hat{t}_{ix\pi}, \hat{t}_{i'x\pi})\right]
\end{aligned} \tag{6.16}$$

Combining equations (6.14) and (6.16) gives the approximate variance. □

The approximate variance can be calculated for any design using the population quantities appearing in the formula. It can also be estimated by estimating any population quantity appearing in the formulae by the corresponding sample quantity.

For illustration, the approximate variance of the π-weighted local ratio estimator under SSRS has been considered. In the formula for the approximate variance, design-based quantities such as $\text{var}_\pi(\hat{t}_{iy\pi})$ and $\text{cov}_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'y\pi})$, etc., are required.

In the notation of Särndal et al. (1992) we write

$$\text{var}_\pi(\hat{t}_{iy\pi}) = \sum\sum_{U} \Delta_{kl} W_b(x_i, x_l) W_b(x_i, x_k) \frac{y_k y_l}{\pi_k \pi_l}.$$

The above can also be written as

$$\text{var}_\pi(\hat{t}_{iy\pi}) = \sum\sum_{U} \left(\frac{\pi_{kl}}{\pi_k \pi_l} W_b(x_i, x_k) W_b(x_i, x_l) y_k y_l \right) - \left(\sum_{U} y_k W_b(x_i, x_k) \right)^2 \quad (6.17)$$

Here, the particular design is introduced, by substituting the π_{kl}, $k \neq l, k$, $l = 1, \ldots, N$ and π_k, $k = 1, \ldots N$ inclusion probabilities into (6.17) above. Using the SSRS inclusion probabilites, then

$$\text{var}_\pi(\hat{t}_{iy\pi}) = \sum_{h=1}^{H} \frac{N_h(N_h - n_h)}{n_h} S^2_{hyi},$$

where

$$S^2_{hyi} = \frac{\sum_{k=1}^{N_h} W_b(x_i, x_{hk})^2 y^2_{hk} - (\sum_{k=1}^{N_h} W_b(x_i, x_{hk}) y_{hk})^2 / N_h}{(N_h - 1)}.$$

This is the locally weighted variance of the y's weighted according to the target x_i, calculated within each stratum. Similar expressions can be derived for the other terms appearing in the approximate variance formula, all of which are summarised below. The $\check{y}_k, \check{x}_k$ are the π-weighted y and x observations in the sample respectively. The covariance terms appearing in the approximate variance formula can be written as:

$$\begin{aligned}
\text{cov}_\pi(\hat{t}_{iq\pi}, \hat{t}_{i'r\pi}) &= \sum\sum\nolimits_U \Delta_{kl} W_b(x_i, x_k) W_b(x_{i'}, x_l) \check{q}_k \check{r}_l \\
&= \sum\nolimits_{h=1}^{H} \tfrac{N_h(N_h - n_h)}{n_h} \Big(\sum\nolimits_{k=1}^{N_h} W_b(x_i, x_{hk}) W_b(x_{i'}, x_{hk}) q_{hk} r_{hk} \\
&\quad - \tfrac{(\sum_1^{N_h} W_b(x_i, x_{hk}) q_{hk})(\sum_1^{N_h} W_b(x_{i'}, x_{hk}) r_{hk})}{N_h} \Big) / (N_h - 1),
\end{aligned}$$

with appropriate substitution of q and r to obtain $\text{cov}_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'y\pi})$, $\text{cov}_\pi(\hat{t}_{ix\pi}, \hat{t}_{i'x\pi})$, $\text{cov}_\pi(\hat{t}_{ix\pi}, \hat{t}_{i'y\pi})$ and $\text{cov}_\pi(\hat{t}_{iy\pi}, \hat{t}_{i'x\pi})$.

The approximate variance of the simplest case, the combined ratio estimator (when all $W_b(x_i, x_j) = 1$), could be computed quite easily if the stratum variances and covariances of the population x and y were known. However, these will be unknown, for the y's at least, and so estimates can be obtained from the within stratum sample x and y variances and covariances. The same estimation procedure can be applied to the approximate variance formula, (6.12) above, substituting the unknown population variances and covariances by their respective sample quantities.

As $n_h \to N_h$, the variance approximation improves. When $N_h = n_h$, the variance is zero. There is an error term associated with the approximate variance which depends on n_h and N_h in some way. The calculation for this has been omitted here.

The result for the approximate variance given for the locally weighted ratio estimator can be extended to a locally weighted regression estimator, and generalised to locally weighted generalised linear models.

6.4 Generalised linear modeling with regression splines

To ensure design unbiasedness in our generalised linear model estimator, as described in Section 4.3.3, a π-weight, associated with each sample value, is introduced.

Recall that the model previously used was:

$$g(\mu_i) = g(x_i) + \beta_0 + \sum_{j=1}^{k} \alpha_j A_j(x_i), \qquad (6.18)$$

where $A_j(x_i)$ are the linear basis functions for a regression spline (B-spline). The above model with canonical link and appropriate variance function leads to a total-preserving estimator. In the case of Normal errors, or Poisson errors with an offset $\log(x)$ and intercept term in the linear predictor, the solution is non-iterative.

