RELEVANCE WEIGHTED SMOOTHING AND A NEW BOOTSTRAP METHOD

by

FEIFANG HU

B.Sc., Hangzhou Normal University, P.R. China, 1985

M.Sc., Zhejiang University, P.R. China, 1988

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1994

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Oct. 13, 1994

DE-6 (2/88)

Relevance Weighted Smoothing and a New Bootstrap Method

Abstract

This thesis addresses two quite different topics. First, we consider several relevance weighted smoothing methods for relevant sample information. This topic can be viewed as a generalization of nonparametric smoothing. Second, we propose a new bootstrap method which is based on estimating functions.

A statistical problem usually begins with an unknown object of inferential of interest. About this unknown object, we may have three types of information (classified in this thesis): direct information, exact sample information and relevant sample information. Almost all classical statistical theory is about direct information and exact sample information. In many cases, relevant sample information is available and useful. But there is no systematic theory about relevant sample information. The problem of this thesis is to extract "the relevant information " contained in the relevant samples. Three general methods have been developed under three different lines of approach (parametric, nonparametric and semiparametric approach). In the parametric approach, we propose the idea of relevance weighted likelihood (REWL). For the nonparametric approach, we develop our theory based on the relevance weighted empirical distribution function (REWED). In the semiparametric approach, the relevance weighted estimating functions are used to extract "relevant information" from relevant samples. From asymptotic results, we find that these proposed methods have many desirable properties.

We apply these proposed methods as well as some adjusted methods to generalized smoothing problems. Theoretical results as well as simulation results show our methods to be promising.

We also present a new bootstrap method. It has computational and theoretical advantages over conventional bootstrap methods when the data obtain from non-identically distributed observables. And it differs from conventional methods in that it resamples components of an estimating function rather than the data themselves.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor James Zidek for his excellent guidance, for his constant encouragement and his patience. Only the support from James Zidek made this thesis possible.

I wish to thank the people whose suggestions and comments about this thesis have been most helpful: Harry Joe, Nancy Heckman and Jean Meloche.

This thesis benefitted from the helpful references and comments of Nancy Heckman.

Many thanks go to Harry Joe, Nancy Heckman, Jian Liu and John Petkau for their encouragement and support during my stay at UBC.

Many thanks go to my dear wife Ying Liu for her constant encouragement and patience. Our son Leon made the life more enjoyable by so much fun he gave to me.

Finally, I would like to thank James Zidek for his financial support. The financial support from the Department of Statistics is acknowledged with great appreciation. I also acknowledge the support of the University of British Columbia through a University Graduate Fellowship.

Chapter 1

Introduction

This thesis addresses two quite different topics. First, we consider several relevance weighted smoothing methods for relevant sample information. This topic can be viewed as a generalization of nonparametric smoothing. Second, we propose a new bootstrap method which is based on estimating functions.

A statistical problem usually begins with an unknown object of inference. About this unknown object of inference, we may have three types of information (classified in this thesis): direct information, exact sample information and relevant sample information. Almost all classical statistical theory is about direct information and exact sample information. In many cases, the relevant sample information is available and useful. There is no systematic theory about relevant sample information.

The problem of this thesis is to extract "the relevant information " contained in the relevant samples. Three general methods have been developed under three different lines of approach (parametric, nonparametric and semiparametric approach). For the parametric approach, we propose the idea of relevance weighted likelihood (REWL). The REWL plays the same role in relevant sample analysis as the likelihood in classical statistical in-

ference. For the nonparametric approach, we develop our theory based on the relevance weighted empirical distribution function (REWED). In the semiparametric approach, the relevance weighted estimating functions are used to extract "relevant information" from relevant samples. We use estimating functions because of their generality.

We show that the maximum relevance weighted likelihood estimator (MREWLE) is consistent and asymptotically normal. The asymptotic theory of the nonparametric approach is also developed. [For the semiparametric approach, the asymptotic theory is omitted.] By these asymptotic results, we find that these proposed methods have many desirable properties.

We also apply these proposed methods to generalized smoothing methods. We find that: (i) the MREWLE has some advantages over current nonparametric regression methods; (For example, the MREWLE always has a smaller variance which depends on the Fisher information function. This also indicates that the MREWLE is a kind of efficient estimator.) (ii) the relevance weighted quantile estimator (based on the REWED) is usually robust and quite efficient. For generalized smoothing models, we get some good estimators by some locally polynomial adjustments. Simulations support our methods.

Various papers (Efron 1979, Bickel and Freedman 1981 and Singh 1981) speak to the quality of the bootstrap resampling procedure for estimating a sampling distribution in situations where the sampled observables are independent and identically distributed. In this thesis we present a new bootstrap method. It has computational and theoretical advantages when the data obtain from non-identically distributed observables. And it differs from conventional bootstrap methods in that it resamples components of an estimating function rather than the data themselves. We apply this bootstrap method

to ordinary linear regression. By comparing with Efron's method, Freedman's pairs method and Wu's method, our method gets support from theoretical results as well as simulations.

Because our new bootstrap method is based on estimating functions, applying this bootstrap method to relevance weighted smoothing is possible and reasonable. This is a further research topic.

We organize the thesis as following.

In Chapter 2, we classify the different types of information about the unknown object of inference (parameter) in statistical way. In classical statistical inference and Bayesian inference, statisticians usually focus on direct information and exact sample information. We learn that relevant sample information is very important in statistical inference. Two possible generalizations of Fisher's information and Kullback-Leibler's information for relevant samples are considered.

For relevant samples, we propose the idea of the relevance weighted likelihood (REWL) in Chapter 3. Our idea generalizes that of the likelihood function in that the independent samples going into the likelihood function may be discounted according to their degree of relevance. The classical likelihood obtains in the special case where the independent samples are all from the study population whose parameters are of inferential interest. But more generally, as in metaanalysis, for example, the value of such samples may be reduced in that their relevance is in doubt because for example, they are noisy or biased. The relevance weights, which enter as exponents of the component factors in the sampling density, enable us to tradeoff information against such things as bias in the samples which may be relevant even if not drawn from the study population.

We show how the REWL can obtain from a generalization of the entropy maximization principle. Using the REWL we define the notion of weak sufficiency and the maximum REWL estimator (MREWLE). By using the MREWLE in a normal example, we find the MREWLE has some advantages.

In Chapter 4, we establish the weak and strong consistency of the MREWLE under a wide range of conditions. My results generalize those of Wald (1948) to both nonidentically distributed random variables and unequally weighted likelihoods (when dealing with independent data sets of varying relevance to the inferential problem of interest). Asymptotic normality of the MREWLE is also proved.

We apply the REWL methods to generalized smoothing models in Chapter 5. By choosing different weights, four estimators are considered. They are the: Nadaraya-Watson MREWLE, Gasser-Muller MREWLE, k-NN weights MREWLE and locally linear smoother MREWLE. Asymptotic results for these four estimators are developed. We also compare them by theoretical results as well as simulations.

Chapter 6 concerns situations in which a sample $X_1 = x_1, \cdots, X_n = x_n$ of independent observations is drawn from populations with different CDF's $F_1, \cdots, F_n$, respectively. Inference is about a quantile of another population with CDF F_0 when the data from the other populations are thought to be "relevant". Nonparametric smoothing of a quantile function would typify situations to which our theory applies. We define the relevance weighted quantile (REWQ) estimator derived from the relevance weighted empirical distribution (REWED) function. We show that the estimator has desirable asymptotic properties. A simulation study is also included. It shows that the median estimator is a robust alternative to the locally weighted averages used in conventional smoothing.

Some further results appear in Chapter 7. We propose a locally polynomial MREWLE.

By comparing with locally constant MREWLE (in Chapter 5) and current nonparametric regression methods, we find that the locally linear MREWLE has a simple bias and smaller variance. Some simulation results also support this locally linear MREWLE. For our semiparametric approach, we present the method of relevance weighted estimating functions. We find that the locally linear MREWLE is the best among all locally linear REW estimating equations (with kernel weights) and the locally linear quasi-MREWLE (maximum relevance weighted quasi-likelihood estimator) is the best among all locally linear REW linear estimating equations.

Chapter 8 presents a method of bootstrap estimation. Rather than resampling from the original sample, as is conventional, the proposed method resamples summands of the estimating function used to produce the original estimate. The result is computation simpler than existing competitors. However, its main advantages lie in its treatment of non-identically distributed observations. Shortcomings of conventional methods are overcome. An application to ordinary linear regression is worked out in detail along with the appropriate asymptotic theory. We report as well a simulation study which provides support for this new bootstrap method.

Chapter 2

Relevant Sample Information

2.1 Introduction

Information is a key word in statistics. After all, this is what the subject is all about. As Basu (1975) says:

- *A problem in statistics begins with a state of nature, a parameter of interest θ about which we do not have enough information. In order to generate further information about θ, we plan and then perform a statistical experiment $\mathcal{E}$. This generates the sample x. By the term 'statistical data' we mean such a pair $(\mathcal{E},x)$ where $\mathcal{E}$ is well-defined statistical experiment and x the sample generated by a performance of the experiment. The problem of data analysis is to extract 'the whole of relevant information'–an expression made famous by R. A. Fisher–contained in the data $(\mathcal{E},x)$ about the parameter θ.*

However, statistical theory has traditionally been concerned with a narrow interpretation of the word embraced by Basu's description. Given data, statisticians would typically

construct a sampling model with a parameter θ to describe a population from which the data were supposedly drawn. Information in the sample about the population comes out through inference about θ. Alternatively, given a θ of interest the classical paradigm sees the statistician as conducting a statistical experiment to generate a sample from a population defined by a sampling distribution with parameter θ. The sample then provides information about θ. In either case, statistical inference will be based on these observations and their directly associated sampling model. This is the frequency theory viewpoint. (See Lehmann 1983).

Bayesians think that we always have some prior distribution for the unknown parameter. Then we combine the prior information and the 'statistical data' information. (See Lindley 1965).

But these two sources of information are not the only source of information for the parameter. There is another source of information (relevant sample information as defined in Section 2.2). This information is very useful in many cases. This can be clearly seen by the following examples. (We use an example similar to that of Basu (1975) but for a different purpose).

Example 2.1 *Suppose an urn A contains 100 tickets that are numbered consecutively as* $\theta+1, \theta+2, \cdots, \theta+100$ *where* θ *is an unknown number. Let* $\mathcal{E}_n$ *stand for the statistical experiment of drawing a simple random sample of n tickets from the urn A without replacement and then recording the sample as a set of n numbers* $x_1 < x_2 < \cdots < x_n$. *Suppose we know that* θ *is bounded by 50, this means* $|\theta| \leq 50$. *This information is a kind of direct information (or prior information). Consider now the hypothetical situation where* $\mathcal{E}_{25}$ *has been performed and has yielded the sample* $x = (55, 57, \cdots, 105)$, *where 55 and 105 are respectively the smallest and largest number drawn. To be specific, with*

data $\mathcal{E}_n, x = (x_1, x_2, \cdots, x_n)$, *we know without any shadow of doubt that the true value of* θ *must belong to the set*

$$S = \{x_1 - 1, x_1 - 2, \cdots, x_1 - m\}$$

where $m = 100 - (x_n - x_1)$. *Now with the information from the data we can now assert that* θ *is an integer that lies somewhere in the interval* $[5, 54]$. *Combine the direct information and the 'statistical data' information, we could conclude that* θ *is an integer that lies somewhere in the interval* $[5, 50]$.

Now suppose another urn B *contains other 100 tickets that are numbered consecutively as* $\theta_1 + 1$, $\theta_1 + 2, \cdots, \theta_1 + 100$ *where* θ_1 *is another unknown number. But we known that the difference between* θ *and* θ_1 *is smaller than 5, meaning that* $|\theta - \theta_1| < 5$. *Suppose now that we draw a simple random sample of 5 tickets from the urn* B *and find the sample* $x' = (51, 80, \cdots, 149)$. *With this information, we can assert that* $\theta_1 = 50$ *or* 49. *Now with the information from the data* x', *we can say that* θ *is an integer that lies somewhere in the interval* $[45, 54]$. *This is the information from the data* x' *for the parameter* θ.

Combining the information from these different sources, we can finally conclude that θ *is an integer somewhere in the interval* $[45, 50]$.

From the above example, the data x' contain some useful information for the parameter θ. But x' are not from the experiment $\mathcal{E}$ with the parameter θ (the experiment we plan and perform). The x' come from the experiment with the unknown parameter θ_1, which has some relation with the parameter θ (in this example, $|\theta - \theta_1| < 5$). This is the difference between x' and x. The classical statistical inference usually focuses on the information from x alone.

In this chapter, we will focus on the discussion of this third source of information. We come out very strongly in support of the use this information for statistical inference when it is available.

In Section 2.2, we try to classify information about the unknown parameter into two main types. The first is "direct information" and the second we call "sample information" (as defined in Section 2.2). We classify sample information into exact sample information and relevant sample information. Almost all classical statistical theory is about direct information and exact sample information. Relevant sample information has only been used in some special contexts of statistics. But systematic theory is not available and needs to be developed.

We have used the word information many times in this section. But, what is information? Or how to measure information? As Basu (1975) remarks no other concept in statistics is more elusive in its meaning and less amenable to a generally agreed definition. The measure of information plays a very important role in statistics and communication science. A lot of well-known work has been done in this area (See Fisher (1925), Shannon (1948), Wiener (1948) and Kullback (1959)). We will review three of the most important information measures in the Section 2.3.

All these information measurements are used for direct information and exact sample information for different purposes. In Section 2.4, we discuss how to measure the relevant sample information. A possible generalization of Fisher information is proposed.

The use of the relevant sample information in the test of hypotheses or discrimination will be considered in Section 2.5. We propose a reasonable measure of information for discrimination. This measure is a generalization of the Kullback-Leibler information measure.

2.2 Classification of Information

Statistics is concerned with the collection of data and with their analysis and interpretation. Here we do not consider the problem of data collection but take the data as given and ask what they tell us. The answer depends not only on the data, but also on background knowledge of the situation; the latter is formalized in the assumptions with which the analysis is entered. We review two traditional principal lines of approach.

Classical inference and decision theory. The observations are now postulated to be the values taken on by random variables which are assumed to follow a joint probability distribution, P, belonging to some known class $\mathcal{P}$. Frequently, the distributions are indexed by a parameter, say θ, taking values in a set, Ω, so that

$$\mathcal{P} = \{P_\theta, \theta \in \Omega\}.$$

The aim of the analysis is then to specify a plausible value for θ (this is the problem of point estimation) or at least to determine a subset of Ω of which we can plausibly assert that it does, or does not, contain θ (estimation by confidence sets or hypothesis testing). Such a statement about θ can be viewed as a summary of the information provided by the data and may be used as a guide to action.

In this approach, statistical inference about θ is based on both this directly associated sampling model (here $\{P_\theta, \Omega\}$) and the observations.

Bayesian analysis. In this approach, it is assumed in addition that θ is itself a random variable (though unobservable) with a known distribution. This prior distribution (specified prior to the availability of the data) is modified in light of the data to determine a posterior distribution (the conditional distribution of θ given the data), which

summarizes what can be said about θ on the basis of assumptions made and the data.

This Bayesian approach about the parameter θ is based on both the directly associated model, the prior distribution and the data.

It is frequently reasonable to assume that we get some other observations which are not from P_θ, but from some P_{θ_1} where P_{θ_1} is related to P_θ. These observations do contain some information about θ. But the above two traditional principal lines of approach do not include these observations.

Example 2.2 *Assume that we wish to estimate the probability (a parameter θ_A) of a penny A showing heads when spun on a flat surface. We usually consider n spins of the penny as a set of n binomial trials with an unknown probability θ_A of showing heads. Suppose, however, that we have m spins of the penny B. If we believe this penny B is similar to penny A (meaning θ_A and θ_B are close to each other), to estimate θ_A, it might be reasonable to use the information from the m spins of penny B.*

The above discussion leads to the following classification of information about the unknown object of inference (here parameter θ).

Definition 2.1 (Direct information). *All the information which is directly related to the unknown object of inference (parameter) is called* **direct information** *of the parameter θ.*

Definition 2.2 (Sample information). *We call* **sample information**, *the information about θ from the sample or the data. If the sample is from the experiment (model) which is direct to the parameter θ, we call it an* **exact sample** *for this parameter θ.*

The information in the exact sample is called **exact sample information**. *The sample which is from the experiment (model) related to the parameter θ (not direct) is called a* **relevant sample** *for parameter θ. The* **relevant sample information** *is defined as the information from a relevant sample.*

As in Example 2.1, $|\theta| \leq 50$ is direct information about θ. The data of $\mathcal{E}_{25}$ is exact sample information. The data drawn from urn B are relevant sample information.

In *classical inference and decision theory*, statistical inference about θ is based on both the directly associated sampling model $\{P_\theta, \Omega\}$ (direct information) and the observations (exact sample information). The Bayesian approach based on the directly associated sampling model (direct information), the prior distribution (direct information) and the data (exact sample information).

In some cases, we may have relevant sample information about the inferential objective. This can be well-illustrated by the examples of the following Chapters. Examples 3.1, 3.2, 3.6 and 6.1 all indicate that relevant sample information is available and useful.

In the following examples, we show how information is classified.

Example 2.3 Linear Model. *Observations y, considered as an $n \times 1$ column vector, is a realization of the random vector Y with $E(Y) = x\beta$, where x is a known $n \times q$ matrix of rank $q \leq n$, and β is a q dimensional column vector of unknown parameters. β are the parameters of interest.*

Because the model involves the unknown parameters directly, the information from the observations y about β is exact sample information.

More generally, for the **Generalized Linear Model** (See McCullagh and Nelder 1989), the information from the sample is still exact sample information, because in that case, the experiment involves the unknown parameters directly.

Example 2.4 *Let X be from $N(\theta, 1)$, and let the estimand be θ, θ has a convenient prior distribution, say $N(0, 1)$.*

Now let us assume Y is from $N(\theta_1, 1)$, with θ_1 unknown. But we do know that $\theta - \theta_1$ has a prior distribution, say $N(0, 1)$.

The data vaule X is an exact sample for θ. The prior distribution of θ is direct information for θ. The data value Y is a relevant sample for θ.

The statistical methods using direct information and exact sample information are well developed. We can easily find these methods in standard textbooks. However there are no systematic methods about how to use relevant sample information.

2.3 Measure of Information

In Section 2.2, we classify the different types of information. But what is information? How can we measure the information? Answers of these questions seem controversial. The definition of information or entropy goes backed to 1870, and a series of papers by L.Boltzmann. Since then, statisticians have proposed many different definitions for different targets. Now we review the three most important definitions of statistical information.