Introducing the weights, $1/\pi_j$, into (6.18) above, makes no difference to the total-preserving property, this still remains. The estimating equation is

$$\sum_{i=1}^{n}\left(\frac{y_i}{\pi_i}-\frac{\hat{\mu}_i}{\pi_i}\right)=0,$$

i.e.

$$\sum_{i=1}^{n}\frac{y_i}{\pi_i}=\sum_{i=1}^{n}\frac{\hat{\mu}_i}{\pi_i}.$$

The model described above, with and without prior weights of $1/\pi_j$, is compared numerically in Section 6.5 with the parametric ratio estimators and the other nonparametric regression estimators, using the India dataset of Section 1.4. The calculations for the generalised linear model estimator were performed in S-PLUS (see Becker et al., 1988 and Chambers and Hastie, 1990 for more on modelling using S-PLUS). The **glm** function was used; prior weights were included using the **weight=** option.

The following remarks can be made about this estimator:

1. It is fairly easily implemented; the π weights can be introduced as prior weights into the generalised linear model if required.

2. It has the feature of a smooth underlying function which is often more realistic of the underlying model than the discontinous underlying model of the separate ratio estimator.

3. The variance under repeated sampling is considerably smaller than some of the parametric estimators, and the bias is comparable. Introducing the π-weight improves the estimator to a certain extent. (See numerical results in Section 6.5).

4. There is no problem with deciding how many knots and where to place them for the regression spline part, if stratified random sampling is used; the number of strata and stratum boundaries should be used. This is discussed in Section 4.3.3 in more detail.

6.5 Numerical results

In this section we establish, empirically, how well the π-weighted estimators perform in terms of bias and variance, under repeated sampling. The empirical performance is also compared with the approximations derived from above. The approximations worsen as the sample size decreases relative to the population size, as expected.

Monte Carlo simulation was performed by selecting 1000 samples and calculating the bias and variance of the 1000 samples.

Example 6.1 *India dataset*

The India dataset of Section 1.4, is used. From the population, 1000 SSRSs with 3 strata were selected using optimal stratification. The bias and variance of the ratio estimator, separate ratio estimator, combined ratio estimator and various π-weighted nonparametric estimators under SSRS were calculated and are found in Tables 6.1 , 6.2, and 6.3. The experiment was performed for varying sample sizes, $n = 30$ and $n = 60$, and for a range of values of the smoothing parameter.

The theoretical approximate bias and variance calculations for this dataset can be found in Table 6.1. The combined and separate ratio estimators are approximately unbiased as these are π-weighted versions of the ratio estimator. Approximate unbiasedness also applies to the π-weighted averaged running ratio and GLM with spline estimators. However, the π-weighted local ratio estimators with kernel or nearest-neighbour weights are not approximately unbiased under repeated sampling. The first term in the approximate bias is $O(1)$ for these estimators and so is fixed for any sample size n. This is because these estimators do not possess the total-preserving property of the other estimators. In order to be total-preserving the following condition should be satisfied:

$$\mathbf{1}^T\mathbf{S}\mathbf{y} = \mathbf{1}^T\mathbf{y},$$

where $\mathbf{S}$ is the smoother matrix associated with the estimator. This is not the case for these estimators and is discussed more in Section 4.3.1. In Tables 6.2 and 6.3 the empirical results from selecting 1000 SSRSs are given. The results based on $n = 60$ give smaller variances and therefore more precise estimates of bias than for the $n = 30$ case, as expected. The bias and variance of the π-weighted running ratio estimator and Gaussian weighted estimators appear to increase and then decrease again for varying smoothing parameter. The estimators behave as might be expected, the ratio estimator performing the worst. The locally weighted Gaussian kernel and running line estimators improve on the ratio estimator, in some instances, but are only unbiased as b (or k) approaches infinity, when they become the combined ratio estimator. The combined and separate ratio estimator perform well in terms of bias but are outperformed further by the averaged running ratio and generalised linear model estimators. Both are unbiased (or approximately so) under repeated sampling and both are more precise and efficient than either the combined or separate ratio estimator. For the $n = 60$ case the empirical MSE is reduced from 0.435 in the separate ratio estimator to 0.412 and 0.317 in the ARRE and GLM with spline estimator respectively. The generalised linear model with spline estimator has a smaller variance than any of the other estimators, particularly in the $n = 60$ case.

The nonparametric regression estimators appear more efficient than the unbiased parametric estimators. The variance term dominates the mean squared error in this example; this is not, however, always the case. These results are very encouraging and show the potential of nonparametric regression in the prediction of finite population measures.

6.6 Design-model based approach

6.6.1 Introduction

In Section 2.2.4, the idea of joint expectation (and variance) under the design and superpopulation model was introduced. Here, properties of the estimators under these joint methods are considered further, motivated by the work of Royall and Herson (1973a) on ratio estimators and separate ratio estimators. They consider simple random sampling compared with balanced sampling, and in particular the superpopulation variance under both designs. They conclude that the variance under balanced sampling is always smaller than the variance under SRS for the ratio and separate ratio estimators. Interesting joint properties to consider include

$$E_{\pi}E_{\xi}(\hat{T} - T) \quad \text{and} \quad E_{\pi}(\text{var}_{\xi}(\hat{T})), \tag{6.19}$$

	Approximate		
Estimator	Bias	Variance $n = 60$	Variance $n = 30$
Ratio Estimator (n=60)	-0.3622	0.3870	5.4437
π-weighted ratio estimators			
Combined R.E.	0.0000	0.4452	5.6602
Separate R.E.	0.0000	0.4403	5.7004
Locally weighted R.Es			
with π weighting			
Gaussian kernel			
b=200	0.1135	0.4083	4.5435
b=400	-0.2321	0.5215	5.7438
b=600	-0.3681	0.5346	5.9239
Running line			
k=10	-0.1946	0.3591	4.7071
k=20	-0.2384	0.3875	4.9922
k=30	-0.2609	0.3885	5.1065
Averaged running R.E.			
k=10	0.0000	0.3735	4.7420
k=20	0.0000	0.4121	5.3670
k=30	0.0000	0.4378	5.7537

Table 6.1: Theoretical approximation to design bias and variance under SSRS for the India dataset

Estimator	Actual bias	Actual Variance	Actual MSE
Ratio estimator	-0.3891	0.3807	0.5321
π-weighted ratio estimator			
Combined R.E.	-0.0309	0.4387	0.4397
Separate R.E	-0.0256	0.4341	0.4348
Locally weighted R.Es			
with π weighting			
Gaussian kernel			
b=200	0.0937	0.4065	0.4153
b=400	-0.2619	0.5129	0.5815
b=600	-0.3998	0.5258	0.6856
b=800($\approx$ 3 df)	-0.5061	0.5157	0.7718
b=10000	-0.0309	0.4387	0.4397
Running line			
k=1	0.0904	0.7018	0.7099
k=10	-0.1955	0.4089	0.4471
k=20($\approx$ 3 df)	-0.2804	0.4105	0.4891
k=30	-0.2161	0.3896	0.4363
k=500	-0.0309	0.4387	0.4397
Averaged running R.E			
k=1	0.0904	0.7018	0.7099
k=10	-8.0e^{-5}	0.4146	0.4146
k=20	-0.0168	0.4114	0.4117
k=30($\approx$ 3df)	-0.0229	0.4167	0.4172
GLM with spline			
with π weight	-0.0575	0.3134	0.3167
without π weight	0.1672	0.3012	0.3292

Table 6.2: Empirical results selecting 1000 stratified SRSs with n=60, India dataset

Estimator	Actual bias	Actual variance	Actual MSE
Ratio estimator	-0.4305	5.4282	5.6135
Combined R.E.	0.0212	5.6424	5.6428
Separate R.E.	0.0191	5.6668	5.6672
Locally weighted R.Es			
with π weighting			
Gaussian Kernel			
b=200	-0.4193	5.4736	5.6494
b=400	-0.6646	6.3283	6.7700
b=600($\approx$ 3df)	-0.6826	6.3669	6.8328
b=800	-0.6827	6.1734	6.6395
b=1000	-0.6910	5.9831	6.4606
b=2000	-0.7005	5.7170	6.2077
b=10000	-0.0155	5.6257	5.6259
Running line			
k=1	-0.2928	7.6158	7.7015
k=10($\approx$ 3 df)	-0.2690	5.5131	5.5855
k=20	-0.4072	5.7849	5.9507
k=26	-0.7981	6.9365	7.5735
k=30	-1.0212	8.3133	9.3561
k=32	-0.9706	8.2032	9.1453
k=36	-0.7473	7.3754	7.9339
k=40	-0.6429	7.0293	7.4426
k=500	0.0212	5.6424	5.6428
Averaged running R.E.			
k=1	-0.2928	7.6158	7.7015
k=10	-0.1004	5.5141	5.5242
k=15($\approx$3 df)	-0.0722	5.5313	5.5365
k=20	-0.0552	5.6126	5.6156
k=30	-0.0313	5.6365	5.6375
k=35	-0.0238	5.6353	5.6358
k=40	-0.0182	5.6348	5.6351
k=50	-0.0103	5.6348	5.6349
GLM with spline			
with π weight	-0.4301	6.3102	6.4952
without π weight	-0.1588	5.1397	5.1649

Table 6.3: Empirical results selecting 1000 stratified SRSs with n=30, India dataset

also the variances associated with these expected values, i.e.

$$\text{var}_\pi\, E_\xi(\hat{T} - T) \quad \text{and} \quad \text{var}_\pi(\text{var}_\xi(\hat{T})). \tag{6.20}$$

The theoretical results for the first expression of (6.19) above can be easily derived from those for $E_\pi(\hat{T} - T)$, by replacing every occurence of y_i in the formula by the superpopulation mean, $m(x_i)$. If the estimator is unbiased under either the design or model, then there is overall unbiasedness.