(I) Shannon and Wiener Information (Entropy).

The statistical interpretation of thermodynamic entropy, a measure of the unavailable energy within a thermodynamic system, was developed by L.Boltzmann around 1870. His first contribution was the observation of the monotone decreasing behavior in time of a quantity defined by

$$E = \int_0^\infty f(x,t) \log\{\frac{f(x,t)}{\sqrt{x}}\} dx,$$

where $f(x,t)$ denotes the frequency distribution of the number of molecules with energy between x and $x + dx$ at time t (Boltzmann, 1872). When the distribution f is defined in terms of the velocities and positions of the molecules, the above quantity takes the form

$$E = \int f \log f dx dy,$$

where x and y denote the vectors of the position and velocity, respectively. Boltzmann showed that for some gases this quantity, multiplied by a negative constant, was identical to the thermodynamic entropy.

Shannon (1948) proposed the definition of the entropy of a probability distribution: (the negative of the above quantity)

$$H = -\int p(x) \log p(x) dx, \tag{2.1}$$

where $p(x)$ denotes the probability density with respect to the measure dx.

Shannon entropy plays a very important role in modern communication theory. There are almost uncountably many papers and books about the use of Shannon entropy. The quantity H is simply referred to as a measure of information, or uncertainty, or randomness.

This definition may be used in the measure of direct information, when the direct information is Bayes prior. Also it may be used in the case of a posterior distribution.

However, Savage (1954, page 50) remarks: 'The ideas of Shannon and Wiener, though concerned with probability, seem rather far from statistics. It is, therefore, something of an accident that the term 'information' coined by them should be not altogether inappropriate in statistics.'

(II) Fisher's Information.

R. A. Fisher's (1925) measure of the amount of information supplied by data about an unknown parameter is well-known to statisticians. This measure is the first use of 'information' in mathematical statistics, and was introduced especially for the theory of statistical estimation.

For a real parameter and a density function satisfying Cramer-Rao regularity conditions it has the form

$$I(\theta) = \int \{\frac{\partial \log f(x,\theta)}{\partial \theta}\}^2 f(x,\theta)dx \tag{2.2}$$

We know that Fisher's Information has a lot of optimal properties as a measure of information: (i) Fisher's information, being specific to a parameter, will stay the same if we reduce to a sufficient statistic; (ii) Fisher's information is additive over different sets of independent data; and (iii) Fisher's information gives a lower variance bound for the estimation of the parameter provided some regularity conditions are satisfied.

(III) Kullback-Leibler's Information.

Kullback and Leibler (1951) consider a definition of information for 'discrimination in favor of H_1 against H_2'. Here H_i, $i = 1, 2$, is the hypothesis that X is from the statistical population with probability measure μ_i, with $d\mu_i(x) = f_i(x)dx$. They define the logarithm of the likelihood ratio, $\log\{f_1(x)/f_2(x)\}$, as **the information in $X = x$ for**

discrimination in favor of H_1 against H_2. The mean information for discrimination in favor of H_1 against H_2 per observation from H_1 is defined as

$$I(1:2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx. \tag{2.3}$$

This definition is a departure from Shannon and Wiener's information. It is widely used in statistics for discrimination. This can be easily seen from Kullback (1959).

The above three definitions of information are all about direct information and exact sample information. This is because classical statistics is focused on these two types of information.

In the following two sections, we will discuss the information measure for relevant sample information.

2.4 Fisher's Information of Relevant Sample in Point Estimation.

It is well-known that Fisher's information plays a very important role in the theory of statistical estimation. In the last section, we have discussed Fisher's information for exact samples. Now we will propose a possible generalization of Fisher's information to the relevant sample case. The bias function and information function in Chapters 4, 5, and 7 offer other possible generalizations.

Let us begin with the simplest case. Assume that X has density $f(x, \theta_o)$, where θ_o is a real parameter. Suppose $f(x, \theta_o)$ satisfies the Cramer-Rao regularity conditions; then the Fisher's Information for θ_o from X as defined in Section 2.3 is (2.2). We are interested in the parameter θ. As we have claimed in Section 2.2, there is some

information in X for the parameter θ, if we know that $|\theta - \theta_o| \leq c$ for some constant c. Now it is natural to ask how to measure the information in X for the parameter θ.

Definition 2.3 *The information for the parameter θ in X is defined as $Inf_X(\theta, \theta_o) = (Inf_X(\theta_o), \theta - \theta_o)$. Here $Inf_X(\theta_o)$ is the Fisher's information.*

In the above definition, we can see that the information in X for the parameter θ contains two parts, one is the information part; another is the bias part. If the bias part is 0, which means that the $\theta = \theta_o$, then the above information measure becomes Fisher information. If we know the bias exactly, then the bias can be eliminated. This is because then we can transform the parameter. Generally, if θ and θ_o have a one-to-one relationship, then we always can do the parameter transformation. This will be the case of exact sample information.

Now we discuss the properties of the above definition of information. In the following discussion, we always assume the bias is the same, unless we specify otherwise. When we say $Inf_X(\theta, \theta_o)$ equals $Inf_Y(\theta, \theta_1)$, we mean $\theta_o = \theta_1$ and $Inf_X(\theta_o) = Inf_Y(\theta_1)$. Only when the bias is the same, can we order the two information indices.

- **a. $Inf(\theta, \theta_o)$ is independent of the σ-measure μ.**

  As we know $Inf(\theta, \theta_o)$ is calculated from the $f(x, \theta_o) = dP_{\theta_o}(x)/d\mu$. If we can find another σ–measure ν such that $\{P_\theta : \theta \in \Theta\} \ll \nu$, then we can replace $f(x, \theta_o)$ by $f^*(x, \theta_o) = dP_{\theta_o}(x)/d\nu$. The value $Inf(\theta, \theta_o)$ is not changed.

  The proof is easy and we omit it here. □

- **b. The information of several independent observations is the sum (appropriately defined) of the information in these observation, if the bias of these observations are the same.**

Mathematically, the above statement says suppose $X_1, \cdots, X_k$ are independent and $X = (X_1, \cdots, X_k)$. If $f_i(x_i, \theta_o)$ is the density of X_i, and they satisfy the Cramer-Rao regularity conditions, then

$$f(x, \theta_o) = f_1(x_1, \theta_o) \cdots f_k(x_k, \theta_o)$$

satisfies the Crame-Rao regular conditions, and

$$Inf(\theta, \theta_o) = \{Inf_1(\theta_o) + \cdots + Inf_k(\theta_o), \theta - \theta_o\}, \tag{2.4}$$

here $Inf(\theta, \theta_o)$ are the information function for θ in X.

The proof of the above result is exactly similar to that of Fisher's Information. We omit the proof here. □

- **c. The information will not increase, when we transform the data.**

  Let $Y = T(x)$ be a statistic, that is , T is a function with domain $\mathcal{X}$ and range $\mathcal{Y}$, and let $\mathcal{T}$ be an additive class of subsets of $\mathcal{Y}$. We assume that T is measurable. Let $g(t, \theta_o)$ be the density of $T(X)$. If $f(x, \theta_o)$ and $g(t, \theta_o)$ satisfy the Cramer-Rao regularity conditions, then

$$Inf_X(\theta, \theta_o) \geq Inf_T(\theta, \theta_o), \tag{2.5}$$

  here $(x_1, y) \geq (x_2, y)$ means $x_1 \geq x_2$.

  We omit the proof here because it is a direct result from the Fisher's Information. □

- **d. Under the conditions of c), we have inequality in (2.5), with equality if and only if the statistic $Y = T(x)$ is sufficient for the parameter θ_o.**

  The proof is omitted. □

In connection with the basic properties of information, we have the following comments.

1. From (2.4), when we have n iid observations from $f(x, \theta_o)$, then $Inf(\theta, \theta_o) = (nInf(\theta_o), \theta - \theta_o)$.

2. Above, d) implies that we should use the sufficient statistic to do the statistic analysis for the parameter θ, although this sufficient statistic is for the population indexed by parameter θ_o. This result tells us how to reduce the dimension of the data.

3. If we do a one-to-one transformation of the parameter, that is $\eta = h(\theta)$ and h is differentiable, the information that X contains about η is

$$Inf_X(\eta, \eta_o) = (Inf_X(\theta_o)[h'(\theta_o)]^2, \eta - \eta_o),$$

where $\eta_o = h(\theta_o)$.

Now we are going to discuss the generalized information inequality, which generalizes the information inequality to the relevant sample case.

Theorem 2.1 (Relevant Information Inequality) *Suppose that Cramer-Rao regularity conditions are satisfied. Let δ be any statistic with $E_{\theta_o}(\delta^2) < \infty$ for which the derivative with respect to θ_o of $h(\theta_o) = E_{\theta_o}(\delta)$ exists and can be obtained by differentiating under the integral sign. Then*

$$E(\delta - \theta)^2 \geq \frac{h'(\theta_o)^2}{Inf(\theta_o)} + (h(\theta_o) - \theta)^2. \tag{2.6}$$

Proof. The result follows directly from

$$E(\delta - \theta)^2 = Var(\delta) + (h(\theta_o) - \theta)^2$$

and the Information Inequality. □

Usually we can control the value of $|\theta - \theta_o|$, but not $|h(\theta_o) - \theta|$. We would like to choose $h(\theta_o) = \theta_o$. From the above Theorem, we obtain.

Corollary 2.1 *Under the conditions of Theorem 1, for any θ_o's unbiased estimator δ,*

$$E(\delta - \theta)^2 \geq \frac{1}{Inf(\theta_o)} + (\theta_o - \theta)^2. \tag{2.7}$$

From the Corollary, we can see that our definition of information for relevant samples is reasonable. The lower bound of the mean square error depends both on the information and the bias.

Now we consider how to combine the information from relevant samples having different bias parts. Let X have density $f(x, \theta_o)$ and Y, the density $g(y, \theta_1)$. Both $f(x, \theta_o)$and $g(y, \theta_1)$ satisfy the Cramer-Rao regularity conditions. θ is the parameter of interest. Both X and Y contain some information about θ, so we need to combine their information.

From Definition 2.3, we get $Inf_X(\theta, \theta_o)$ and $Inf_Y(\theta, \theta_1)$. Then the information from (X, Y) is defined as $\{I(\theta_o, \theta_1), B(\theta_o, \theta_1)\}$, here $I(\theta_o, \theta_1) = diag(Inf_X(\theta_o), Inf_Y(\theta_1))$ and $B(\theta_o, \theta_1) = (\theta - \theta_o, \theta - \theta_1)^t$. This is similar to Fisher's information for the multiparameter case except for our inclusion of the bias.

We can easily obtain the following result.

Theorem 2.2 *Let $\delta(X, Y)$ be any statistic with $E(\delta(X, Y)^2) < \infty$ such that derivative with respect to θ_o and θ_1 of $h(\theta_o, \theta_1) = E(\delta(X, Y))$ exists and can be obtained by differentiating under the integral sign. Then*

$$E(\delta(X, Y) - \theta)^2 \geq \beta^t I^{-1}(\theta_o, \theta_1)\beta + (h(\theta_o, \theta_1) - \theta)^2, \tag{2.8}$$

where $\beta^t = (\partial h(\theta_o, \theta_1)/\partial\theta_o, \partial h(\theta_o, \theta_1)/\partial\theta_1)$.

We stop this section here. Further theory is under development but not yet complete.

2.5 Kullback and Leibler's Information in Relevant Sample in Discrimination

Let us begin with the following example.

Example 2.5 Simple null hypothesis. *Assume* X_1 *is a sample from* $N(\theta, 1)$. *The null hypothesis is* H_0: $\theta = 0$ *and the alternative* H_a: $\theta = 2$. *From the Neyman-Pearson Lemma, we can easily get the most powerful test of level* $\alpha = .05$ *as: if* $X_1 > 1.645$, *we reject the null hypothesis* H_0. *Otherwise, we accept the* H_0. *The power of this test is* $\beta = .639$.

Now suppose we get another sample X_2 *from* $N(\theta_o, 1)$. *We know that* $|\theta - \theta_o| \leq .5$. *We construct a new test: if* $X_1 + X_2 > 2.826$, *we reject the null hypothesis* H_0. *Otherwise we accept* H_0.

For the second test, we have $\sup P_{H_0}(reject\ H_0) \leq .05$ *and* $\inf P_{H_A}(reject\ H_0) \geq .683 > .639$. *So the second test is more powerful than the first one. This means the observation* X_2 *contains some information about the simple null hypothesis.*

Example 2.5 tells us that the relevant sample is useful for testing. In this section, we will consider the information of the relevant sample for discrimination. We will use an idea similar to that underlying Kullback-Leibler information.

As we know, Kullback and Leibler (1951) define the logarithm of the likelihood ratio, $\log\{f_1(x)/f_2(x)\}$, as the information in $X = x$ for discrimination in favor of H_1 against H_2. Now we suppose that the sample X is from some density distribution $g_1(x)$ which may have some relationship to $\{f_1(x), f_2(x)\}$ (X is a relevant sample).

Definition 2.4 *The mean information for discrimination in favor of H_1 against H_2 per observation from $g_1(x)$ is defined as*

$$I(1:2;X) = \int g_1(x) \log \frac{f_1(x)}{f_2(x)} dx. \square \tag{2.9}$$

This definition is a departure from Kullback and Leibler information; Kullback and Leibler information is about the sample from $f_1(x)$. We generalize this to the relevant sample case. If $I(1:2;X) > 0$, then the sample from $g_1(x)$ favors H_1.

We can easily see that

$$I(1:2;X) = \int g_1(x) \log \frac{g_1(x)}{f_2(x)} dx - \int g_1(x) \log \frac{g_1(x)}{f_1(x)} dx. \tag{2.10}$$

The term $\int g_1(x) \log(g_1(x)/f_1(x))dx$ is the bias part of this information. If $g_1(x) = f_1(x)$, then the bias part vanishes.

Now we discuss the properties of this information.

- **a. $Inf(\theta, \theta_o)$ is independent of the σ-measure μ.**

  This is similar to the property a) in Section 2.4.

- **b. The information could be negative, and the negative value means this sample favors H_2.**

- **c. The information for discrimination in independent observations is additive.**

  This means that if we have some independent observations $X_1, \cdots, X_k$ from densities $g_1(x), \cdots, g_k(x)$ and let $X = (X_1, \cdots, X_k)$; then

$$I(1:2;X) = I(1:2;X_1) + \cdots + I(1:2;X_k).$$

  The proof is easy and we omit it here.

2.6 General Remarks

As suggested in Section 2.2, there are three types of information. What is the relationship among these types of information? The relevant sample information contains two parts: one is the relation between the two experiments (models); another is the observations. The relation between two experiments is a kind of direct information with an unknown parameter. When the number of relevant samples goes to infinity, the relevant sample information becomes direct information. For example, in Example 2.2, the similarity of θ_A and θ_B indicates a relationship between the two experiments. When m (spins of the penny B) goes to infinity, we can get an exact value of θ_B. Then the relevant sample information would mean θ_A is close to some known value (this is direct information).

The relation between θ_A and θ_B can be of several types. Here we list some of them: (1) $|\theta_A - \theta_B| \leq c$ for some constant c (Example 3.1); (2) $\theta_A - \theta_B$ is small (Example 2.2 and Example 3.2); and (3) $\theta_A - \theta_B$ is a random variable with a known distribution.

The information measures proposed in Section 2.4 and 2.5 need further study.

Chapter 3

Relevance Weighted Likelihood Estimation

3.1 Introduction

In this chapter we generalize the classical likelihood as the *Relevance Weighted Likelihood* (REWL). The REWL arises in parametric inference when in addition to (or instead of) the sample from the study population, relevant but independent samples from other populations are available. By down-weighting them according to their relevance, the REWL incorporates the information from these other samples. We have characterized such "relevant " sample information in Chapter 2. To motivate the likelihood theory presented below, we merely illustrate situations where such information arises.

Example 3.1 *Let* $Y_i \sim N(\mu_i, 1)$, $i = 1, 2$, *be independent random variables, the* $\{\mu_i\}$ *being unknown parameters. We want to estimate* μ_1. *Two estimators present themselves. (i) Classical likelihood-based estimation theory suggests the MLE* μ_1 *which uses just* Y_1.

Then we get

$$\hat{\mu}_1 = Y_1. \tag{3.1}$$

(ii) However, if μ_2 was deemed to be "close" to μ_1, intuition suggests we use the information in Y_2 in some way. Yet the classical theory still yields the result in (i). Even when we add the structural condition that $|\mu_1 - \mu_2| \leq c$ for a specified constant $c > 0$, the MLE still uses just Y_1 unless the condition $|Y_1 - Y_2| \leq c$ is violated. If that condition fails, $\hat{\mu}_1 = \{Y_1 + (Y_2 - c)\}/2$ or $\hat{\mu}_1 = \{Y_1 + (Y_2 + c)\}/2$ according as $Y_1 < (Y_2 - c)$ or $Y_1 > (Y_2 + c)$. So then the MLE does bring Y_2 into the estimation of μ_1. But it does so crudely only through truncation. So instead we turn to a seemingly more natural alternative which uses Y_2 more fully:

$$\hat{\mu}_1^* = \frac{1 + c^2}{2 + c^2} Y_1 + \frac{1}{2 + c^2} Y_2. \tag{3.2}$$

Under the mean squared error criterion, we find

$$E(\hat{\mu}_1 - \mu_1)^2 = 1 > \frac{1 + c^2}{2 + c^2} \geq E(\hat{\mu}_1^* - \mu_1)^2. \tag{3.3}$$

From (3.3), we can conclude that the estimator (3.2) based on both Y_1 and Y_2 always has a smaller mean squared error than that (3.1) based on Y_1 alone.

The last example demonstrates that in certain cases, we can profitably incorporate the information from samples drawn out of populations different from that under study. In other words, all "relevant" information must be used in inferences about the parameters of interest.

Example 3.2 Nonparametric Regression Model. *If n data points $\{(X_i, Y_i)\}_1^n$ have been collected, the regression relationship between Y and X can be modeled as*

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \cdots, n. \tag{3.4}$$

using the unknown regression function m and observation errors ϵ_i. Assume that $\epsilon_1, \cdots, \epsilon_n$ are iid from some unknown density function $f(x)$ with $E(\epsilon_1) = 0$. The function m are the parameters we want to estimate.