Recall that

$$\text{var}_\xi(\hat{T}_{\text{ARRE}}) = \sigma^2 \sum_{j\in s} x_j \left(\sum_{i=1}^{N} \frac{x_i}{k} \sum_{m=1}^{k} \frac{I_{ijm}}{\sum_{l\in s} I_{ilm} x_l} \right)^2,$$

for our averaged running ratio estimator. Then, after some manipulation, it can be shown that

$$E_\pi\left(\text{var}_\xi(\hat{T}_{\text{ARRE}})\right) \doteq \frac{N}{n} \sum_{i=1}^{N} x_i \left(\sum_{l=1}^{N} \frac{x_l}{k} \sum_{m=1}^{k} \frac{I_{ijm}}{\sum_{l=1}^{N} I_{ilm} x_l} \right)^2,$$

expectation taken with respect to SRS. Similar results can be derived for the quantities given in (6.20) above but have been omitted here.

6.6.2 An example

The India dataset of Section 1.4 has been used in this example. The ratio estimator, separate ratio estimator and averaged running ratio estimator are used to predict the population total. In Table 6.4 the results from selecting 1000 stratified (3 strata) simple random samples and computing the joint model-design properties, $E_\pi E_\xi(\hat{T} - T)$ and $E_\pi\left(\text{var}_\xi(\hat{T})\right)$, are given. Varying sample sizes and spans are used. The standard errors associated with these expectations are given in brackets in the table. The averaged running ratio estimator performs best in terms of reducing the standard error associated with the joint expectation. The ratio estimator performs best in expected variance under repeated sampling, however this is not the case for the expected bias under repeated sampling.

Sample size	Estimator(Span)	$E_\pi(\mathrm{var}_\xi(\hat{T}))$		$E_\pi(E_\xi(\hat{T}-T))$	
10	$\hat{T}_{ARRE}$ k=5	22197	(48)	5.4	(1.51)
	$\hat{T}_{RE}$	20963	(1892)	66.6	(93.67)
	$\hat{T}_{SRE}$	23183	(2194)	-9.1	(69.74)
15	$\hat{T}_{ARRE}$ k=7	14939	(26)	3.9	(1.05)
	$\hat{T}_{RE}$	14385	(956)	5.2	(68.60)
	$\hat{T}_{SRE}$	153282	(1047)	-6.2	(52.58)
20	$\hat{T}_{ARRE}$k=10	10874	(33)	11.7	(0.55)
	$\hat{T}_{RE}$	10436	(400)	71.6	(118.65)
	$\hat{T}_{SRE}$	11371	(518)	-2.5	(28.67)
27	$\hat{T}_{ARRE}$ k=13	8326	(3)	8.5	(0.03)
	$\hat{T}_{RE}$	8185	(86)	46.9	(3.70)
	$\hat{T}_{SRE}$	8507	(177)	-0.3	(3.85)
34	$\hat{T}_{ARRE}$ k=17	7426	(0)	0	(0)
	$\hat{T}_{RE}$	7426	(0)	0	(0)
	$\hat{T}_{SRE}$	7426	(0)	0	(0)

Table 6.4: Expected $\mathrm{var}_\xi(\hat{T})$ and bias under stratified simple random sampling

Chapter 7

Conclusions and further research

7.1 Conclusions

In this thesis, the problem of predicting finite population totals using nonparametric regression estimators has been addressed. The nonparametric estimators described lead to greater efficiency than standard estimators because of their ability to reflect the actual structure of the data. In particular, comparisons with the ratio estimator and separate ratio estimator were most beneficial; it was shown that nonparametric regression estimators have genuine gains in efficiency over the ratio estimator and separate ratio estimator. This was emphasised by numerical results in Chapter 5 and 6. Efficiency gains, over the separate ratio estimator, were shown for all of the nonparametric regression estimators described. The fact that nonparametric regression estimators perform better than standard estimators suggests that they are methods that should be considered in any prediction of finite population measures.

A large class of nonparametric regression estimators were described in Chapter 4, dichotomised into operational and model-based estimators. Motivation for these nonparametric regression estimators was also given. Of particular interest are those which possess a 'total-preserving' property, found to be of greatest benefit in removing design-based bias while keeping the design-variance small. The total-preserving nonparametric estimators were more efficient, for a suitable choice of the smoothing parameter, than standard total-preserving estimators such as the separate and combined ratio estimators, because of the smooth underlying curve of the nonparametric estimators as opposed to the underlying discontinuous model of the parametric separate and combined ratio estimators. This was particularly true of the averaged running ratio estimator and the model-based, generalised linear model (GLM) with spline estimator. In particular, the GLM with spline estimator, for the example given in Chapter 5, outperformed most of the other estimators, in reducing the design-variance. There is much scope for considering other generalised linear model estimators such as this, incorporating a nonparametric component into a parametric model thereby providing a *semi-parametric* approach. This semi-parametric approach could be extended to many different types of data. Also in Chapter 4 an alternative form for the degrees of freedom was derived which allowed for the possibility of a heterogenous variance.

In Chapter 6, the estimators described were π-weighted to ensure approximate design unbiasedness and properties of these π-weighted estimators under repeated sampling were derived. In particular, the approximate design-based bias of a π-weighted local generalised linear model estimator was obtained and shown to be zero when the estimator was total-preserving. The approximate variance of the π-weighted local ratio estimator was also derived. For the generalised linear model with regression spline estimator π weights were introduced as prior weights into the model. The total-preserving property still holds in this case. An example based on the India dataset was included to illustrate the advantage of total-preserving nonparametric regression estimators over parametric estimators.

Finally, the question of what value to choose for the smoothing parameter was covered in Chapter 5. Crossvalidation, in the context of predicting a population total, performed disappointingly; a modified crossvalidation criterion did not improve on this any further. Work on the asymptotic bias, variance and mean squared error of the locally weighted ratio estimator was enlightening. It was shown how the bias depends on functionals of the design density, the underlying superpopulation mean and their derivatives. The asymptotic variance has a first term which does not depend on b, the smoothing parameter. We did not derive an asymptotically optimal bandwidth from this, but an intuitive 'local' bandwidth was considered based on some aspects of the asymptotic variance. The most practical choice of smoothing parameter was using the degrees of freedom formula, as described by Tibshirani and Hastie (1987) and the alternative form derived in Section 4.5, from the modified residual sums of squares. There appears to be much scope for further research into this very important question of smoothing parameter selection with a growing interest in the literature for new or modified approaches.

7.2 Recommendations for further research

The following are some possible areas of further research related to the work carried out in this thesis:

(1) The important issue of smoothing parameter selection requires much further investigation. In particular, more work on the choice of a local or variable bandwidth in the kernel estimators is required. It would also be useful to provide an automatic method of selection, perhaps by considering an approach similar to penalising functions as described in Section 5.2.2.

(2) The nonparametric estimators described have not been exhaustive. The class of operational estimators referred to as averaged running estimators can be extended to include any estimator (in particular total-preserving) as the running estimator. These may include, for example, running generalised linear model estimators. Semi-parametric estimators, such as the GLM with spline estimator, are well worth investigating further. Provided an intercept term and canonical link are used in the generalised linear model part of the model the total-preservation property holds. The smooth part of the model (as in the regression spline) could be replaced by any smooth function, such as a smoothing spline. This semi-parametric approach leads to a flexible and also fairly efficient class of models.

(3) Variance estimation has not been covered extensively in this thesis but is an important area that also needs addressing. It would be of interest to find methods of estimating the function $\sigma(x_i)$ in order to yield a variance estimator for the prediction error of the total.

(4) Robustification to outliers by downweighting influential observations as described by Chambers (1993), is worth consideration. Also no allowance for missing population x values has been given; the examples we have looked at have contained the x value for the whole population.

(5) Extensions to other types of data. An obvious extension is to binary data. For example, suppose $y_i = 1$ if an event occurs, and zero otherwise and x_i is some measure of size. Interest may be in a population total such as $T = \sum_{i=1}^{N} x_i y_i$.

Here, since $\pi_i = Pr(Y_i = 1)$ typically depends on x_i, a linear logistic model such as

$$\log \frac{\pi_i}{1-\pi_i} = \sum_{j=0}^{J} \beta_j x_i^j,$$

might be appropriate, for some Jth order polynomial ($J \geq 1$) to ensure total-preservation. This has been covered to some extent in Section 6.3 with the introduction to the local generalised linear model estimator. In the case of multivariate data the generalised additive model already described in Section 3.4, is analogous to nonparametric regression in the univariate setting. More recently Hastie and Tibshirani (1993) have described a general class of varying coefficient models. These models are linear in the regressors, but their coefficients are allowed to change smoothly with the value of other variables. Generalised additive models and nonparametric regression are examples of varying coefficient models.

(6) A related problem to finite population prediction is that of nonparametric regression estimation of finite population distribution functions as studied recently by Dorfman and Hall (1993), Kuk (1993) and Chambers, Dorfman and Wehrly (1993). In particular the introduction of bias calibration has been considered to take account of bias incurred by model misspecification. Bias calibration ideas could also be applied to nonparametric regression estimation of finite population totals.

Appendix A

Approximation to model mean squared error

Below, the coefficients a_1, a_2 and a_3, of the large-b (or small-k) approximation to the model mean squared error $(a_1 + a_2 k + a_3 k^2)$ are given, for the locally weighted ratio estimator with a Gaussian kernel weight function. These are based on using the Taylor series expansion of the Gaussian kernel function and the subsequent expansion of the locally weighted ratio estimator.