About the situation embraced by this last example, Eubank,R.(1988, p.7) says "If m is believed to be smooth, then the observations at X_i near x should contain information about the value of m at x. Thus it should be possible to use something like a local average of the data near x to construct an estimator of $m(x)$." Reasoning like this and the last example itself motivates our work. Developments of recent years in the theory and application of nonparametric regression validate Eubank's argument. These developments show nonparametric regression to be a useful explanatory and diagnostic tool. Eubank (1988), Hardle (1990) and Muller (1988) discuss nonparametric regression where observations at x_i near x are used to infer m at x because they contain relevant information.

However, the domain encompassed by the heuristics of nonparametric regression theory appears much broader than that currently encompassed by that theory. In fact, Example 3.2 immediately suggests a number of unanswered questions. (i) If the error-distribution associated with (3.4) had a known (parametric) form with unspecified parameters, how would all "relevant" information be used in inference about $m(x)$? More specifically, is there a likelihood based approach which would permit the use of that information?

(ii) If we were interested in estimating some unknown x- population attribute other than its mean, $m(x) = E(Y|X = x)$, what could we do in either the parametric case of (i) or more generally in the nonparametric case?

(iii) How can the information about $m(x)$ in the observations at x_i near x be described or quantified?

We have explored possible answers to question (iii) in Chapter 2. We will address question (ii) for the nonparametric case in Chapter 6; a solution for the parametric case implicitly derives from the theory of this chapter. Finally, the method of this chapter gives an answer to question (i) (see details in Chapter 5).

In this Chapter, we propose the idea of the relevance weighted likelihood (REWL) for the relevant sample situations. In Section 3.2, we construct the REWL function. After looking at several examples, we discuss there the relationship between the usual and relevance weighted likelihood function.

For the traditional purpose of data reduction, we define "weakly" sufficient statistics using the REWL in Section 3.3 (weakly because of their dependence on the relevance weights chosen in the construction of the REWL). We show that weakly sufficient statistics have some of the properties of their sufficient relatives.

In Section 3.4, we introduce the MREWLE, the *maximum relevance weighted likelihood estimator.* The MREWLE's obtained in some specific applications are new. In others, the theory merely enables us to rederive known estimators albeit from a more basic starting point.

We can justify the idea of the REWL by appealing in entropy and prediction much as in the classical case. This we do in Section 3.5. We first extend there the entropy-maximization principle to embrace all relevant samples. The resulting extension then yields the relevance-weighted likelihood principle.

We use relevance weighted likelihood under normal theory assumptions in Section 3.6. We show that maximizing the REWL yields a reasonable estimator. In this example, the MREWLE has advantage over MLE which is available in some special cases. The

application of the MREWL estimation to generalized smoothing models appears in Chapter 5. Some general remarks are made in Section 3.7.

3.2 The Relevance Weighted Likelihood

3.2.1 Definition.

Let $y = (y_1, \ldots, y_n)$ denote a realization of the random vector $Y = (Y_1, \ldots, Y_n)$ and $f_i \epsilon \mathcal{F}_i$, $i = i, \cdots, n$, the unknown probability density functions (PDF) of the Y_i which are assumed to be independent. We are interested in the PDF $f \epsilon \mathcal{F}$ of a study variable X measurable on items of our study population (with PDF f). At least in some qualitative sense, the f_i are thought to be "like" f. Consequently the y_i's are thought to be of value in our inferential analysis even though the Y_i are independently drawn from a population different from our study population.

In the familiar paradigm of repeated sampling, we impose the condition $f_i = f$ for all i in deriving the likelihood. In reality, this condition represents an approximation which may be more plausible for some of the i's than others. It may even seem desirable to downweight certain of the likelihood components, $f(y_i)$ in some way, when the quality of the associated approximation seems low. But how should those components be weighted?

A heuristic Bayesian analysis suggests a way of assigning relative weights to the likelihood components. This analysis leads us to the REWL. Suppose we take logarithms of the various PDF's (assumed to be positive for these heuristics) to put them on an affine scale: $(-\infty, \infty)$. In the log-likelihood, the correct term associated with y_i is $\log[f_i(y_i)]$. If we are to replace this with a term involving only $\log[f(y_i)]$, we might plausibly use

the best linear predictor (BLP) of $\log[f_i(y_i)]$ based on $\log[f(y_i)]$. This gives us $f(y_i)^{p_i}$ in place of $f_i(y_i)$ in the likelihood. Here p_i represents the coefficient of $\log f(y_i)$ in the BLP. In other words, p_i represents the covariance between $\log f_i(y_i)$ and $\log f(y_i)$ (if we ignore a multiplicative rescaling factor). [Our analysis also ignores an irrelevant additive factor in the BLP.]

This leads us to define the REWL at $y = (y_1, \cdots, y_n)$:

$$Wlik\{f(\cdot), y\} = \prod_1^n f(y_i)^{p_i}, \ for \ f \in \mathcal{F}. \tag{3.5}$$

The REWL, like the classical likelihood, allows the data to jointly assess the credibility of any hypothesized candidate f for the role of the study population PDF. But here the y_i's from the study population itself would be given the greatest weight in this joint assessment. As the relevance of the other y_i's decline in their relevance (measured by their p_i's) so does the weight accorded to them in the assessment.

Usually it is convenient to work with the natural logarithm, denoted by $Wl\{f(\cdot), y\}$ and called the *log relevance weighted likelihood*:

$$Wl\{f(\cdot), y\} = \sum_1^n p_i \log f(y_i), \ for \ f \in \mathcal{F}. \tag{3.6}$$

Conventionally we take $\mathcal{F}_i = \mathcal{F}$ for all i and index $\mathcal{F}$ by a finite dimensional (unknown) parameter $\theta = (\theta_1, \ ..., \ \theta_q) \in \Omega$. Then $f(t) = f(t;\theta)$, $f(\cdot;\cdot)$ having a known form. Then (3.5) and (3.6) become for $\theta \in \Omega$,

$$Wlik\{\theta, y\} = \prod_1^n f^{p_i}(y_i;\theta), \quad Wl\{\theta, y\} = \sum_1^n p_i \log f(y_i;\theta). \tag{3.7}$$

3.2.2 Examples.

The following examples illustrate the REWL and reveal differences between likelihood and the REWL.

Example 3.3 Continuation of Example 3.1. *The usual likelihood function in this problem would be*

$$lik(\mu; y) = (2\pi)^{-1} \exp(-[(y_1 - \mu_1)^2 + (y_2 - \mu_2)^2]/2). \tag{3.8}$$

This likelihood would ignore prior information like $|\mu_1 - \mu_2|$ is "small". Now define the REWL by putting $p_1 = 1$ and $p_2 = (1 + c^2)^{-1}$ as the relevance weights when inference is about the study population parameter μ_1. Then

$$Wlik(\mu_1; y) = (2\pi)^{-1} \exp(-\frac{(y_1 - \mu_1)^2}{2})((2\pi)^{-1} \exp(-\frac{(y_2 - \mu_1)^2}{2}))^{\frac{1}{1+c^2}} \tag{3.9}$$

The likelihood (3.8) contains the parameters for both of the populations from which the data were obtained. None of the information from Y_2 would be used to estimate μ_1 even when that information was deemed relevant. In contrast, the REWL in (3.9) contains only the parameter of inferential interest, μ_1. And Y_2 would be used in estimating μ_1 to the extent determined by the size of c.

Example 3.4 Continuation of Example 3.2. *Here define the likelihood to be*

$$\prod_1^n f(y_i - m(x_i)). \tag{3.10}$$

We find the result of no use in estimating $m(x)$. Yet as we argued earlier, if $m(\cdot)$ were thought to be smooth, observations near x should contain information about the PDF, $f(y - m(x))$. The REWL can reflect this heuristic through the relevance weights p_i in

$$\prod_1^n f^{p_i}(y_i - m(x)). \tag{3.11}$$

The next example differs from the last two in that we allow the relevance weights to depend on the data themselves.

Example 3.5 Robustness. *Assume $Y_1, \cdots, Y_n$ are iid observations with PDF $f(y, \theta)$ parametrized by θ so that the likelihood is*

$$\prod_1^n f(y_i, \theta).$$

We may believe some of the y_i to be outliers effectively coming from some other population than that under study. The information from such data needs to be downweighted. To do this we may: (i) order the data as $y_{(1)}, \cdots, y_{(n)}$; and (ii) assign relevance weights p_i depending on the degree to which we regard the associated data as outlying. The REWL becomes

$$\prod_1^n f^{p_i}(y_{(i)}, \theta). \tag{3.12}$$

In the extreme case, when a fraction 2ϵ are deemed to be outliers, we could choose $p_i = 0$, when $i \leq [n\epsilon]$ or $i \geq [n(1-\epsilon)]$. [Here $[\cdot]$ denotes the greatest integer less than x.] Then the REWL becomes a "trimmed likelihood." The trimmed mean would then be obtained in certain cases as a MREWLE. This case will be discussed again in Section 3.4.

In "parametric-nonparametric regression" we would postulate PDF's in parametric form with a smoothly varying regression (mean) function. Yet other parameters like quantiles and variances for example may also be of interest in this setting. The following general approach to smoothing through the REWL enables us to deal with this diversity of possible inferential objectives within a unified framework.

Example 3.6 Generalized Smoothing *Suppose $\{Y_i, X_i\}_1^n$ are n data pairs, for given X_i, Y_i has PDF $f\{y, \theta(X_i)\}$ with parameter $\theta(X_i)$. Interest lies in the study population*

corresponding to a fixed value $X = x$ and fitting the associated PDF. The relevance weights p_i enable us to represent the degree to which the information from the populations corresponding to X_i should be used in fitting $f\{y, \theta(x)\}$. The REWL becomes

$$\prod_1^n f^{p_i}\{y_i, \theta(x_i)\}. \tag{3.13}$$

Generally choosing the $\{p_i\}$ will be like choosing a kernel and bandwidth in nonparametric regression theory. Indeed, in the domain of that theory, we can find the $\{p_i\}$ directly from the corresponding kernels [and their bandwidths] making our task easy in that case.

3.2.3 Remarks

Here we discuss the relationship between the likelihood and the REWL. Then we consider some properties of the MLE retained by the REWL.

3.2.1 The likelihood is obtained from the REWL as a special case when all the data are independently drawn from the study population. However, even here there may be a role for the REWL as Example 3.5 demonstrates.

2.2 The likelihood is usually derived from the sampling density by inverting what is fixed and what is varying. In particular, conditional on f (or the parameter of f), the likelihood integrates over the sample space to 1. This duality between likelihood and sampling density may be useful for determining the likelihood. However, it does not seem intrinsic. We could in the usual case of iid sampling, take the *nth* root of the inverted sampling density without apparent loss and without preserving the aforementioned property. Moreover, in the Bayesian framework the

sample space need not even be specified, once the data have been obtained. Yet the likelihood can certainly be defined.

So we do not see the lack of duality with sampling as a shortcoming of our proposed extension of the likelihood. The usual asymptotic theory, appropriately modified still obtains as shown in the next Chapter.

3.2.3 The REWL depends on the relevance weights and work remains to be done on how these may be chosen. As noted above, in the case of nonparametric regression, standard theory for kernel smoothers suggests reasonable possibilities.

3.2.4 We can easily show that the REWL is preserved under arbitrary differentiable data-transformations (with non- vanishing Jacobian) when the sampling densities are absolutely continuous with respect to Lebesgue measure. So the REWL inherits this important property of the likelihood.

3.3 Weakly Sufficient Statistics

3.3.1 Definition.

The very important likelihood principle of classical statistical inference tells us that the likelihood embraces all relevant information about the parameter. Indeed, according to the factorization theorem sufficiency may be defined through the likelihood. Standard constructions of minimally sufficient statistics rely on the likelihood.

The counterpart of the likelihood theory in our setting would be the REWL principle. Lacking the invertibility of likelihood and sampling density found in standard frequency based theory of the likelihood, we must resort to a REWL based definition of sufficiency.

Lacking a basis for claiming our likelihood captures all relevant information in the data, we call our notion of sufficiency "weak sufficiency". That notion enables us to reduce the dimension of the observation vector to that of any (weakly) sufficient vector-valued function of the data, while retaining all information in the REWL.

Definition. We call *weakly sufficient* any vector- valued statistic which determines the REWL up to an arbitrary multiplicative factor which does not depend on f. *Weakly minimal-sufficient statistics* are functions of every other weakly sufficient statistic.

A weakly minimal-sufficient statistic yields the maximal data reduction. Such a statistic need not be unique. The *factorization theorem* remains true for weak sufficiency.

Theorem 3.1 *A necessary and sufficient condition that S be weakly sufficient for the parametric family, $\mathcal{F}$, indexed by θ is that there exist functions $m_1(s,\theta)$ and $m_2(y)$ such that for all $\theta \in \Omega$,*

$$Wlik(\theta, y) = m_1(s,\theta)m_2(y). \tag{3.14}$$

Accepting the REWL as the basis for inference makes reliance on weakly sufficient statistics inevitable. The seemingly reasonable estimators we obtain below depend on the data only through a weakly sufficient statistic, thereby offering "empirical support" for our principle. Just as the conventional likelihood (regarded as a function) is sufficient, the REWL is weakly sufficient. [This fact follows from the factorization theorem.] However, weakly sufficient statistics lack the property of conventional sufficiency which renders the conditional sampling distribution of the data given a sufficient statistic independent of θ. [Our REWL does not derive from a sampling density function.]

3.4 Maximum Relevance Weighted Likelihood Estimation (MREWLE)

In this section we generalize the MLE.

Definition: Call any $\hat{\theta} \in \Omega$ which maximizes the REWL, a *maximum REWL estimator* (MREWLE).

Before discussing the properties of MREWLE, we reconsider one of our examples.

Continuation of Example 3.5. *Assume the density $f(y|\theta)$ is that of a normal distribution with mean θ. Then the MREWLE of θ is*

$$\hat{\theta} = [n - 2[n\epsilon]]^{-1} \sum_{[n\epsilon]}^{[n(1-\epsilon)]} Y_{(i)},$$

when we choose the $\{p_i\}$ in Example 3.5. This is a trimmed mean. Other choices yield L-statistics as the MREWLE's.

The MREWLE inherits some of the properties of the MLE.

- Under certain weak conditions, the MREWLE is consistent. We will prove this fact in the next Chapter by generalizing to the non-iid and more general weighted case the well-known theory of Wald (1949).
- The asymptotic normality of MREWLE under certain conditions is proved in Chapter 4.
- The MREWLE always relies on the data only through weakly sufficient statistics.
- the MREWLE possesses the familiar property of invariance under one-to-one parameter transformations.

- The goal of establishing the asymptotic efficiency of MREWLE in some appropriate general sense has eluded us. At this time we can give that property only in the special case of nonparametric regression (Chapter 5 and Chapter 7).

3.5 The Entropy Maximization Principle.

In a series of papers, Akaike (1977, 1978, 1983,1985) discusses the importance of the entropy maximization principle in unifying conventional and Bayesian statistics. We generalize this principle to the framework of relevant samples in this section. This generalization enables us to prove that the method of MREWL may be viewed as a realization of that principle to an important but limited extent.

3.5.1 The generalized entropy maximization principle.

To recall the conventional entropy maximization principle, suppose we draw $x = (x_1, x_2, \cdots, x_k)'$ from a multivariate distribution with density f. Suppose we intend to estimate f by $g(\cdot : x)$ and view this estimate as a predictive distribution for a future vector drawn from f. As the index of the quality of $g(\cdot : x)$, use the entropy of f with respect to $g(\cdot : x)$:

$$B(f;g) = -\int f(z) \log \frac{f(z)}{g(z:x)} dz.$$

The entropy maximization principle asks us to find the $g(\cdot : x)$ which maximizes the expected entropy $E_x B(f;g) = \int B(f;g) f(x) dx$. We may view the result as giving us an "optimum" estimator of f, regarded as the object of inferential interest. We would note in passing that Fisher's maximum likelihood method and the AIC (Akaike information criterion) are two very important implications of the entropy maximization principle.

For simplicity of exposition, consider now only the univariate case. [The vector variable case is an obvious generalization.] Suppose $y_1, \cdots, y_n$ respectively, are independently drawn from distributions with densities $f_1(y), \cdots, f_n(y)$ thought to be related to the density of inferential interest f. [The relevance of f to f_i could be described by $B(f_i; f) > -c_i$ for all i where the $\{c_i\}$ are positive constants. This inequality means f is not far from f_i. For the special iid case, $f = f_i$ for all i or equivalently, $B(f_i; f) = 0$ for all i.] Let $g(\cdot : y)$ denote an estimate of f where $y = (y_1, \cdots, y_n)$. Once again we may view g as an estimated predictive distribution of a future observation z from f.

Because the relevance of the f_i to f varies with i, we assign different relevance weights, p_i, to them. We then get the weighted entropy measure:

$$\sum_1^n p_i B(f_i; g) = -\sum_1^n p_i \int f_i(z) \log \frac{f_i(z)}{g(z : y)} dz.$$

[Because we do not know f, we choose the above index to force g to lie "close" to the densities we do know, $\{f_i\}$, and which we deem to be close to f.]

Our generalized entropy maximization principle may now be stated. *All inference about f may be based on the g obtained by maximizing the expected weighted entropy of the predictive distribution where the expected weighted entropy is*

$$\sum_1^n E_{y_i} p_i B(f_i; g) = -\sum_1^n \int p_i B(f_i; g) f_i(y) dy.$$

3.5.2 The MREWLE and generalized entropy maximization principle.

We know that

$$\sum_1^n p_i B(f_i; g) = \sum_1^n p_i \int f_i(z) \log g(z : y) dz - \sum_1^n p_i \int f_i(z) \log f_i(z) dz.$$

The second term on the right, a constant, depends on only $\{f_i\}$. For assessing g we need only consider the first term. However, we cannot evaluate that unknown term so estimate it by $\sum_1^n p_i \log g(y_i : y)$. This amount uses what in Chapter 6 we call the Relevance Weighted Empirical Distribution function which puts mass p_i at y_i for all i. If we specify a family of feasible $g(\cdot : y)$'s the one which maximizes the estimated expected weighted entropy at y defines the maximum REWL estimate of f. Obviously the performance of the MREWLE of f depends on both the choice of the feasible family and the statistical characteristics of the simple and natural estimator. If for the feasible family we choose the parametric family of the $f(y|\theta)$, then we find that the estimate obtained from the generalized entropy maximization principle is just the MREWLE.

For brevity we will not pursue further our discussion of the generalized entropy maximization principle. However many questions about the generalized entropy maximization principle remain to be answered.

3.6 MREWLE for Normal Observations

In this section, we develop a method of estimating the mean of a normal population using data from relevant samples, thereby extending Example 1 above. We use the MREWLE and compare it with other estimators.