The model-MSE or PMSE can be written, up to terms of $O(k^2)$, as:

$$a_1 + a_2 k + a_3 k^2,$$

where the coefficients a_1, a_2 are obtained from contributions of the model-variance and squared bias terms. These coefficients are:

$$\begin{aligned}
a_1 &= \sigma^2 \left[\frac{\sum_{i=1}^N x_i}{\sum_{j=1}^n x_j} \left(\sum_{i=1}^N x_i - \sum_{j=1}^n x_j \right) \right] + (\sum_{i=1}^N x_i)^2 \left(\frac{\sum_{j=1}^n m(x_j)}{\sum_{j=1}^n x_j} \right)^2, \\
a_2 &= -2 \sum_{i=1}^N x_i \left[\frac{\sum_{j=1}^n m(x_j)}{\sum_{j=1}^n x_j} - \frac{\sum_{i=1}^N m(x_i)}{\sum_{i=1}^N x_i} \right] \times \\
&\quad \left[\frac{\sum_{i=1}^N x_i}{2(\sum_{j=1}^n x - j)^2} \left(\sum_{j=1}^n (x_i - x_j)^2 \right) \left(m(x_j) \sum_{k \in s} x_k - x_j \sum_{k \in s} m(x_k) \right) \right],
\end{aligned}$$

and

$$a_3 = \frac{\sum_{j=1}^n \sigma^2(x_j)}{4} \left[\frac{\sum_{i=1}^N x_i}{\sum_{j=1}^n x_j} \left(\frac{\sum_{j=1}^n x_j (x_i - x_j)^2}{\sum_j x_j} - (x_i - x_j)^2 \right) \right]^2$$

$$
\begin{aligned}
&+\frac{1}{4(\sum_{j=1}^{n} x_j)^4}\left\{\sum_{i=1}^{N} x_i\left[\sum_{j=1}^{n}(x_i-x_j)^2\left(m(x_j)\sum_{k\in s} x_k - x_j\sum_{k\in s} m(x_k)\right)\right]\right\}^2\\
&+\frac{\sum_{i=1}^{N} x_i}{2(\sum_{j=1}^{n} x_j)^2}\left(\frac{\sum_{j=1}^{n} m(x_j)}{\sum_{j=1}^{n} x_j}\right)\times\\
&\sum_{i=1}^{N} x_i\left(\frac{\sum_{j=1}^{n}(x_i-x_j)^4}{2!}-\frac{\sum_{j=1}^{n} x_j(x_i-x_j)^2\sum_{j=1}^{n}(x_i-x_j)^2}{\sum_{j=1}^{n} x_j}\right)\\
&\left(m(x_j)\sum_{k\in s} x_k - x_j\sum_{k\in s} m(x_k)\right).
\end{aligned}
$$

Bibliography

[1] Abramson, I.S., (1982). On bandwidth variation in kernel estimates - a square root law. *Annals Statist.* **10** 1217-1223.

[2] Akaike, H., (1974). A new look at statistical model identification. *IEEE Trans. Auto. Cont.* **19** 716-723.

[3] Azzalini, A., Bowman, A.W. and Härdle, W., (1989). On the use of nonparametric regression for model checking. *Biometrika* **76** 1-11.

[4] Becker, R.A., Chambers, J.M. and Wilks, A.R., (1988). *The New S Language.* Wadsworth & Brooks/Cole Advanced Books and Software: California.

[5] Benedetti, J., (1977). On the nonparametric estimation of regression functions. *J. Roy. Statist. Soc. B* **39** 248-253.

[6] Bierens, H.J., (1987). Kernel estimators of regression functions. In: Advances in Econometrics - Fifth world congress. Vol. 1. Cambridge University Press: New York.

[7] Bolfarine, H. and Zacks S., (1991). *Prediction Theory for Finite Populations.* Spinger-Verlag: New York.

[8] de Boor, C., (1978). *A Practical Guide to Splines.* Spinger-Verlag: New York.

[9] Bowman, A.W., (1984). An alternative method of crossvalidation for smoothing of density estimates. *Biometrika* **71** 353-360.

[10] Brewer, K.R.W., (1963). Ratio estimation and finite population: some results deductible from the assumption of an underlying stochastic process. *Aust. J. Statist.* **5** 93-105.

[11] Buja, A., Hastie, T.J. and Tibshirani, R.J., (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 435-555.

[12] Burman, P., (1989). A comparative study of ordinary cross-validation, ν-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76** 503-14.

[13] Carroll and Ruppert (1988). *Transformation and weighting in regression.* Chapman and Hall: New York.

[14] Cassel, C.M., Särndal, C.E. and Wretman, J.H., (1977). *Foundations of Inference in Survey Sampling.* Wiley: New York.

[15] Chambers, R.A. and Hastie, T.J., (1990). *Statistical Models in S.* Wadsworth and Brooks: California

[16] Chambers, R.L., (1993). Outlier robust sample survey inference. In: Proceedings of 49th session of the ISI conference, Italy.

[17] Chambers, R.L., Dorfman A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibaration. *J. Amer. Statist. Assoc.* **88** 268-277.

[18] Cheng, K.F. and Lin, P.E., (1981). Nonparametric estimation of a regression function. *Z. Wahrsh. verw. Gebiete* **57** 223-233.

[19] Chu, C-K. and Marron, S., (1991). Choosing a kernel regression estimator. *Statist. Science* **6** 404-436.

[20] Clark, R.M., (1975). A calibration curve for radiocarbon dates. *Antiquity* **49** 251-66.

[21] Clark, R.M., (1980). Calibration, Cross-validation and Carbon 14 II. *J. R. Statist. Soc. A* **143** 177-94.

[22] Cleveland, W.S., (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829-36.