Let Y and Y_1 be observations from normal populations with known variances and unknown means θ and θ_1, respectively. Without essential loss of generality, suppose $Var(Y) = 1$ and $Var(Y_1) = \sigma_1^2$. Assume $\theta - \theta_1 \in [-c, c]$ for some fixed $c > 0$, θ being the parameter of interest. We readily find the MREWLE of θ to be

$$\hat{\theta} = \frac{c^2 + \sigma_1^2}{1 + c^2 + \sigma_1^2} Y + \frac{1}{1 + c^2 + \sigma_1^2} Y_1, \tag{3.15}$$

if we choose the relevance weights $p_1 = \frac{c^2+\sigma_1^2}{1+c^2+\sigma_1^2}$ and $p_2 = \frac{1}{1+c^2+\sigma_1^2}$ for Y and Y_1, respectively. Here we choose the relevance weights by minimaxing the mean square error of MREWLE.

Now we compare the MREWLE with the maximum likelihood estimator.

In agreement with intuition, we find that the MREWLE loses the advantage over the MLE as $\sigma_1^2 \to \infty$ or $c \to \infty$. The extra information in Y_1 becomes useless in these extreme cases because of the uncontrolled bias or noise in the second sample. When $c = 0$, the MREWLE becomes the MLE for the full data set. In all these cases, the MREWLE is the minimax estimator.

If $\sigma_1^2 \to 0$, then the problem under consideration becomes that of estimating a bounded normal mean. However, the MREWLE differs from the MLE. Without loss of generality, assume $\theta_1 = 0$. From (3.15), the MREWLE of θ is

$$\hat{\theta} = \frac{c^2}{1+c^2} Y.$$

and the MLE is

$$\hat{\theta}(MLE) = \begin{cases} -c, & \text{if } Y < -c \\ Y, & \text{if } -c \le Y \le c \\ c, & \text{if } c < Y. \end{cases}$$

The mean square error for these two estimators are

$$\max_{\theta} E(\hat{\theta} - \theta)^2 = c^2/(1+c^2),$$

and

$$E(\hat{\theta}(MLE) - \theta)^2 = 2\Phi(c)(1-c^2) + 2c^2 - 1 + 2c\exp(-c^2/2)/\sqrt{2\pi}$$

for only $\theta = \theta_1$. There is no closed form of the maximum mean square error for the MLE over all the parameter space. We compare them in the following Figure 3.1.

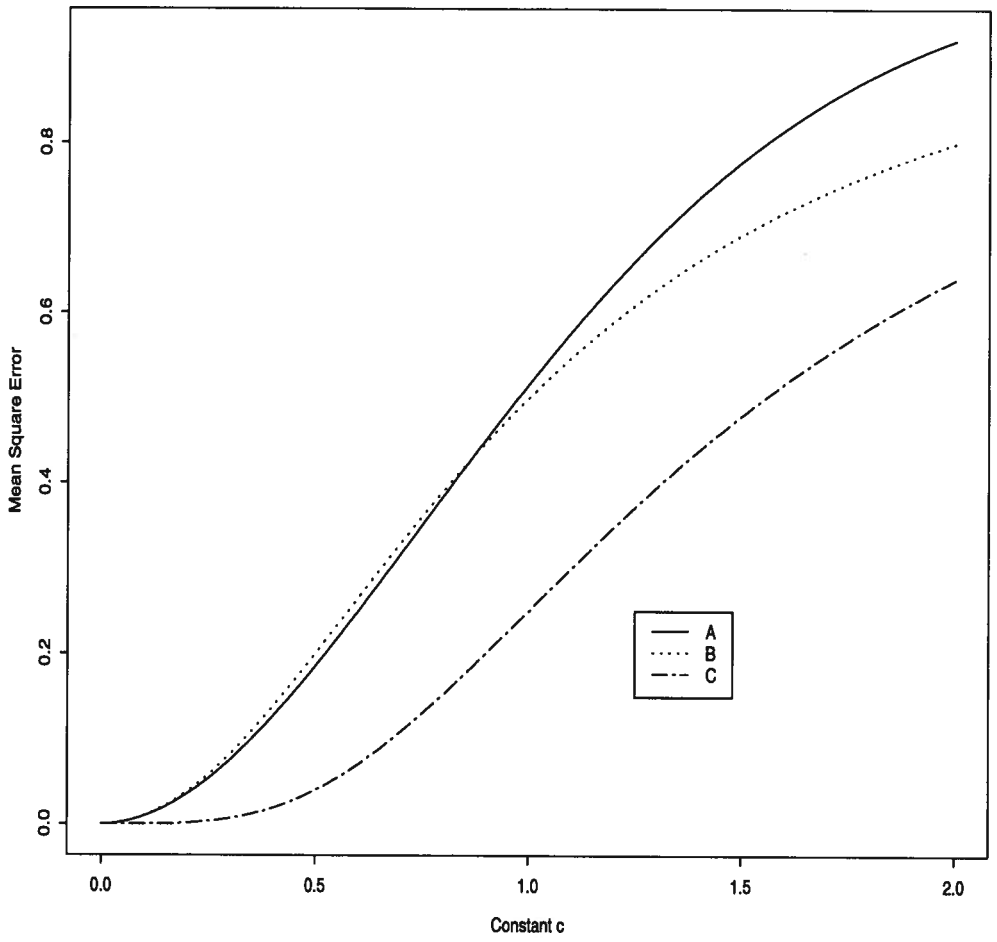

Figure 3.1: *A comparison of MREWLE and MLE under* $\sigma_1 = 0$. *Curve A represents the MSE of MLE at line* $\theta = \theta_1$. *Curve B represents the Maximum MSE of MREWLE over whole parameter space. Curve C represents the MSE of MREWLE at line* $\theta = \theta_1$.

Assume $\sigma_1^2 = 1$. From (3.15), the MREWLE of θ is

$$\hat{\theta} = \frac{1+c^2}{2+c^2}Y + \frac{1}{2+c^2}Y_1,$$

and the MLE becomes

$$\hat{\theta}(MLE) = \begin{cases} (Y+Y_1+c)/2, & \text{if } Y_1 < Y-c \\ Y, & \text{if } Y-c \le Y_1 \le Y+c \\ (Y+Y_1-c)/2, & \text{if } Y_1 > Y+c. \end{cases}$$

The comparison of the MREWLE with the MLE is showing in the Figure 3.2.

From the above comparison, we find that the MREWLE has the advantage over the MLE for the two normal relevant samples, when the mean square error criterion is used.

With several relevant samples for θ with differing means, the analytical calculation of the MLE of θ proves nearly impossible. By choosing relevance weights, we can easily calculate the MREWLE.

3.7 General Remarks

In Example 3.5 of Section 3.2, we show that we can use the REWL for robustness. This idea is similar to the work of weighted partial likelihood by Sasieni (1992) for the Cox model (exact sample case). In that paper, he considers robustness and efficiency of the weighted partial likelihood method. So even for the iid sample case, we may used the REWL for both robustness and efficiency.

The weak sufficient statistics defined in Section 3.3 depend on the relevance weights. This agrees with our intuition. For different relevance weights (i.e. different views of relative importances), the weak sufficient statistics should be different.

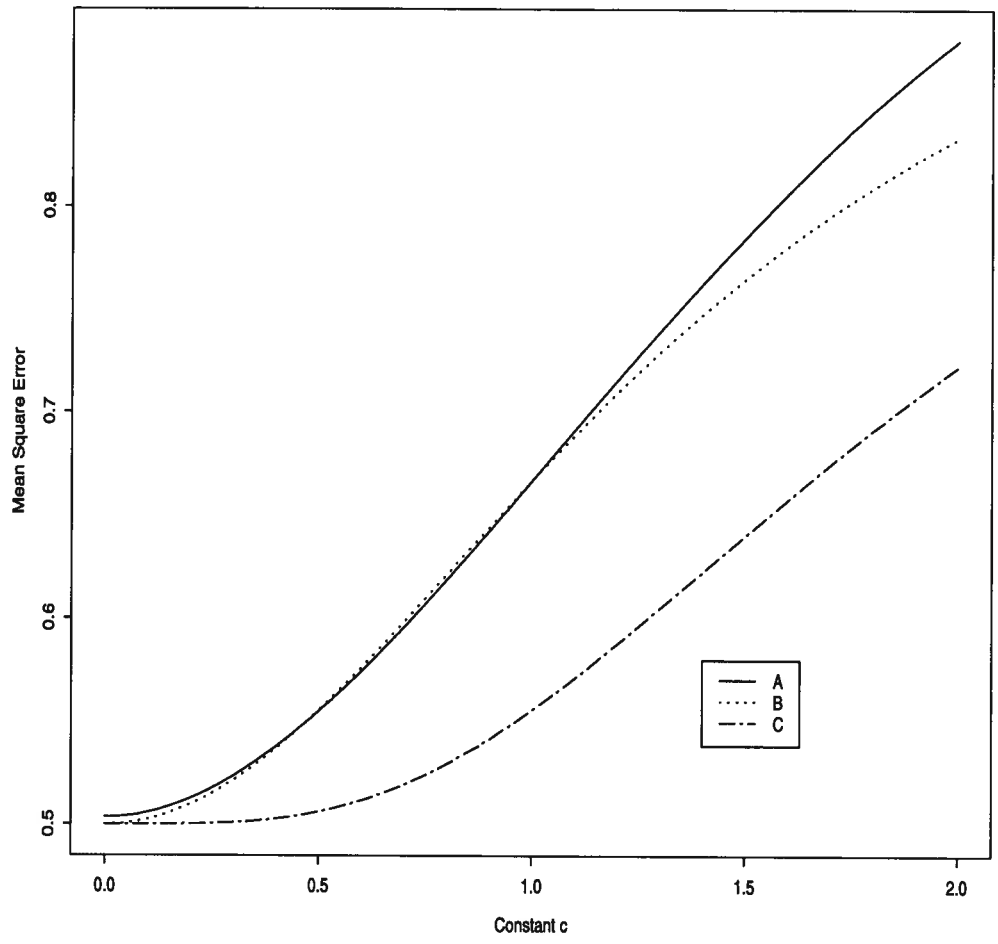

Figure 3.2: *A comparison of MREWLE with MLE under* $\sigma_1 = 1$. *Curve A represents the MSE of MLE at line* $\theta = \theta_1$. *Curve B represents the Maximum MSE of MREWLE over whole parameter space. Curve C represents the MSE of MREWLE at line* $\theta = \theta_1$.

The REWL proposed in this Chapter depends on the relevance weights, $\{p_i\}$. These weights express the statistician's perceived relationships among the populations and usually can be chosen on intuitive grounds. For different problems, the relevance weights were chosen in different ways. In Section 3.6, we choose these weights by minimaxing the mean square error. In the Example 3.5, we may choose these weights by considering both robustness and efficiency. For the generalized smoothing model, we will used the weights similar to these of nonparametric regressions. This is in Chapter 5.

The asymptotic theory of REWL in the next Chapter also give a guide line for the choice of relevance weights. In the following Chapters, we will further discuss the choice of these weights.

Chapter 4

The Asymptotic Properties of the Maximum Relevance Weighted Likelihood Estimator

4.1 Introduction

In Chapter 3, we gave a very general method for using relevant sample information in statistical inference. We base the theory what we call the *relevance weighted likelihood* (REWL). The REWL function plays the same role in the case of relevance sample information as the likelihood in that of exact sample information.

The maximum relevance weighted likelihood estimate (MREWLE) studied in this paper, plays the role of the maximum likelihood estimator (MLE) in conventional point estimation. The consistency of the MLE has been investigated by several authors (*c.f.* Cramer 1946 and Wald 1948). Here we prove the consistency of MREWLE under general conditions. But in many cases, the consistency of the MREWLE is not enough; we need to get the asymptotic distribution of the MREWLE. The asymptotic normality of

the MREWLE is considered in this Chapter.

I first consider the weak consistency of the MREWLE and show: (i) there exists a weakly consistent sequence of roots for the log REWL equation (Theorem 4.2); (ii) the MREWLE is weakly consistent (Theorem 4.4). I then go on to the strong consistency of the MREWLE (Theorem 4.5). My analysis relies heavily on the work of Chow and Lai (1973) as well as Stout (1968) who deal with the almost sure behavior of weighted sums of independent random variables. Finally, we prove the asymptotic normality of the MREWLE (Theorem 4.7).

I organize the paper as follows. The main results on weak and strong consistency are stated in Sections 4.2 and 4.3, respectively. We state the asymptotic normality in Section 4.4. [The proofs of these theorem are in Section 4.7]. Section 4.5 proposes two estimators of the variance of the MREWLE. In Section 4.6, I discuss possible extensions and make some concluding remarks.

4.2 Weak Consistency

In Chapter 3, we have defined the REWL and the MREWLE. In this Chapter, we will treat only the parametric case so that interest focuses on a single parameter θ for simplicity.

Let $X_1, \ldots, X_n$ be random variables with probability density functions (PDF's) $f_1, f_2, \ldots, f_n$. We are interested in the PDF $f(x, \theta) : \theta \in \Omega$ of a study variable X. θ is an unknown parameter. To state the Theorem, we begin with the following assumptions.

Assumptions 4.1 **4.1.1** $\{P_\theta : \theta \in \Omega\}$ *represents a family of distinct distributions*

with common support and dominating measure μ.

Let $f(x,\theta)$ denote the PDF of P_θ.

4.1.2 *The distributions of the independent sample observations, $X_i, i = 1,\ldots,n$, have the same support as the $\{P_\theta\}$.*

4.1.3 *The relevance weights p_{ni} corresponding to $X_i, i = 1,\ldots,n$, and incorporated in the vector $P_n = (p_{n1}, p_{n2}, ..., p_{nn})$ play a central role in our theory. They satisfy the formal requirements $p_{ni} \geq 0$ and $\sum_1^n p_{ni} = 1$. As well, with the "true" value of θ denoted by θ_0, we require that*

$$\sum_1^n p_{ni} E \log\{f_i(X_i)/f(X_i,\theta_0)\} \to 0 \text{ as } n \to \infty \tag{4.1}$$

and for any θ

$$\sum_1^n p_{ni}^2 Var[\log\{f_i(X_i)/f(X_i,\theta)\}] \to 0 \text{ as } n \to \infty. \tag{4.2}$$

4.1.4 *Ω contains an open interval Θ of which the true parameter θ_0 is an interior point.*

Let $\underline{X} = (X_1, X_2, ..., X_n)$. For fixed $\underline{X} = \underline{x}$, the function, $\theta \rightsquigarrow \prod_1^n f^{p_{ni}}(x_i,\theta)$ will be called the REWL function.

Theorem 4.1 *Assumptions 4.1.1-4.1.3 imply*

$$P_{\theta_0}\{f^{p_{n1}}(X_1,\theta_0)\ldots f^{p_{nn}}(X_n,\theta_0) > f^{p_{n1}}(X_1,\theta)\ldots f^{p_{nn}}(X_n,\theta)\} \to 1 \tag{4.3}$$

as $n \to \infty$ for any fixed $\theta \neq \theta_0$. □

From (4.3), the value of the REWL function at θ_0 (regarded as depending of $\underline{X}$) exceeds its value at any other fixed θ with high probability when n is large. We do not know θ_0, but we can determine the point, $\hat{\theta}$ called the MREWLE, at which the REWL function for fixed $\underline{X} = \underline{x}$ is maximized. Suppose the observations are from distributions with PDF's like that of the true sampling distribution $f(x, \theta_0)$ (and that Conditions (4.1) and (4.2) hold). Then the last theorem suggests that the MREWLE of θ should be close to the true value of θ if the REWL function of $\underline{x}$ varies smoothly with θ. Hence the MREWLE should be a reasonable estimator.

Remark A.

(i) Assumption 4.1.2 seems quite reasonable. If the distributions did not have the same support, we could not construct a useful REWL function. For example, if X_1 had support $[0, 2]$ and $f(x, \theta)$ had support $[0, 1]$, the REWL function would be identically 0 when X_1 was in $(1, 2]$.

(ii) The independence assumed in 4.1.2 greatly simplifies our problem which would otherwise be insurmountable.

(iii) Condition (4.1) underlies the construction of a useful REWL function. Recall that [the Kullback-Leibler (KL) functional] $E \log\{f_i(X_i)/f(X_i, \theta_0)\}$ measures the discrepancy between f_i and $f(\cdot, \theta_0)$. That condition insures that the weighted KL discrepancy of the observations converges to zero when the sample size grows large. When the PDF's of the observations are quite different from $f(x, \theta_0)$, we usually cannot get a good estimator of the true parameter. Our difficulty arises then because we do not get enough information about the unknown parameter from the observations.

This condition is easily satisfied as when $E \log\{f_i(X_i)/f(X_i, \theta_0)\}$ is uniformly bounded

while $\lim_{i\to\infty} E\log\{f_i(X_i)/f(X_i,\theta_0)\} \to 0$. For then P_n can easily be chosen to make (4.1) hold.

(iv) Condition (4.2) commonly holds as when $\max_i\{p_{ni}\} \to 0$ while $Var(\log(f_i(X_i)/f(X_i,\theta)))$ is uniformly bounded for each θ. This is because that

$$\sum p_{ni}^2 Var[\log\{f_i(X_i)/f(X_i,\theta)\}] \leq \sum p_{ni}^2 C(\theta) \leq \max_i\{p_{ni}\}C(\theta).$$

Here the $C(\theta)$ is a constant depend on θ. The first inequality follows from uniform boundedness and the second inequality because $\sum p_{ni} = 1$.

(v) Conditions (4.1) and (4.2) hold respectively, when $(X_1, X_2, ..., X_n)$ are independent and identically distributed with PDF $f(\cdot,\theta_0)$ and when $Var[\log\{f(X_1,\theta_0)/f(X_1,\theta)\}]$ exists while $\max_i\{p_{ni}\} \to 0$.

Corollary 4.1 *If Ω is finite, Assumptions 4.1.1-4.1.3 imply that the MREWLE $\hat{\theta}_n$: (i) exists; (ii) is unique with probability tending to 1 and (iii) is weakly consistent.*

Proof: The result follows immediately from the Theorem and the fact that $P(A_{1n}\bigcap\ldots\bigcap A_{kn}) \to 1$ as $n \to \infty$ if $P(A_{in}) \to 1$ for $i = 1,\ldots,k$. □

Theorem 4.2 *Suppose: (i) $\underline{X} = (X_1, X_2, ..., X_n)$ satisfies Assumptions 4.1; and (ii) for $\theta \in \Theta$ and almost all x, the function $\theta \rightsquigarrow f(x,\theta)$ is differentiable with derivative $f'(x,\theta)$. Then with probability tending to 1 as $n \to \infty$, the relevance weighted likelihood (REWL) equation*

$$\partial/\partial\theta\{\prod_1^n f^{p_{ni}}(x_i,\theta)\} = 0$$

has a root, $\hat{\theta}_n = \hat{\theta}_n(x_1,\ldots,x_n)$, which tends to θ_0. □

Note that the REWL equation in the last theorem may equivalently be stated as

$$WL'(\theta, \underline{x}) = \sum_{1}^{n} p_{ni} \frac{f'(x_i, \theta)}{f(x_i, \theta)} = 0. \tag{4.4}$$

The following comments also relate to Theorem 4.2.

1. Its proof shows incidentally that with probability tending to 1, the $\{\hat{\theta}_n\}$, can be chosen to be local maxima. Therefore we may take the θ_n^* to be the root closest to a maximum.

2. But the Theorem does not establish the existence of a consistent estimator sequence since, with the true value unknown, the data do not enable us to pick a specific consistent sequence.

3. Theorem 4.2 only gives us the existence of a consistent root of the REWL equation. But only in very special cases is this root the MREWLE, in which case it is then consistent (see Corollary 4.2 below)

4. To prove Theorem 4.2, we require that $\theta \rightsquigarrow f(x, \theta)$ be differentiable, $\theta \in \Theta$. We will give some conditions (similar to the conditions given by Wald (1949) for the iid case) which avoid the requirement that $\theta \rightsquigarrow f(x, \theta)$ be differentiable.

Corollary 4.2 *If the weighted likelihood equation has a unique root δ_n for each n and all $\underline{x}$ the hypotheses of Theorem 4.2 imply that $\{\delta_n\}$ is a consistent sequence of estimators of θ. If in addition, the parameter space is any open interval (a, b) then with probability tending to 1, δ_n maximizes the weighted likelihood, (δ_n is the MREWLE), and is therefore consistent.*

Proof: The first statement is obvious. To prove the second suppose the probability

of δ_n being MREWLE does not tend to 1. Then for sufficiently large n, the weighted likelihood must, with positive probability, tend to a supremum as θ tends toward a or b. Now with probability tending to 1, δ_n is a local maximum of the weighted likelihood, which must then also possess a local minimum. This contradicts the assumed uniqueness of the root. □

The conclusion of Corollary 4.2 holds when the probability of multiple roots tends to 0 as $n \to \infty$.

We have already discussed the consistency of a root of the REWL equation. Now we are going to study the consistency of the MREWLE.

Before formulating our assumptions, we introduce some notation. For any θ and $\rho, r > 0$ let: $f(x,\theta,\rho) = \sup\{f(x,\theta') : |\theta'-\theta| \le \rho\}$; $\varphi(x,r) = \sup\{f(x,\theta) : |\theta| > r\}$; $f^*(x,\theta,\rho) = f(x,\theta,\rho)$ *or* 1 according as $f(x,\theta,\rho) > 1$ *or* ≤ 1, respectively; $\varphi^*(x,r) = \varphi(x,r)$ *or* 1 according as $\varphi(x,r) > 1$ *or* ≤ 1, respectively.

Assumptions 4.2 **4.2.1** *For any θ and ρ, $x \rightsquigarrow f(x,\theta,\rho)$ is measurable.*

4.2.2 *For any $\theta, \theta_0 \in \Theta$, there exists a $\rho_{\theta,\theta_0} > 0$, such that if $0 < \rho < \rho_{\theta,\theta_0}$, the expected value of $\int_{-\infty}^{\infty} \log f^*(x,\theta,\rho) dF(x,\theta_0)$ is finite. Similarly, $\int_{-\infty}^{\infty} \log \varphi^*(x,r) dF(x,\theta_0) < \infty$ for sufficiently larger r (depending on θ_0); here $F(\cdot,\theta_0)$ represents the CDF for P_{θ_0}.*

4.2.3 *For any $\theta \in \Theta$, $\int_{-\infty}^{\infty} |\log f(x,\theta)| f(x,\theta) d\mu(x) < \infty$.*

4.2.4 *There exists A, a Borel set, such that for any $\theta \in \Theta$, $\int_A f(x,\theta) d\mu(x) = 0$ and for $x \in \bar{A}$, $\lim_{|\theta|\to\infty} f(x,\theta) = 0$.*

4.2.5 *If* $\theta_1 \neq \theta_2$ *then* $\mu(\{x : f(x,\theta_1) \neq f(x,\theta_2\}) > 0$.

4.2.6 *For any* $\theta \in \Theta$, *there exists* B_θ, *such that* $\int_{B_\theta} f(x,\theta_0)d\mu(x) = 0$ *for any* $\theta_0 \in \Theta$, *and for* $x \in \bar{B}_\theta$, $f(x,\theta') \to f(x,\theta)$ *for any* $\theta' \to \theta$.

4.2.7 *In the observation vector,* $\underline{X} = (X_1, X_2, ..., X_n)$, *the* X_i *are independent with PDF* $f_i(x)$ *with respect to the same dominating measure* μ.

4.2.8 *Let* $P_n = (p_{n1}, p_{n2}, ..., p_{nn})$ *denote the respective important weights satisfying* $p_{ni} \geq 0$ *and* $\sum_1^n p_{ni} = 1$. *Assume*

$$\sum_1^n p_{ni} E \log\{f_i(X_i)/f(X_i,\theta_0)\} \to 0 \text{ as } n \to \infty. \tag{4.5}$$

4.2.9 *Let* $g_n(x) = \sum_1^n p_{ni} f_i(x)$. *Assume there exists a Borel measurable function* G, *such that* $g_n(x) \leq G(x)$, $\int G(x)|\log f(x,\theta_0)|d\mu(x) < \infty$ *and* $\int G(x)|\log \varphi(x,r)|d\mu(x) < \infty$ *for sufficiently large* r.

4.2.10 *For each* θ *and* $\rho(\theta)$,

$$\sum_1^n p_{ni}^2 Var[\log\{f(X_i,\theta_0)/f(X_i,\theta,\rho(\theta))\}] \to 0 \quad as\ n \to \infty. \tag{4.6}$$

For each r,

$$\sum_1^n p_{ni}^2 Var[\log\{f(X_i,\theta_0)/\varphi(X_i,r)\}] \to 0 \quad as\ n \to \infty. \tag{4.7}$$

The Assumptions 4.2.1-4.2.6 are similar to Wald's assumptions for the i.i.d case. These assumptions insure the validity of the lemmas of Wald (1948). Assumption 4.2.8 is essential for constructing a useful REWL. Assumption 4.2.9 and Assumption 4.2.10 are for the Dominated Convergence Theorem and Weak Law of Large Numbers.

Theorem 4.3 *If $\theta_0 \notin \omega$, ω a closed subset of Θ, Assumptions 4.2 imply that for any $\epsilon > 0$,*

$$P_n[\{\frac{\sup_{\theta\in\omega} f^{p_{n1}}(X_1,\theta)\dots f^{p_{nn}}(X_n,\theta)}{f^{p_{n1}}(X_1,\theta_0)\dots f^{p_{nn}}(X_n,\theta_0)}\}^n \geq \epsilon] \to 0. \tag{4.8}$$

Here the probability P_n denotes that of the $\{X_i\}$ obtained from $\{f_i(x) : i = 1,\dots,n\}$. □

Theorem 4.4 *Suppose Assumptions 4.2 obtain. Let $\bar{\theta}_n(X_1,\dots,X_n)$ be any function of the observations, $X_1,\dots,X_n$, such that*

$$\{\frac{f^{p_{n1}}(X_1,\bar{\theta}_n)\dots f^{p_{nn}}(X_n,,\bar{\theta}_n)}{f^{p_{n1}}(X_1,\theta_0)\dots f^{p_{nn}}(X_n,\theta_0)}\}^n \geq c > 0 \text{ for all } n \text{ and all } X_1,\dots,X_n. \tag{4.9}$$

Then $\bar{\theta}_n$ is a weakly consistent estimator of θ_0. □

Observe that the MREWLE always satisfies the conditions of Theorem 4.4 if we choose $c = 1$. So we have proved the weak consistency of MREWLE.

4.3 Strong Consistency

For the strong consistency of MREWLE, we use almost sure results on linear combinations of independent random variables such as those of (see Chow and Lai (1973), Stout (1968)).

For simplicity, define: (i) $A_n = \sum_1^n p_{ni}^2$; (ii) $D_{ij} = \log f(X_i,\theta_0) - \log f(X_i,\theta_j,\rho(\theta_j)) - E(\log f(X_i,\theta_0) - \log f(X_i,\theta_j,\rho(\theta_j)))$ from (4.24) and the proof of Theorem 4.3; (iii) $D_i^* = \log f(X_i,\theta_0) - \log\varphi(X_i,r_o) - E(\log f(X_i,\theta_0) - \log\varphi(X_i,r_o))$ from (4.25) and the proof of Theorem 4.3, where $i = 1,\dots,n$ and $j = 1,\dots,k$. In the next theorem, D will generically denote all D_{ij} and D_i^*.

Theorem 4.5 *Suppose Assumptions 4.2.1-4.2.9 obtain, and there exists a constant $K > 0$, such that $p_{ni} \leq Kn^{-\alpha}$ for some $\alpha > 0$ and that one of the following conditions hold:*

(i) $E|D|^{(1+\alpha+\beta)/\alpha}(\log^+ |D|)^{1+\xi} \leq K$ *for some* $\xi > 0$ *and* β,

$$(1+\alpha+\beta)/\alpha \geq 2, \quad ED^2(\log^+ |D|)^{2+\xi} \leq K, \quad A_n \leq Kn^{\beta-\alpha};$$

(ii) $ED^2 \leq K$, $ED = 0$, *and* $\sum_1^n p_{ni}^{\delta} \leq Kn^{\alpha(2-\delta)-1-\xi}$ *for some* $0 < \delta < 2$ *and* $\xi > 0$;

(iii) $E|D|^{(1+\alpha+\beta)/\alpha}(\log^+ |D|)^{1+\xi} \leq K$ *for some* $\xi > 0$ *and* β,

$$1 \leq (1+\alpha+\beta)/\alpha < 2, \quad \sum_1^n p_{ni}^{(1+\alpha+\beta)/\alpha} \leq Kn^{-\gamma}, A_n \leq Kn^{\beta-\alpha}$$

for some $\gamma > 0$;

(iv) $E|D|^{(1+\alpha+\beta)/\alpha}(\log^+ |D|)^{1+\xi} \leq K$ *for some* $\xi > 0$ *and* β,

$$0 < (1+\alpha+\beta)/\alpha < 1, \quad \sum_1^n p_{ni}^{(1+\alpha+\beta)/\alpha} \leq Kn^{-\gamma}, A_n \leq Kn^{\beta-\alpha}$$

for some $\gamma > 0$ *and* $p_{ni} = 0$ *for* $i > n^{\xi}$ *where* $\xi < \gamma(1+\alpha+\beta)/\alpha$. *Then*

$$P[\lim_{n\to\infty} \{\frac{\sup_{\theta\in\omega} f^{p_{n1}}(X_1,\theta)\dots f^{p_{nn}}(X_n,\theta)}{f^{p_{n1}}(X_1,\theta_0)\dots f^{p_{nn}}(X_n,\theta_0)}\}^n = 0] = 1.\square \tag{4.10}$$

Theorem 4.6 *Under the conditions of Theorem 4.5, let $\bar{\theta}_n(X_1,\dots,X_n)$ be any function of the observations $X_1,\dots,X_n$ such that (4.9) holds. Then $\bar{\theta}_n$ is a strongly consistent estimator of θ_0.*

We now state without proof a direct corollary of the last Theorem.

Corollary 4.3 *Under the conditions of Theorem 4.5, the MREWLE is strongly consistent.*

The weak conditions of Theorem 4.5 are not easily verified. In contrast, the stronger conditions in the following corollaries are easy to verify (but the results are not then as general as those of the Theorem).

Corollary 4.4 *Under the assumptions of Theorem 4.4, let* $p_{ni} \leq Kn^{-\alpha}$ *for some* $\alpha > 0$. *If* $E|D|^{2/\alpha} \leq K$ *for some* $0 < \alpha < 1$ *then (4.10) holds.*

Proof: Let $\beta = 0$. Then (i) of Theorem 4.5 is satisfied. □

In the case of independently and identically distributed observations, the assumptions of Theorem 4.5 can be quite unrestrictive. We will not go into detail here because our observation follows immediately from Theorem 1 of Stout (1968).

4.4 The Asymptotic Normality of the MREWLE

In the last section, we have shown that under regularity conditions, the MREWLEs are consistent and strongly consistent. In this section, we shall show that the MREWLEs are asymptotically normal under some conditions.

As we know, the asymptotic normality of MLE have been discussed by Cramer (1946) among others. Our results generalize Cramer's to both non-i.i.d and unequal p_{ni}.

Assumptions 4.3 **4.3.1** *For each* $\theta \in \Theta$, *the derivatives*

$$\frac{\partial \log f(x,\theta)}{\partial \theta}, \frac{\partial^2 \log f(x,\theta)}{\partial \theta^2}, \frac{\partial^3 \log f(x,\theta)}{\partial \theta^3}$$

exist, all x.

4.3.2 *For each* $\theta_0 \in \Theta$, *there exist functions* g, h *and* H *(possibly depending on* θ_0*) such that for* θ *in a neighborhood* $N(\theta_0)$ *the relations*

$$|\frac{\partial f(x,\theta)}{\partial \theta}| \leq g(x), |\frac{\partial^2 f(x,\theta)}{\partial \theta^2}| \leq h(x), |\frac{\partial^3 \log f(x,\theta)}{\partial \theta^3}| \leq H(x)$$

hold, all x, *and*

$$\int g(x)d\mu(x) < \infty, \int h(x)d\mu(x) < \infty, E_{\theta_0}H(x) < \infty.$$

4.3.3 *For each* $\theta \in \Theta$

$$0 < I(\theta) = E_\theta[\{\frac{\partial \log f(x;\theta)}{\partial \theta}\}^2] < \infty.$$

4.3.4 $\max_i(p_{ni}^2)/\sum p_{ni}^2 \to 0$.

4.3.5 *Define* $b_i = \int\{\partial \log f(x;\theta_0)/\partial\theta_0\}f_i(x)d\mu(x)$ *as the biased functions for each observation. Then*

$$(\sum p_{ni}^2)^{-1}\sum p_{ni}^2 b_i^2 \to 0.$$

4.3.6 *Let* $g_n(x) = \sum p_{ni}f_i(x)$. *Assume there exists a Borel measurable function* G *such that* $g_n(x) \leq G(x)$,

$$\int G(x)\,|\,\frac{\partial \log f(x,\theta_0)}{\partial \theta_0}\,|\,d\mu(x) < \infty,$$

$$\int G(x) \mid \frac{\partial^2 \log f(x,\theta_0)}{\partial \theta_0^2} \mid d\mu(x) < \infty,$$

and

$$\int G(x)H(x)d\mu(x) < \infty.$$

4.3.7 *Let* $g_n^*(x) = (\sum p_{ni}^2)^{-1} \sum p_{ni}^2 f_i(x)$. *Assume* $g_n^*(x) \to f(x,\theta_0)$ *almost surely and there exists a Borel measurable function* $G^*(x)$ *such that* $g_n^*(x) \leq G^*(x)$ *and*

$$\int G^*(x)\{\frac{\partial \log f(x,\theta_0)}{\partial \theta_0}\}^2 d\mu(x) < \infty.$$

4.3.8

$$\sum p_{ni}^2 Var(\frac{\partial^2 \log f(X_i,\theta_0)}{\partial \theta_0^2}) \to 0$$

and

$$\sum p_{ni}^2 Var(H(X_i)) \to 0$$

as $n \to \infty$.

Some interpretations of these conditions as following. Assumption 4.3.1 insures that the function $\partial \log f(x,\theta)/\partial\theta$ has, for each x, a Taylor expansion as a function of θ. Assumption 4.3.2 insures (justifies) that $\int f(x,\theta)d\mu(x)$ and $\int\{\partial \log f(x,\theta)/\partial\theta\}d\mu(x)$ may be differentiated with respect to θ under the integral sign. Assumption 4.3.3 states that Fisher's information is positive. Assumption 4.3.4 insures a sufficiently fast asymptotic rate and Assumption 4.3.5 insures that the weighted bias go to 0. Assumptions 4.3.6, 4.3.7 and 4.3.8 insure the validity of the Dominated Convergence Theorem and Weak Law of Large Numbers.

Theorem 4.7 *Under the Assumptions 4.1 and 4.3, the relevance weighted likelihood equations admit a sequence of solutions* $\{\hat{\theta}_n\}$ *satisfy:*

(i) $\hat{\theta}_n \to \theta_0$ *in probability* $n \to \infty$*;*

(ii) $(\sum p_{ni}^2)^{-1/2}\{\hat{\theta}_n - \theta_0 - b/I(\theta_0)\} \to N(0, \frac{1}{I(\theta_0)})$.

Here $b = \sum p_{ni} b_i$ *as the asymptotic bias of the MREWLE.*

Further, for any consistent sequence $\hat{\theta}_n$ *of roots of the REWL equations, (ii) is true.*

Theorems 4.5 and 4.7 insure that the MREWLEs are asymptotically normal, when the MREWLEs are roots of REWL equations. By comparing this result with the asymptotic normality of the MLE for the case of iid sampling, we find that the MREWLE has a kind of asymptotic efficiency, because the variance is the inverse of Fisher Information. The convergence rate of the MREWLE is $(\sum p_{ni}^2)^{1/2}$, which depends on the relevance weights. The best convergence rate is $n^{-1/2}$ obtained by choosing $p_{ni} = 1/n$, But in most of relevant sample information cases, we cannot use $p_{ni} = 1/n$ for condition (4.1) and Assumption 4.3.5.

A straight forward result from the above Theorem is

Corollary 4.5 *When* $X_1, ..., X_n$ *are iid from* $f(x, \theta_0)$ *and the Assumptions 4.1 and 4.3 obtain, then any consistent sequence* $\{\hat{\theta}_n\}$ *of roots of the REWL equation (4.4) satisfies*

$$(\sum p_{ni}^2)^{-1/2}(\hat{\theta}_n - \theta_0) \to N(0, \frac{1}{I(\theta_0)}). \tag{4.11}$$

To get the best convergent rate, we always choose $p_{ni} = 1/n$ for the iid exact sample case.