[23] Cleveland, W.S. and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 597-610.

[24] Cochran, W.G., (1977). *Sampling Techniques.* Third Edition. Wiley: New York.

[25] Collomb, G., (1977). Properties of the kernel method for nonparametric regression estimation at a fixed point. *C. R. Acad. Sc. Paris* 285-289.

[26] Collomb, G., (1981). Estimation non-parametric de la regression: revue bibliographique. *Internat. Statist. Rev.* **49** 75-93.

[27] Collomb, G., (1985). Nonparametric regression: an up-to-date bibliography. *Statistics* **16** 309-24.

[28] Craven, P. and Wahba, G., (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377-403.

[29] Dorfman, A.H., (1993). Nonparametric regression for estimating totals in finite populations. Unpublished manuscript.

[30] Dorfman, A.H. and Hall, P., (1993). Estimators of finite population distribution functions using nonparametric regression. *Ann. Statist.* **21** 1452-1475.

[31] Eubank, R.L., (1988). *Spline Smoothing and Nonparametric Regression.* Dekker: New York.

[32] Family Expenditure Survey Annual Base Tapes, (1968-1983). Department of Employment, Statistics Division, Her Majestys Stationary Office, London. The data utilized in this thesis were available by the ESRC Data Archive at the University of Essex.

[33] Fan, J., (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998-1004.

[34] Fan, J., (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196-216.

[35] Fan, J. and Gijbels, I., (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008-2036.

[36] Fan, J., Heckman, N.E. and Wand, M.P., (1992). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. Unpublished manuscript.

[37] Firth, D., (1993). Recent developments in quasi-likelihoood methods. In: Proceedings of 49th session of ISI conference, Italy.

[38] Firth, D., Glosup, J. and Hinkley, D.V., (1991). Model checking with nonparametric curves. *Biometrika* **78** 245-52.

[39] Friedman, J. and Stuetzle, W., (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817-823.

[40] Gasser, Th. and Engel, J., (1990). Choice of weights in kernel regression estimation. *Biometrika* **77** 377-381.

[41] Gasser, Th. and Müller, H.G., (1979). Kernel estimation of regression functions. In: Smoothing techniques for curve estimation. Th. Gasser and M. Rosenblatt (eds.). Springer-Verlag: Heidelberg. pp. 23-68.

[42] Gasser, Th. and Müller, H.G., (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171-185.

[43] Gasser, Th., Müller, H.G. and Mammitzch, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. B* **47** 238-52.

[44] Geisser, S., (1979). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320-328.

[45] Green, P.J., (1985). Penalised likelihood for general semi-parametric regression models. *Internat. Statist. Rev.* **55** 245-260.

[46] Green, P. J. and Silverman, B. W., (1994). *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach.* Chapman and Hall: London.

[47] Green, P.J. and Yandell, B., (1987). Semi-parametric generalized linear models. Proceedings 2nd International GLIM conference. Lancaster, Lecture notes in Statistics. No 32 pp. 44-55. Spinger-Verlag: New York.

[48] Godambe, V.P. and Joshi, V.M., (1965). Admissibility and Bayes estimation in sampling finite populations, 1. *Ann. Math. Statist.* **36** 1707-1722.

[49] Godambe, V.P. and Thompson, M.E., (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Internat. Statist. Rev.* **54** 127-138.

[50] Hall, P., (1990). On the bias of variable bandwidth curve estimators. *Biometrika* **77** 529-536.

[51] Hall, P. and Marron, J., (1987). Extent to which least squares crossvalidation minimises integrated squared error in nonparametric density estimation. *Prob. Theory Rel. Fields* **74** 567-582.

[52] Hansen, M.H., Maddow, 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.

[53] Hanurav, (1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *J. Roy. Statist. Soc. B* **29** 374-91.

[54] Härdle, W., (1990a). *Applied Nonparametric Regression.* Cambridge University Press: New York.

[55] Härdle, W., (1990b). *Smoothing Techniques with Implementation in S.* Spinger-Verlag: New York.

[56] Härdle, W. and Marron, S., (1985a). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72** 481-484.

[57] Härdle, W. and Marron, S., (1985b). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465-81.

[58] Härdle, W. , Hall, P. and Marron, S., (1988). How far are the optimally chosen smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86-95.

[59] Hastie, T.J. and Loader, C., (1993). Local regression: automatic kernel carpentry. (with comments) *Statist. Science* **8** 120-143.

[60] Hastie, T.J. and Tibshirani, R.J., (1986). Generalised Additive Models (with discussion). *Statist. Science* **1** 297-318.

[61] Hastie, T.J. and Tibshirani, R.J., (1990). *Generalized Additive Models.* Chapman and Hall: London.

[62] Hastie, T.J. and Tibshirani, R.J., (1993). Varying coefficient models (with discussion) *J. Roy. Statist. Soc. B* **55** 757-796.

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

[64] Horvitz, D.G. and Thompson, D.J., (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663-685.

[65] Jennen-Steinmetz, C. and Gasser, Th., (1988). An unifying approach to nonparametric regression estimation. *J. Amer. Statist. Assoc.* **83** 1084-1089.