4.5 Estimated Variance of the MREWLE

It was seen in Subsections 4.2, 4.3 and 4.4 that the MREWLE's are consistent and asymptotically normal under certain conditions. The variance of this asymptotically normal distribution provides a reasonable measure of the accuracy of the estimator sequence. But in most cases, we do not know this variance, so an estimator of this variance will be useful.

From Theorem 4.7, there are several possible estimators of the variance. We only discuss two of them. They are

$$\hat{v}_1 = \frac{\sum p_{ni}^2}{\sum p_{ni} \partial^2 \log f(x_i, \theta)/\partial\theta^2} \mid_{\theta=\hat{\theta}}, \tag{4.12}$$

and

$$\hat{v}_2 = \frac{\sum p_{ni}^2 \{\partial \log f(x_i, \theta)/\partial\theta\}^2}{\{\sum p_{ni} \partial^2 \log f(x_i, \theta)/\partial\theta^2\}^2} \mid_{\theta=\hat{\theta}} . \tag{4.13}$$

For the estimator $\hat{v}_1$, we try to estimate the Fisher Information. The $\hat{v}_2$ is obtained directly from the Taylor expansion of the REWL equation (4.4). Under the conditions of Theorem 4.7, we can show that both estimators are consistent.

In Theorem 4.7, the $b \to 0$, so we can use the above variance estimator to construct asymptotic confidence intervals. We do not discuss these variance estimators further in this thesis.

4.6 Some Possible Extensions and Remarks

The method given in this report can be extended to establish the consistency of the MREWLE's for certain types of dependent random variables for which the weak and

strong law of large numbers remain valid.

Assumption 4.1.4 about the importance weights P_n can be extended to general cases where we drop the requirements $p_{ni} \geq 0$ and $\sum_1^n p_{ni} = 1$. But then Conditions (4.1) and (4.2) need to be changed. We could for example replace condition (4.1) by the following stronger conditions:

$$(|\sum_1^n p_{ni}|)^{-1} \sum_1^n |p_{ni}| E \log\{f_i(X_i)/f(X_i, \theta_0)\} \to 0 \text{ as } n \to \infty. \tag{4.14}$$

In Chapter 5, we describe plausible situations where we might want to choose negative weights.

The consistency of MREWLE's can be extended to REWME's defined in Chapter 6. The results of this Chapter apply in several important subdomains of estimation theory indicated by Chapter 5. Of particular note is that of nonparametric smoothing methods.

For simplicity, our treatment in this chapter is confined to the case of a one-dimensional parameter. The multivariate extension is similar and we omit it for brevity.

4.7 Proofs

To prove Theorem 4.1, we need the following two important lemmas about Kullback-Leibler information (KLI; see Kullback 1959).

Lemma 4.1 *Let $f_i(x)$ $i = 1, \ldots, n$ and $g(x)$ be $n+1$ general PDF's with the same support and $q_i \geq 0$ $i = 1, \ldots, n$ be such that $q_1 + q_2 + \ldots + q_n = 1$. If $f(x) = q_1 f_1(x) + \ldots + q_n f_n(x)$, then*

$$\sum_1^n q_i \int f_i(x) \log\{\frac{f_i(x)}{g(x)}\} d\mu(x) \geq \int f(x) \log\{\frac{f(x)}{g(x)}\} d\mu(x) \tag{4.15}$$

with equality if and only if $f_1(x) = f_2(x) = \ldots = f_n(x)$ *almost everywhere with respect to measure* μ.

Lemma 4.2 *If the PDF's* $\{g_n(x)\}$ *and* $g(x)$ *have the same support then* $\lim_{n\to\infty} \int g_n(x) \log\{g_n(x)/g(x)\} d\mu(x) = 0$ *if and only if* $\lim_{n\to\infty} g_n(x)/g(x) = 1$ $[\mu]$ *uniformly.*

We can now apply these results.

Lemma 4.3 *Let* $f_i(x)$ $i = 1, \ldots, \infty$ *and* $g(x)$ *be PDF's with common support and* $P_n = (p_{n1}, p_{n2}, ..., p_{nn})$ *denote the respective importance weights,* $p_{ni} \geq 0$ *and* $\sum_1^n p_{ni} = 1$. *Let* $g_n(x) = \sum p_{ni} f_i(x)$. *If*

$$\sum_1^n p_{ni} \int f_i(x) \log\{\frac{f_i(x)}{g(x)}\} d\mu(x) \to 0 \tag{4.16}$$

then $g_n(x) \to g(x)$ *almost surely.*

Proof: By Lemma 4.1,

$$\sum_1^n p_{ni} \int f_i(x) \log\{\frac{f_i(x)}{g(x)}\} d\mu(x) \geq \int g_n(x) \log\{\frac{g_n(x)}{g(x)}\} d\mu(x) \geq 0.$$

The second inequality follows from the positivity of the KLI.

The assumptions of Lemma's 4.1 and 4.2 imply that $\lim_{n\to\infty}(g_n(x)/g(x)) = 1$ $[\mu]$, uniformly. Now for each $x \in A = \{x : \lim_{n\to\infty}(g_n(x)/g(x)) = 1\}$, we have $\lim_{n\to\infty} g_n(x) = g(x)$. Because $\mu(\bar{A}) = 0$, the conclusion follows. □

Lemma 4.4 *Let* $f_i(x)$ $i = 1, \ldots, \infty$ *and* $g(x)$ *be PDF's with common support and* $P_n = (p_{n1}, p_{n2}, ..., p_{nn})$ *denote the respective importance weights. Suppose condition* (4.16)

holds and $\mu\{x : h(x) \neq g(x)\} > 0$. *Here* $h(x)$ *is another PDF with the same support as* $g(x)$. *Then there exists a* $\delta > 0$ *and an* $N(\delta)$ *such that for* $n > N(\delta)$

$$\sum_1^n p_{ni} \int f_i(x) \log\{\frac{f_i(x)}{h(x)}\} d\mu(x) > \delta. \tag{4.17}$$

Proof: The KLI is always positive. If the inequality (4.17) is not true, then there exists a subsequence $n(j)$ such that

$$\sum_1^{n(j)} p_{n(j)i} \int f_i(x) \log\{\frac{f_i(x)}{h(x)}\} d\mu(x) \to 0 \text{ as } j \to \infty.$$

Now let $g_n(x) = \sum p_{ni} f_i(x)$. By the last Lemma, we get $g_{n(j)}(x) \to h(x)$ almost surely. Also by the same Lemma, $g_n(x) \to g(x)$ almost surely, so $g_{n(j)}(x) \to g(x)$ almost surely. Therefore $\mu\{x : h(x) \neq g(x)\} = 0$. This contradicts the assumption and completes the proof. □

Proof of Theorem 4.1: The inequality in (4.3) is equivalent to

$$\sum_1^n p_{ni} \log f(X_i, \theta_0) - \sum_1^n p_{ni} \log f(X_i, \theta) \geq 0.$$

Now

$$\begin{aligned} & \sum_1^n p_{ni} \log f(X_i, \theta_0) - \sum_1^n p_{ni} \log f(X_i, \theta) \\ = & \sum_1^n p_{ni} \log\{\frac{f_i(X_i)}{f(X_i, \theta)}\} - \sum_1^n p_{ni} \log\{\frac{f_i(X_i)}{f(X_i, \theta_0)}\} \\ = & (I) + (II) \end{aligned}$$

From (4.1) and lemma 4.2, $\sum p_{ni} f_i(x) \to f(x, \theta_0)$ almost surely. Now because $f(x, \theta)$ and $f(x, \theta_0)$ are distinct densities, then by Lemma 4.4, there exists a $\delta > 0$ such that for n large enough

$$\sum_1^n p_{ni} E \log\{\frac{f_i(X_i)}{f(X_i, \theta)}\} \geq \delta.$$

Now from Assumption 4.1.4,

$$(I) - \sum_{1}^{n} p_{ni} E \log\{\frac{f_i(X_i)}{f(X_i,\theta)}\} \to 0 \; \textit{in probability.}$$

Therefore $(I) \geq \delta$ in probability. Similarly, $(II) \to 0$ in probability. This means

$$\sum_{1}^{n} p_{ni} \log f(X_i,\theta_0) - \sum_{1}^{n} p_{ni} \log f(X_i,\theta) \geq 0$$

in probability. That observation completes the proof. □

Proof of Theorem 4.2: Let a be small enough so that $(\theta_0 - a, \theta_0 + a)$ contains in Θ and let

$$S_n = \{\underline{x} : WL(\theta_0,\underline{x}) > WL(\theta_0 - a,\underline{x}) \; and \; WL(\theta_0,\underline{x}) > WL(\theta_0 + a,\underline{x})\}.$$

By Theorem 4.1, $P_{\theta_0}(S_n) \to 1$. For any $\underline{x} \in S_n$ there thus exists a value $\theta_0 - a < \hat{\theta}_n < \theta_0 + a$ at which $WL(\theta)$ has a local maximum, so that $WL'(\hat{\theta}_n) = 0$. Hence for any $a > 0$ sufficiently small, there exists a sequence $\hat{\theta}_n = \hat{\theta}_n(a)$ of roots such that $P_{\theta_0}(|\hat{\theta}_n - \theta_0| < a) \to 1$.

It is remains to show that we can determine such a sequence, which does not depend on a. Let θ_n^* be the closest root to θ_0. [This root exists because the limit of a sequence of roots is again a root by the assumed continuity of $WL(\theta)$.] Then $P_{\theta_0}(|\theta_n^* - \theta_0| < a) \to 1$ and this completes the proof. □

Before we prove Theorem 4.3, we state the following Lemmas.

Lemma 4.5 *$\varphi(\theta,r), f^*(x,\theta,\rho)$ and $\varphi^*(\theta,r)$ are Borel measurable functions.*

This last lemma follows from immediately from Assumption 4.2.1 and we omit it for brevity. The next two lemmas follow from Wald (1949).

Lemma 4.6 *For any $\theta \neq \theta_0$ in Θ we have*

$$\lim_{\rho \to 0} E_{\theta_0} \log f(X, \theta, \rho) = E_{\theta_0} \log f(X, \theta). \tag{4.18}$$

Lemma 4.7 *For any $\theta_0 \in \Theta$ we have*

$$\lim_{r \to \infty} E_{\theta_0} \log \varphi(X, r) = -\infty. \tag{4.19}$$

Lemma 4.8 *For any $\delta > 0$, there exist $r_o(\delta) > 0$ and $N(\delta, r_o)$, such that for every $n > N(\delta, r_o)$ and $r \geq r_o$,*

$$\sum_1^n p_{ni} E \log f(X_i, \theta_0) - \sum_1^n p_{ni} E \log \varphi(X_i, r) > \delta. \tag{4.20}$$

Proof: From (4.15), we have

$$\begin{aligned}
& \sum_1^n p_{ni} E \log f(X_i, \theta_0) - \sum_1^n p_{ni} E \log \varphi(X_i, r) \\
= & \sum_1^n p_{ni} E \log\{\frac{f_i(X_i)}{\varphi(X_i, r)}\} - \sum_1^n p_{ni} E \log\{\frac{f_i(X_i)}{f(X_i, \theta_0)}\} \\
\geq & \int g_n(x) \log\{\frac{g_n(x)}{\varphi(x, r)}\} d\mu(x) - \sum_1^n p_{ni} E \log\{\frac{f_i(X_i)}{f(X_i, \theta_0)}\} \\
= & \int g_n(x) \log\{\frac{g_n(x)}{f(x, \theta_0)}\} d\mu(x) + \int g_n(x) \log f(x, \theta_0) d\mu(x) \\
& - \int g_n(x) \log \varphi(x, r) d\mu(x) - \sum_1^n p_{ni} E \log\{\frac{f_i(X_i)}{f(X_i, \theta_0)}\} \\
= & (I) + (II) - (III) - (IV)
\end{aligned}$$

By Assumption 4.2.8, $(I) \to 0$ and $(IV) \to 0$.

Now we prove $(II) \to E_{\theta_0} \log f(X, \theta_0)$. From Lemma 4.3, we know that $g_n(x) \to f(x, \theta_0)$. The result now follows from the Assumption 4.2.9 and the Dominated Convergence Theorem. We can prove $(III) \to E_{\theta_0} \log \varphi(X_i, r)$. in a similar fashion. From (4.19), we can choose r_o such that $E_{\theta_0} \log \varphi(X_i, r_o) < -(5\delta + |E_{\theta_0} \log f(X, \theta_0)|)$. Now choose

N_1, N_2, N_3 and N_4 such that: (i)when $n > N_1$, $|(I)| < \delta$; (ii)when $n > N_2$, $|(II) - E_{\theta_0} \log f(X, \theta_0)| < \delta$; (iii)when $n > N_3$, $|(III) - E_{\theta_0} \log \varphi(X_i, r_o)| < \delta$; and (iv)when $n > N_4$, $|(IV)| < \delta$, respectively. Let $N(\delta, r_o) = \max\{N_1, N_2, N_3, N_4\}$, we have prove (4.20) for $r = r_o$. Because $\varphi(X_i, r)$ decreases with r, the proof is complete. □

Lemma 4.9 *For any $\theta \neq \theta_0$ in Θ, there exist $\delta > 0$, $\rho(\theta, \delta) > 0$ and $N(\rho(\theta, \delta), \delta)$ such that for $n > N(\rho(\theta, \delta), \delta)$,*

$$\sum_1^n p_{ni} E \log f(X_i, \theta_0) - \sum_1^n p_{ni} E \log f\{X_i, \theta, \rho(\theta, \delta)\} > \delta. \tag{4.21}$$

Proof: The proof of this lemma is similar to that of the last. The only difference is that we use (4.18) instead of (4.19). □

Proof of Theorem 4.3: From Lemma 4.8, we know there exist r_o and $N(r_o)$ such that

$$\sum_1^n p_{ni} E \log f(X_i, \theta_0) - \sum_1^n p_{ni} E \log \varphi(X_i, r_o) > 1. \tag{4.22}$$

Let $\omega_1 = \omega \cap \{\theta : \|\theta\| \leq r_o\}$. Then by Lemma 4.9, for any $\theta \in \omega_1$, there exist $\rho(\theta) > 0$, $\delta(\theta) > 0$ and $N(\theta, \rho(\theta), \delta(\theta))$ such that $n > N(\theta, \rho(\theta), \delta(\theta))$ and

$$\sum_1^n p_{ni} E \log f(X_i, \theta_0) - \sum_1^n p_{ni} E \log f(X_i, \theta, \rho(\theta, \delta)) > \delta(\theta). \tag{4.23}$$

Now let $S(\theta, \rho)$ denote the sphere with center θ and radius ρ. Since ω_1 is compact, by the Finite Covering Theorem, there exists a finite number of points $\{\theta_1, \ldots, \theta_k\}$ in ω_1 such that $S(\theta_1, \rho(\theta_1)) \cup \ldots \cup S(\theta_k, \rho(\theta_k))$ contains ω_1 as a subset. Clearly, we have

$$\begin{aligned} & 0 \leq \sup_{\theta \in \omega} f^{p_{n1}}(X_1, \theta) \ldots f^{p_{nn}}(X_n, \theta) \\ \leq & \sum_1^k f^{p_{n1}}\{X_1, \theta_i, \rho(\theta_i)\} \ldots f^{p_{nn}}\{X_n, \theta_i, \rho(\theta_i)\} + \varphi^{p_{n1}}(X_1, r_o) \ldots \varphi^{p_{nn}}(X_n, r_o). \end{aligned}$$

Hence, the theorem is proved if we can show that

$$P_n[\{\frac{f^{p_{n1}}(X_1,\theta_i,\rho(\theta_i))\dots f^{p_{nn}}(X_n,\theta_i,\rho(\theta_i))}{f^{p_{n1}}(X_1,\theta_0)\dots f^{p_{nn}}(X_n,\theta_0)}\}^n \geq \epsilon/(k+1)] \to 0, \quad i=1,\dots,k.$$

and

$$P_n[\{\frac{\varphi^{p_{n1}}(X_1,r_o)\dots\varphi^{p_{nn}}(X_n,r_o)}{f^{p_{n1}}(X_1,\theta_0)\dots f^{p_{nn}}(X_n,\theta_0)}\}^n \geq \epsilon/(k+1)] \to 0.$$

Proving these last results is equivalent to showing that for $i=1,\dots,k$

$$n\{\sum_1^n p_{ni}\log f(X_i,\theta_0) - \sum_1^n p_{ni}\log f(X_i,\theta_j,\rho(\theta_j))\} \to \infty \ \textit{in probability.} \tag{4.24}$$

and

$$n\{\sum_1^n p_{ni}\log f(X_i,\theta_0) - \sum_1^n p_{ni}\log \varphi(X_i,r_o)\} \to \infty \ \textit{in probability.} \tag{4.25}$$

Under our assumptions, (4.22), (4.23) and the weak law of large numbers, we can prove (4.24) and (4.25). This completes the proof of this theorem. □

Proof of Theorem 4.4: For any $\varepsilon > 0$, let

$$A_\varepsilon = \{(X_1,X_2,\dots) : \bar{\theta}_n(X_1,\dots,X_n) \in S(\theta_0,\varepsilon) \ \textit{for sufficient large } n\}.$$

From (4.9), we obtain

$$A_\varepsilon^c \subset C_\varepsilon = \{(X_1,X_2,\dots) :$$
$$\{\frac{\sup_{|\theta-\theta_0|\geq\varepsilon} f^{p_{n1}}(X_1,\theta)\dots f^{p_{nn}}(X_n,\theta)}{f^{p_{n1}}(X_1,\theta_0)\dots f^{p_{nn}}(X_n,\theta_0)}\}^n \geq c \ \textit{for infinitely many } n\}$$

By Theorem 4.3, we have $\lim_{n\to\infty} P_n(C_\varepsilon) = 0$, then $\lim_{n\to\infty} P_n(A_\varepsilon^c) = 0$. Therefore, $\lim_{n\to\infty} P_n(A_\varepsilon) = 1$. This completes the proof. □

Proof of Theorem 4.5: If we can prove (4.24) and (4.25) with probability 1, then from the proof of Theorem 4.3, we obtain the asserted result. But Theorem 4 of Stout (1968) and our conditions on P_n imply this result and hence the conclusion of our theorem. □

The proof of Theorem 4.6 is similar to Theorem 4.4 and hence we omit this for brevity.