[66] Jones, M.C., (1990). Variable kernel density estimates and variable kernel density estimates. *Aust. J. Statist.* **32** 361-371.

[67] Jones, M.C., Davies, S.J. and Park, B.U., (1994). Versions of kernel-type regression estimators. *J. Amer. Statist. Assoc.*, to appear.

[68] Jones, M.C., (1993). Simple boundary correction for Kernel Denisty Estimates. *Statistics and Computing* **3** 135-146.

[69] Jones, M.C. and Wand, M.P., (1994). *An Introduction to Kernel Smoothing.* Chapman and Hall, London.

[70] Kendall, M.G. and Stuart, A., (1972). *The Advanced Theory of Statistics.* Second Edition. Volume 2. Griffin: London.

[71] Kuk, A.Y.C., (1993). A kernel method for estimating finite population distribution functions using auxilary information. *Biometrika* **80** 385-92.

[72] Mallows, C., (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

[73] Marron, S., (1988). Automatic smoothing parameter: a survey. *Empirical Econom.* **13** 187-208.

[74] McCullagh, P. and Nelder, J.A., (1989). *Generalized Linear Models.* Second Edition. Chapman and Hall: London.

[75] Müller, H.G., (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231-8.

[76] Müller, H.G. and Stadtmüller, U., (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15** 182-201.

[77] Nadaraya, E.A., (1964). On estimating regression. *Theor. Probab. Appl.* **9** 141-142.

[78] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Roy. Statist. Soc. A* **135** 370-84.

[79] Njenga, E.G., (1990). Robust estimation of the regression coefficients in complex surveys. Unpublished Doctoral thesis, University of Southampton.

[80] Park, B.U. and Marron, S., (1989). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **84** 66-72.

[81] Pfeffermann, D., (1993). The role of sampling weights when modelling survey data. *Internat. Statist. Rev.* **61** 317-337.

[82] Priestley, M.B. and Chao, M.T., (1972). Nonparametric function fitting. *J. Roy. Statist. Soc. B* **34** 384-392.

[83] Rice, J.A., (1984a). Boundary modifications for kernel regression. *Commun. Statist. A* **13** 893-900.

[84] Rice, J.A., (1984b). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215-1230.

[85] Rice, J.A. and Rosenblatt, M., (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Statist.* **11** 141-56.

[86] Rosenblatt, M., (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 642-69.

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

[88] Royall, R.M., (1988). The Prediction Approach to Sampling Theory. In: P.R. Krishnaiah and C.R.Rao (eds.). *Handbook of Statistics*, Vol. 6. Amsterdam: North Holland, pp. 399-413.

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

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

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

[92] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65-78.

[93] Särndal, C., Swensson, B. and Wretman,J., (1992). *Model Assisted Survey Sampling.* Spinger-Verlag: New York.

[94] Schumaker, L.L., (1981). *Spline Functions: Basic Theory.* John Wiley: New York.

[95] Scott, D.W. and Terrell, G.R., (1987). Biased and unbiased crossvalidation in density estimation. *J. Amer. Statist. Assoc.* **82** 1131-1146.

[96] Sheather, S.J. and Jones, M.C., (1991). A reliable data-based bandwidth selector method for kernel density estimation. *J. Roy. Statist. Soc. B* **53** 683-690.

[97] Shibata, R., (1981). An optimal selection of regression variables. *Biometrika* **68** 45-54.

[98] Silverman, B.W., (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12** 898-916.

[99] Silverman, B.W., (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. B* **47** 1-52.

[100] Smith, T.M.F., (1994). Sample Surveys 1975-1990; An age of Reconciliation? *Inter. Statist. Rev.* **62** 5-34.

[101] Smith, T.M.F. and Njenga, E., (1992). Robust Model-based Methods for Analytic surveys. *Survey Methodology* **18** 187-208.

[102] Staniswallis, J.G., (1989). Local bandwidth selection for kernel estimates. *J. Amer. Statist. Assoc.* **84** 284-288.

[103] Stone, C.J., (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.* **5** 549-645.

[104] Stone, M., (1974). Crossvalidatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. B* **36** 111-47.

[105] Sukhatme, P.V., (1954). *Sampling theory of surveys, with applications.* Iowa State College Press: Ames, Iowa.

[106] Tibshirani, R.J. and Hastie, T.J., (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559-68.

[107] Titterington, D.M., (1985). Common structures of smoothing techniques in statistics. *Internat. Statist. Rev.* **53** 141-170.

[108] Wahba, G., (1975). Smoothing noisy data with spline functions. *Numer. Math.* **24** 383-93.

[109] Wahba, G., (1990). *Spline functions for observational data.* CBMS-NSF. Regional conference series. SIAM. Philadelphia.

[110] Watson, G.S., (1964). Smooth regression analysis. *Sankhya A* **26** 359-372.

[111] Wedderburn, R. W. M., (1974). Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika* **63** 27-32.

[112] Wegman, E.J. and Wright, I.W., (1983). Splines in statistics. *J. Amer. Statist. Assoc.* **78** 351-65.