Proof of Theorem 4.7: By Assumption (4.3.1) and (4.3.2), we have for θ in the neighborhood $N(\theta_0)$ a Taylor expansion of $\partial \log f(x,\theta)/\partial\theta$ about the point $\theta = \theta_0$ as follows:

$$\frac{\partial \log f(x,\theta)}{\partial\theta} - \frac{\partial \log f(x,\theta)}{\partial\theta} \mid_{\theta=\theta_0} = (\theta - \theta_0)\frac{\partial^2 \log f(x,\theta)}{\partial\theta^2} \mid_{\theta=\theta_0} + \frac{1}{2}\xi(\theta - \theta_0)^2 H(x)$$

where $\mid \xi \mid < 1$. Therefore, putting

$$A_n = \sum p_{ni} \frac{\partial \log f(X_i,\theta)}{\partial\theta} \mid_{\theta=\theta_0},$$

$$B_n = \sum p_{ni} \frac{\partial^2 \log f(X_i,\theta)}{\partial\theta^2} \mid_{\theta=\theta_0},$$

and

$$C_n = \sum p_{ni} H(X_i).$$

We have

$$-A_n = B_n(\hat{\theta}_n - \theta_0) + \frac{1}{2}\xi^* C_n(\hat{\theta}_n - \theta_0)^2$$

where $\mid \xi^* \mid < 1$.

From Theorem 4.2, we know that there exist a sequence of $\hat{\theta}_n$ such that $\hat{\theta}_n \to \theta_0$ so

$$\hat{\theta}_n - \theta_0 = \frac{-A_n}{B_n + \frac{1}{2}\xi^* C_n(\hat{\theta}_n - \theta_0)}$$

Now we prove:

(i) $A_n \sim AN(\sum p_{ni} b_i, (\sum p_{ni}^2) I(\theta_0))$.

As we know

$$A_n = \sum p_{ni} \frac{\partial \log f(X_i,\theta)}{\partial\theta} \mid_{\theta=\theta_0} .$$

so $EA_n = -\sum p_{ni} b_i$ and

$$\begin{aligned} Var(A_n) &= \sum p_{ni}^2 E(\frac{\partial \log f(X_i,\theta)}{\partial \theta} \mid_{\theta=\theta_0})^2 - \sum p_{ni}^2 b_i^2 \\ &= (\sum p_{ni}^2) \int (\frac{\partial \log f(x,\theta_0)}{\partial \theta_0})^2 \sum \frac{p_{ni}^2}{\sum p_{ni}^2} f_i(x) d\mu(x) - \sum p_{ni}^2 (\sum \frac{p_{ni}^2}{\sum p_{ni}^2} b_i^2). \end{aligned}$$

By Assumption 4.3.5, 4.3.7 and Dominated Convergence Theorem

$$Var(A_n) \to (\sum p_{ni}^2) I(\theta_0) + o(\sum p_{ni}^2).$$

Now let

$$\begin{aligned} \Gamma_n &= \sum_{i=1}^{n} E \mid p_{ni} \frac{\partial \log f(X_i,\theta_0)}{\partial \theta_0} \mid^3 \\ &= \sum_{i=1}^{n} p_{ni}^3 E \mid \frac{\partial \log f(X_i,\theta_0)}{\partial \theta_0} \mid^3 \\ &\leq max(p_{ni}^2) \int \mid \frac{\partial \log f(x,\theta_0)}{\partial \theta_0} \mid^3 \sum p_{ni} f_i(x) d\mu(x). \end{aligned}$$

By Dominated Convergence Theorem

$$\int \mid \frac{\partial \log f(x,\theta_0)}{\partial \theta_0} \mid^3 \sum p_{ni} f_i(x) d\mu(x) \longrightarrow \int \mid \frac{\partial \log f(x,\theta_0)}{\partial \theta_0} \mid^3 f(x,\theta_0) d\mu(x),$$

so $\Gamma \to 0$ ($max(p_{ni}^2) \to 0$). Then by Theorem 7.1.2 of Chung (1974, p200) and Assumption 4.3.4, we get

$$A_n \sim AN\{\sum p_{ni} b_i, (\sum p_{ni}^2) I(\theta_0)\}.$$

(ii) $B_n \to -I(\theta_0)$ in probability.

$$B_n = \sum p_{ni} \frac{\partial^2 \log f(X_i,\theta_0)}{\partial \theta_0^2}$$

$$\begin{aligned} \sum p_{ni} E \frac{\partial^2 \log f(X_i,\theta_0)}{\partial \theta_0^2} &= \int \frac{\partial^2 \log f(x,\theta_0)}{\partial \theta_0^2} \sum p_{ni} f_i(x) d\mu(x) \\ &\longrightarrow -I(\theta_0) \end{aligned}$$

by Dominated Convergence Theorem.

So from Assumption 4.3.8, $B_n \to I(\theta_0)$ in probability.

(iii) $C_n \to E_{\theta_0} H(X)$ in probability. The proof is similar to (ii).

Therefore $\hat{\theta}_n - \theta_0 = -A_n/[B_n + \frac{1}{2}\xi^* C_n(\hat{\theta}_n - \theta_0)]$. Since $\hat{\theta}_n - \theta_0 \to 0$ in probability.

$$B_n + \frac{1}{2}\xi^* C_n(\hat{\theta}_n - \theta_0) \to -I(\theta_o)$$

Further, $-A_n \sim AN(-\sum p_{ni} b_i, (\sum p_{ni}^2) I(\theta_0))$. Consequently, by Slutsky's Theorem,

$$(\sum p_{ni}^2)\{\hat{\theta}_n - \theta_0 - \frac{\sum p_{ni} b_n}{I(\theta_0)}\} \to N(0, \frac{1}{I(\theta_0)})$$

establishing the theorem.

Chapter 5

The MREWLE for Generalized Smoothing Models

5.1 Introduction

As we have mentioned in the Chapter 2 and 3, the nonparametric regression (NR) paradigm motivates much of our work on relevance weighted inference. Methods developed in NR use relevant information in relevant samples (see Chapter 2). Nonparametric regression provides a useful explanatory and diagnostic tool for this propose. See Eubank (1988), Hardle (1990), and Muller (1988) for good introductions to the general subject area. Several methods have been proposed for estimating $m(x)$: kernel, spline, and orthogonal series. Recently, local likelihood was introduced by Tibshirani and Hastie (1987) as a method of smoothing by 'running lines' in nongaussian regression models. Staniswalis (1989) carried out a similar generalization of the kernel estimator.

In this Chapter, we apply the method of MREWLE to capture the relevant sample information for generalized smoothing models. Both local likelihood and Staniswalis's method can be viewed as special cases of our methods. We wish to demonstrate the

applicability of our methodology, whose primary advantage over individual methods which have been developed in NR, lies in its generality. However, we are able to establish secondary advantages as well. The MREWLE always has a smaller variance which depends on Fisher information.

Usually NR is used for the location parameter and it also assumes the mean and variance of the observations exist. But in many cases, we are interested in some other parameters, for example, the variance of a normal distribution or parameters of the Weibull distribution in Section 5.5. In some cases, even when we are interested in the location parameter, the mean and variance of the observations may not exist. The Cauchy distribution in Section 5.5 is an example. The method of MREWLE works well in these situations.

Under the model of Example 3.6, we know that given X_i, Y_i has density $f(y, \theta(X_i))$ (assumed known up to the unknown parameter $\theta(X_i)$). We seek to estimate $\theta(x)$ at the fixed point $X = x$. After choosing the relevance weights, we get the MREWLE by maximizing over $\theta(x)$,

$$\prod_1^n f(Y_i, \theta(x))^{p_{ni}(x)}.$$

The next straightforward result shows how locally weighted regression estimators obtain.

Theorem 5.1 . *If $f(y, \theta(X_i))$ is the density for a normal distribution with mean $\mu(X_i)$ and variance $\sigma^2(X_i)$ [here $\theta(x) = \{\mu(x), \sigma^2(x)\}$], then the MREWLE is*

$$\hat{\mu}(x) = \sum_1^n p_{ni}(x) Y_i \tag{5.1}$$

and

$$\hat{\sigma}^2 = \sum_1^n p_{ni}(x) \{Y_i - \hat{\mu}(x)\}^2. \square \tag{5.2}$$

Thus the MREWLE in the normal case of the last theorem, is the linear smoother of Fan (1992). Obviously we can get Nadaraya-Watson kernel estimators, k-nearest neighbor estimates, Gasser-Muller estimators and locally linear regression smoothers by choosing appropriate relevance weights. These smoothers include those generated by the use of spline and orthogonal series methods. This means that the MREWLE method subsumes current nonparametric smoothing methods when the error distribution is normal. The variance estimator in (5.2) is the same as the variance estimate in Hardle (1990). Here it is a direct result of the MREWLE.

We now study the MREWLE in relation to current nonparametric smoothing methods in situations where the error distribution family is known. Our discussion addresses both asymptotic and non-asymptotic issues.

We organize this chapter as follows. Small sample properties are considered in Section 5.2. The main asymptotic results are shown in Section 5.3. Also in Section 5.3, we compare the MREWLE with current NR methods. How to choose the relevant weights is considered in Section 5.4. Finally in Section 5.5, we give some simulation results.

5.2 Small samples.

Our earlier discussion leads us to wonder about the difference between the MREWLE and other linear smoothers for nonnormal error models. The following theorem partially addresses this issue. There we refer to conventional sufficiency with respect to the joint sampling distribution of all the data.

Theorem 5.2 . *Suppose the sufficient statistics for the error distribution family are not linear in the data. If $X_i = X_j$ for some $i \neq j$, then with respect to quadratic loss,*

the linear smoother is inadmissible.

Proof. We easily obtain the conclusion using Rao-Blackwell Theorem and sufficiency at the replication points. □

The last theorem shows that if we have replicate observations in a designed experiment, we can achieve a uniformly smaller risk than that of any linear smoother (which depends linearly on the data when sufficiency shows it should not). For instance, the one parameter exponential family has sufficient statistics based on $\{\sum T(Y_{i'})\}$ for some function T) where the sums are taken at the replication points. Obviously basing any smoother on the $\{Y_i\}$ would violate the sufficiency principle in this case.

It could be argued that this claim is unfair. Linear smoothers are proposed in a nonparametric-nonparametric framework where neither the error distribution nor the regression function has parametric form.

That argument ignores Theorem 5.1 which shows that these linear smoothers are consistent with normal error models. Indeed, in Chapter 6 we obtain smoothers in the nonparametric- nonparametric setting which are very different than linear smoothers. So the argument fails to blunt the impact of the last theorem. Rather that theorem points to the nonrobustness of linear smoothing methods. Theorem 5.2 emphasizes the importance of the vehicle which carries the data into a smoothing procedure. And it tells us how to improve on a linear smoother if we have repeated observations at some points.

We know by weak sufficiency that the MREWLE must depend only on the sufficient statistics. So it evades the difficulty confronted above by linear smoothers.

When the $\{X\}$'s are continuous covariates, we cannot (in principle) have repeated observations at any point, so cannot improve on linear smoothers by invoking the last Theorem. But we may nevertheless have near ties among the $\{X\}$'s in which case the heuristics underlying that theorem still obtain. Large sample theory below will lead to further discussion of this issue.

In Chapter 2 and 3, we emphasized the NR paradigm because it provided a context wherein some information from the relevant samples have been used to advantage. The last theorem suggests that these methods fail to use all the relevant information available when the error distribution cannot be assumed to be normal. In this way, these linear smoothers seem analogous to moment estimators in classical estimation theory; the MREWLE would then be analogous to the MLE.

5.3 Asymptotic Properties

We begin with the generalized smoothing models described in Example 3.6. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a random sample from a population (X, Y). For given $X = x$, Y has density function $f\{y, \theta(x)\}$.

Because we have used relevant sample information for estimation, the MREWLE is usually biased. We define the bias function as

$$B_x(z) = E_{\theta(z)} \partial \log f\{Y, \theta(x)\}/\partial\theta(x). \tag{5.3}$$

This bias function indicates the bias when we used the information from $Y|X = z$ to estimate $\theta(x)$. Under some conditions, we can get $B_x(x) = 0$ and

$$B_x(z) = I\{\theta(x)\}\{\theta(z) - \theta(x)\} + o\{\theta(z) - \theta(x)\}, \tag{5.4}$$

where $I\{\theta(x)\}$ is the Fisher information function for $\theta(x)$. Equation (5.4) indicates the meaning of this bias function. Also this bias function is a special case of the bias function in Chapter 4.

In the next four subsections we present classes of possible relevance weights suggested by results in modern NR theory. We study one of these classes, that suggested by Nadaraya-Watson, in some detail. For the rest, some expected results will merely be sketched for brevity.

(I) Kernel weights (Nadaraya-Watson).

The weight sequence for kernel smoothers (for a one dimensional x) is defined by

$$p_{ni}(x) = K_{h_n}(x - X_i)/\{n\hat{g}_{h_n}(x)\}, \tag{5.5}$$

where

$$\hat{g}_{h_n}(x) = n^{-1}\sum_{1}^{n} K_{h_n}(x - X_i) \quad and \quad K_{h_n}(u) = h_n^{-1}K(u/h_n). \tag{5.6}$$

The *kernel* K is a continuous, bounded and symmetric real function which integrates to one,

$$\int K(u)du = 1. \tag{5.7}$$

Because the form (5.5) of kernel weights $p_{ni}(x)$ has been proposed by Nadaraya (1964) and Watson (1964), we call these Nadaraya-Watson weights.

The MREWLE with Nadaraya-Watson weights obtains from maximizing

$$n^{-1}\sum\{\hat{g}_{h_n}(x)\}^{-1}K_{h_n}(x - X_i)\log f\{Y_i, \theta(x)\}.$$

We now consider the asymptotic properties of this MREWLE. In the sequel, we always let

$$c_K = \int_{-\infty}^{\infty} u^2K(u)du, \quad d_K = \int_{-\infty}^{\infty} K^2(u)du. \tag{5.8}$$

We need the following assumptions:

Assumptions 5.1 **5.1.1** *The bias function $B_x(z)$ has a bounded and continuous second derivative for every fixed x.*

5.1.2 *Let $B_z^*(x) = E_{\theta(z)} \log\{f(Y,\theta(z))/f(Y,\theta(x))\}$. $B_z^*(x)$ has a bounded and continuous first derivative for every fixed x.*

5.1.3 *The marginal density $g(x)$ of the covariate X has continuous first derivative and is bounded away from zero in an interval (a_o, b_o).*

5.1.4 *$\int uK(u)du = 0$ and $\int u^4 K(u)du < \infty$.*

5.1.5 *The density function $f(y,\theta))$ satisfies the following regularity conditions:*
(i) $\log f(y,\theta)$ has three continuous partial derivatives with respect to θ;
(ii) for each $\theta_0 \in \Theta$, there exists integrable functions $H_i(y)$ such that for θ in a neighborhood $N(\theta_0)$ the relations

$$|\frac{\partial f(y,\theta)}{\partial\theta}| \leq H_1(y), |\frac{\partial^2 f(y,\theta)}{\partial\theta^2}| \leq H_2(y), |\frac{\partial^3 \log f(y,\theta)}{\partial\theta^3}| \leq H_3(y)$$

hold, for all y, and

$$\int H_1(y)dy < \infty, \int H_2(y)dy < \infty, E_\theta(H_3(Y)) < \infty$$

for $\theta \in N(\theta_0)$;
(iii) the Fisher information $I\{\theta(x)\}$ is continuously differentiable and bounded away from zero.

We state the following pointwise properties of the MREWLE.

Theorem 5.3 *Under the Assumption 5.1, assume $h_n \to 0$ and $nh_n \to \infty$. If x is a fixed point in (a_o, b_o), then the REWL equations admit a sequence of solutions $\{\hat{\theta}_n\}$ satisfying:*

$$\left[\frac{nh_n g(x) I\{\theta(x)\}}{d_K}\right]^{1/2} [\hat{\theta}(x) - \theta(x) - \left\{\frac{B_x''(x)}{2I\{\theta(x)\}} + \frac{B_x'(x)g'(x)}{g(x)I\{\theta(x)\}}\right\} c_K h_n^2\{1 + O(h_n)\}] \to N(0,1). \tag{5.9}$$

Proof: For $x \in (a_o, b_o)$, similar to the proof of Theorem 4.2, we can show that the REWL equations admit a sequence of solutions $\{\hat{\theta}_n\}$ satisfy $\hat{\theta}_n \to \theta(x)$ in probability. As in the proof of Theorem 4.7, putting

$$A_n = \sum p_{ni}(x) \frac{\partial \log f\{Y_i, \theta(x)\}}{\partial \theta(x)},$$

$$B_n = \sum p_{ni}(x) \frac{\partial^2 \log f\{Y_i, \theta(x)\}}{\partial \theta(x)^2},$$

and

$$C_n = \sum p_{ni}(x) H_3(Y_i),$$

yields

$$-A_n = B_n\{\hat{\theta}_n(x) - \theta(x)\} + \frac{1}{2}\xi^* C_n\{\hat{\theta}_n(x) - \theta(x)\}^2.$$

Here $| \xi^* | < 1$. Now

$$\hat{\theta}(x) - \theta(x) = \frac{-A_n}{B_n + \frac{1}{2}\xi^* C_n\{\hat{\theta}(x) - \theta(x)\}}.$$

We prove:

$$A_n \sim AN[(B_x''(x)/2 + \frac{B_x'(x)g'(x)}{g(x)})c_K h_n^2\{1 + O(h_n)\}, \frac{I\{\theta(x)\}d_K}{nh_n g(x)}]. \tag{5.10}$$

As we know

$$A_n = \sum p_{ni}(x) \frac{\partial \log f\{Y_i, \theta(x)\}}{\partial \theta(x)}$$

$$= \sum \frac{K_{h_n}(x - X_i)}{n\hat{g}_{h_n}(x)} \frac{\partial \log f\{Y_i, \theta(x)\}}{\partial \theta(x)}$$
$$= \frac{1}{n\hat{g}_{h_n}(x)} \sum K_{h_n}(x - X_i) \frac{\partial \log f\{Y_i, \theta(x)\}}{\partial \theta(x)}.$$

Here $K_{h_n}(x - X_i)\partial \log f\{Y_i, \theta(x)\}/\partial\theta(x)$ are iid random variables with expectation

$$\begin{aligned} a &= EK_{h_n}(x - X_i) \frac{\partial \log f\{Y_i, \theta(x)\}}{\partial \theta(x)} \\ &= \int K_{h_n}(x - z) \frac{\partial \log f(y, , \theta(x))}{\partial \theta(x)} f\{y, \theta(z)\} g(z) dy dz \\ &= \int K_{h_n}(x - z) B_x(z) g(z) dz \\ &= \int K(u) B_x(x - h_n u) g(x - h_n u) du \\ &= \{B_x''(x)g(x)/2 + B_x'(x)g'(x)\} c_K h_n^2 \{1 + O(h_n)\}. \end{aligned}$$

The variance is

$$\begin{aligned} b &= Var[K_{h_n}(x - X_i) \frac{\partial \log f\{Y_i, \theta(x)\}}{\partial \theta(x)}] \\ &= \int [K_{h_n}(x - z) \frac{\partial \log f(y, , \theta(x))}{\partial \theta(x)}]^2 f\{y, \theta(z)\} g(z) dy dz \\ &- [EK_{h_n}(x - X_i) \frac{\partial \log f\{Y_i, \theta(x)\}}{\partial \theta(x)}]^2 \\ &= \frac{I\{\theta(x)\} d_K g(x)}{h_n} \{1 + O(h_n)\}. \end{aligned}$$

From the Central Limit Theorem, we have

$$\sqrt{nb^{-1}} \{\hat{g}_{h_n}(x) A_n - a\} \to N(0, 1) \tag{5.11}$$

As we know that $\hat{g}_{h_n}(x) = n^{-1} \sum K_{h_n}(x - X_i)$, the

$$\begin{aligned} EK_{h_n}(x - X_i) &= \int K_{h_n}(x - z) g(z) dz \\ &= \int K(u) g(x - h_n u) du = g(x)\{1 + O(h_n)\} \end{aligned}$$

and $Var\{K_{h_n}(x - X_i)\} = d_K g(x) {h_n}^{-1} \{1 + O(h_n)\}$. By the Central Limit Theorem,

$$\sqrt{\frac{nh_n}{g(x) d_K}} \{\hat{g}_{h_n}(x) - g(x)\{1 + O(h_n)\}\} \to N(0, 1). \tag{5.12}$$

Now

$$A_n - \frac{a}{g(x)\{1+O(h_n)\}} = \frac{g(x)\{1+O(h_n)\}(\hat{g}_{h_n}(x)A_n - a) - a[\hat{g}_{h_n}(x) - g(x)\{1+O(h_n)\}]}{\hat{g}_{h_n}(x)g(x)\{1+O(h_n)\}}.$$

So

$$g(x)\sqrt{nb^{-1}}[A_n - \frac{a}{g(x)\{1+O(h_n)\}}] \to N(0,1) \tag{5.13}$$

by (5.11), (5.12) and Slutsky's Theorem ($\hat{g}_{h_n}(x) \to g(x)$ in probability and $a \to 0$). We get result (5.10) from (5.13).

As in the above proof, by the SLLN and Slutsky's theorem,

$$B_n \to -I\{\theta(x)\} \ \textit{in probability} \tag{5.14}$$

and

$$C_n \to E_{\theta(x)}H_3(Y) \ \textit{in probability}. \tag{5.15}$$

By (5.10), (5.14), (5.15) and Slutsky's theorem, we get the conclusion of the theorem. □

In Theorem 5.3, we only prove that the REWL equations admit a sequence of solutions which is asymptotically normal. As we have discovered in Chapter 4, the MREWLE is a consistent sequence of solutions under certain conditions. In the following discussion, we always assume that the MREWLE is a consistent sequence of solutions of the REWL equations.

The last theorem gives us the asymptotic bias and variance for $\hat{\theta}$. Then they can formally be combined to give a conditional MSE (mean square error) result for the MREWLE of θ. The result suggested heuristically through that combination suggests the MREWLE has the conditional MSE

$$E[\{\hat{\theta}(x) - \theta(x)\}^2 | X_1, \ldots, X_n] = [\frac{B_x''(x)}{2I\{\theta(x)\}} + \frac{B_x'(x)g'(x)}{g(x)I\{\theta(x)\}}]^2 c_K^2 h_n^4$$

$$+ \quad \frac{d_K}{nh_n g(x) I\{\theta(x)\}} + o_p(h_n^4 + \frac{1}{nh_n}), \qquad (5.16)$$

when x is any fixed point in (a_o, b_o).

Similar results are suggested for the case of fixed design variables of the form $X_j = G^{-1}(j/n) + o(1/n)$, where the function G is the cdf of $g(x)$. We defer for future work, the formal proof of these and other results which are stated below.

When $\theta(x)$ is a location parameter, then we can use the ordinary Nadaraya-Watson kernel estimator. The MSE of Nadaraya-Watson kernel estimator is (Hardle 1990, P77)

$$MSE = \{\frac{\theta''(x)}{2} + \frac{\theta'(x)g'(x)}{g(x)}\}^2 c_K^2 h_n^4 + \frac{d_K \sigma^2(x)}{nh_n g(x)} + o_p(h_n^4 + \frac{1}{nh_n}). \qquad (5.17)$$

Comparing (5.16) with (5.17) suggests tentatively that:

(i) the MREWLE usually has smaller variance which depends on Fisher information (like the Fisher information bound for exact sample case);

(ii) the bias of the MREWLE depends on the bias function $B_x(z)$ and Fisher information function, while the bias of the Nadaraya-Watson kernel estimator depends on the regression function $\theta(x)$. [But it is hard to compare the biases of these two method.]

(II) Gasser-Muller type weights. Let us now use the weights derived from those of Gasser and Muller (1979). Their results formally suggest the following conclusion for the resulting MREWLE:

$$\begin{aligned} [\frac{2}{3}\frac{nh_n g(x) I\{\theta(x)\}}{d_K}]^{1/2}[\hat{\theta}(x) - \theta(x) \quad &- \quad \frac{B_x''(x)}{2I\{\theta(x)\}} c_K h_n^2 \{1 + O(h_n)\}] \\ &\rightarrow \quad N(0,1), \end{aligned} \qquad (5.18)$$

if x is a fixed point in (a_o, b_o). We leave verification of this result to future work.

(III) k-NN weights. Use of the k-NN weights defined in Hardle (1989, page 42), and his results heuristically suggest that the MREWLE sequence, $\{\hat{\theta}_n\}$ with k-NN weights satisfies:

$$\begin{aligned}[\frac{kI\{\theta(x)\}}{2d_K}]^{1/2}[\hat{\theta}(x)-\theta(x) &- (\frac{k}{n})^2\frac{B_x''(x)g(x)+2B_x'(x)g'(x)}{8g^3(x)I\{\theta(x)\}}c_K\{1+O(\frac{k}{n})\}] \\ &\rightarrow N(0,1) \end{aligned} \tag{5.19}$$

for fixed x.

(IV) Locally linear smoother weights. Locally linear smoother weights are defined by Fan (1992) for nonparametric regression. Using those weights to define the MREWLE for generalized smoothing models, suggests that

$$\begin{aligned}[\frac{nh_ng(x)I\{\theta(x)\}}{d_K}]^{1/2}[\hat{\theta}(x)-\theta(x) &- \frac{B_x''(x)}{2I\{\theta(x)\}}c_Kh_n^2\{1+O(h_n)\}] \\ &\rightarrow N(0,1) \end{aligned} \tag{5.20}$$

by analogy Fan's results, for any fixed x.

Now we compare the MREWLE smoothers with the nonparametric regression smoothers. In the following discussion, assume $\theta(x)$ is a location parameter. So nonparametric regression smoothers are applicable. For comparison, we summarize the asymptotic pointwise bias and variance of nonparametric smoothers in Table 5.1 (See Hardle 1989 and Fan 1992). We also give in Table 5.2 the results stated above (but not yet proved) about the MREWLE smoothers.

Tables 5.1 and 5.2 suggest that the results for the MREWLE smoothers are similar to their corresponding nonparametric regression smoothers. The variances of the MREWLE smoothers would seem to depend on Fisher information function while the

Table 5.1: Pointwise Biases and Variances of Regression Smoothers

Method	Bias	Variance
Nadaraya-Watson	$(\frac{1}{2}\theta''(x) + \frac{\theta'(x)g'(x)}{g(x)})h_n^2 c_K$	$\frac{\sigma^2(x)d_K}{nh_n g(x)}$
Gasser-Muller	$\frac{1}{2}\theta''(x)h_n^2 c_K$	$\frac{3\sigma^2(x)d_K}{2nh_n g(x)}$
K-NN weights	$\frac{\theta''(x)g(x)+2\theta'(x)g)}{8g^3(x)} \cdot \frac{k}{n}c_K$	$\frac{2\sigma^2(x)d_K}{k}$
Local linear smoother	$\frac{1}{2}\theta''(x)h_n^2 c_K$	$\frac{\sigma^2(x)}{2nh_n g(x)}$

Table 5.2: Conjectured Pointwise Biases and Variances of the MREWLE Smoothers

Method	Bias	Variance
Nadaraya-Watson	$(\frac{1}{2}B''(x) + \frac{B'(x)g'(x)}{g(x)})h_n^2 c_K / I\{\theta(x)\}$	$\frac{d_K}{nh_n g(x) I\{\theta(x)\}}$
Gasser-Muller	$\frac{1}{2}B''(x)h_n^2 c_K / I\{\theta(x)\}$	$\frac{3d_K}{2nh_n g(x) I\{\theta(x)\}}$
K-NN weights	$\frac{B''(x)g(x)+2B'(x)g'(x)}{8g^3(x)I\{\theta(x)\}} \cdot \frac{k}{n}c_K$	$\frac{2d_K}{kI\{\theta(x)\}}$
Local linear smoother	$\frac{1}{2}B''(x)h_n^2 c_K / I\{\theta(x)\}$	$\frac{d_K}{2nh_n g(x) I\{\theta(x)\}}$

variances of nonparametric regression smoothers are based on the variance function. This suggests that the MREWLE smoothers usually have smaller variances. But the biases of the MREWLE smoothers depend on the bias function and Fisher information function while the biases of the nonparametric regression smoothers depend on the parameter function $\theta(x)$. It is hard to compare them. Our tentative findings remain to be confirmed by rigorous analysis.

In nonparametric regression models, the comparison of these four smoothers has been considered by several authors (see Chu and Marron 1991; Fan 1992; Hardle 1989; Mack and Muller 1988; and among others).

Now we compare again in a heuristic fashion, the four MREWLE smoothers. The Nadaraya-Watson MREWLE seems essentially similar to the k-NN MREWLE. By

choosing $k = 2g(x)nh_n$, they have the same apparent variance and bias. The bias of the Nadaraya-Watson MREWLE depends on both $B''_x(x)$ and $B'_x(x)g'(x)/g(x)$. Keeping $B''_x(x)$ fixed, we first remark that in a highly clustered design where $\mid g'(x)/g(x) \mid$ is large, the bias of the Nadaraya-Watson MREWLE is large. Note also that when $\mid B'_x(x) \mid$ is large, so is the bias of that estimator. The Gasser-Muller MREWLE appears to have an asymptotic variance 3/2 times as large as that of that of the Nadaraya-Watson MREWLE. On the other hand the bias of the Gasser-Muller MREWLE seems simpler; it does not share the drawbacks mentioned above. The locally linear smoother MREWLE appears to overcome the disadvantages of these two methods. As noted above, the biases of the MREWLE smoothers seem to depend on the bias function, not the parameter function. This makes it hard to compare with other methods. In Chapter 7, we propose a MREWLE which have both simple bias (depends on the parameter function θ) and small variance.

5.4 Bandwidth Selection

For generalized smoothing models, the selection of weights to construct the REWL is analogous to getting a proper bandwidth. It is well known that the bandwidth plays a very important role in the trade-off between reducing bias and variance in nonparametric regression. Often the user will be able to choose the bandwidth satisfactorily by eye with interactive graphics. Moreover it is also desirable to have a reliable data–driven rule for selecting the value of h. Here we list some possible methods.

Cross-Validation (Wahba and Wold, 1975) may be used to select the bandwidth, in which case h is chosen to maximize

$$\sum \log f(Y_i, \hat{\theta}_j).$$

Here $\hat{\theta}_j$ is the maximizer with respect to θ of

$$\sum_{i \neq j} p_{ni}(x) \log f(Y_i, \theta).$$

This is a global bandwidth selection procedure, whose asymptotic optimality properties are not known.

Alternatively the bandwidth may be selected by the 'plug-in' procedure, based on the asymptotic expansion of the squared error for kernel smoothers:

$$MSE = \frac{n^{-1}h^{-1}d_K}{I\{\theta(x)\}g(x)} + \frac{h^4[c_K\{B''(x)/2 + B'(x)(g'/g)(x)\}]^2}{I^2(\theta(x))}.$$

An 'optimal' bandwidth minimizing this expression would be

$$h_n = \{\frac{d_K I^3(\theta(x))/g(x)}{4[c_K(B''(x)/2 + B'(x)(g'/g)(x)]^2}\}^{1/5} n^{-1/5}.$$

This bandwidth is proportional to $n^{-1/5}$ with constants depending on the unknown $I\{\theta(x)\}$, $B'(x)$, $B''(x)$ and so on.

This has been shown to be more stable in both theoretical and practical performance in NR (see e.g. Park and Marrron 1990). Indeed, Fan and Marron (1992) show that, in the density estimation case, the 'plug-in' selector is an asymptotically efficient method from a semiparametric point of view. We expect the 'plug-in' procedure will perform as well in the MREWLE case.

5.5 Simulation

We have shown, via asymptotics, that the MREWLE possesses a number of desirable properties. Now we use three simulated experiments to illustrate its finite sample behavior. Two methods are considered here: the Nadaraya-Watson MREWLE and the locally linear smoother, MREWLE.

Simulation 1. A random sample of size n is simulated from the model

$$Y = X(1 - X) + 0.01\epsilon, \tag{5.21}$$

with $\epsilon \sim$ Cauchy(0,1) independent of $X \sim$ Uniform(0,1). The nonparametric regression kernel smoothers do not work here, because the mean and variance of the Cauchy(0,1) do not exist. We use the MREWLE smoothers here. A typical realization when $n = 200$ is shown in Figure 5.1. The bandwidth used here is $h = 0.1$. Gaussian kernel function is used here and in the sequel. The Nadaraya-Watson MREWLE based on five simulations from Model (5.21) is shown in Figure 5.2. Figure 5.3 shows the locally linear smoother MREWLE based on five simulations from Model (5.21).

From the above Figures, both the Nadaraya-Watson MREWLE and the locally linear smoother MREWLE perform reasonable. The Nadaraya-Watson MREWLE has large boundary effects, and may require boundary modifications. But the locally linear smoother MREWLE seems accurate over the whole interval.

Simulation 2. A random sample of size n is simulated from the model

$$Y = \sin(X) + 0.01\epsilon, \tag{5.22}$$

with $\epsilon \sim$ Cauchy(0,1) independent of $X \sim N(0, \sigma^2)$. When σ is small, the quantity $|g'(x)/g(x)|$ gets large. Thus we anticipate that the Nadaraya-Watson MREWLE does not behave well for small σ.

We estimate the parameter function in the interval $x \in [-2\sigma, 2\sigma]$, two standard deviations away from its normal mean. Figure 5.4 plots the estimates for the case $n = 200$ and $\sigma = 0.25$. The bandwidth is $h = .15$. The Nadaraya-Watson MREWLE has a large bias. The locally linear smoother MREWLE provides a suitable estimation.

Simulation 3. Instead of estimating the location parameter function, consider the fol-

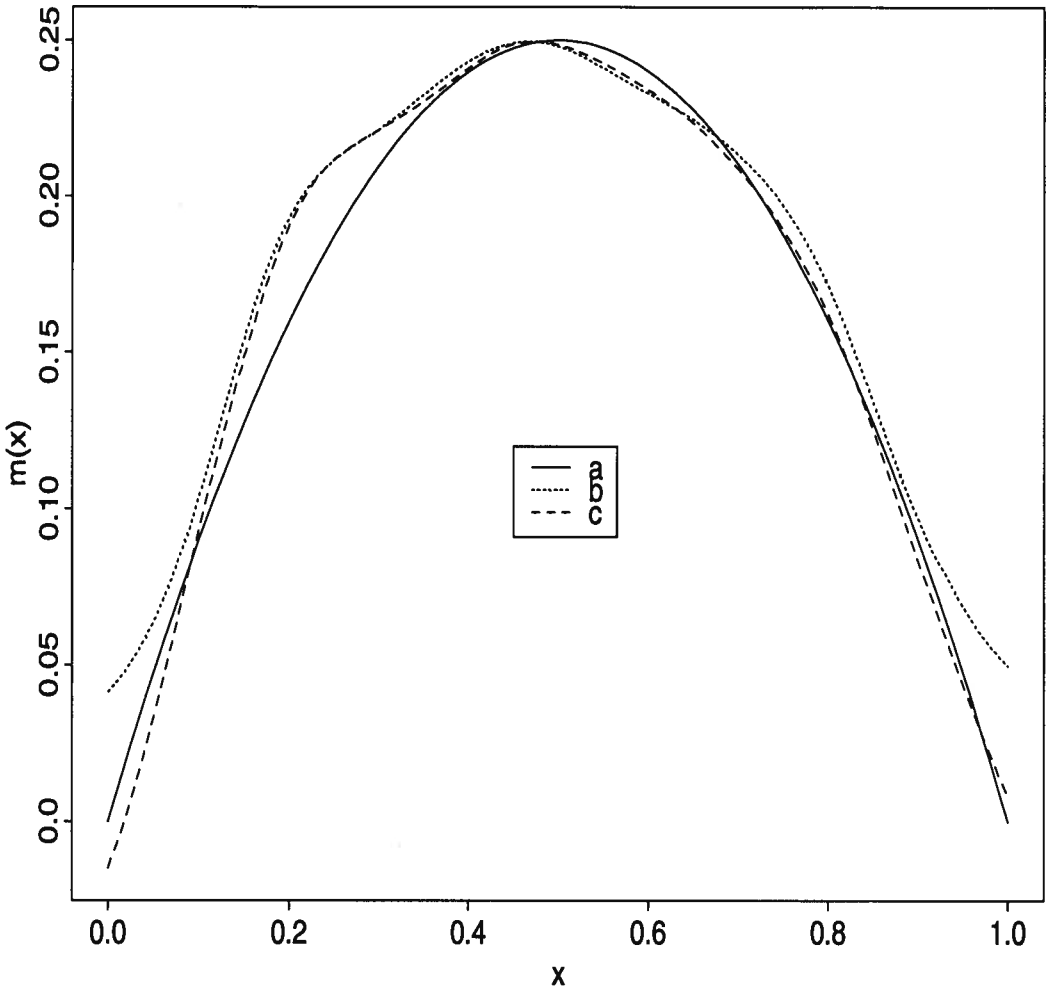

Figure 5.1: *A comparison of the Nadaraya-Watson MREWLE with the locally linear smoother MREWLE from Model (5.21) with* $n = 200$. *The true curve is a, the Nadaraya-Watson MREWLE, b, and the locally linear smoother MREWLE, c.*

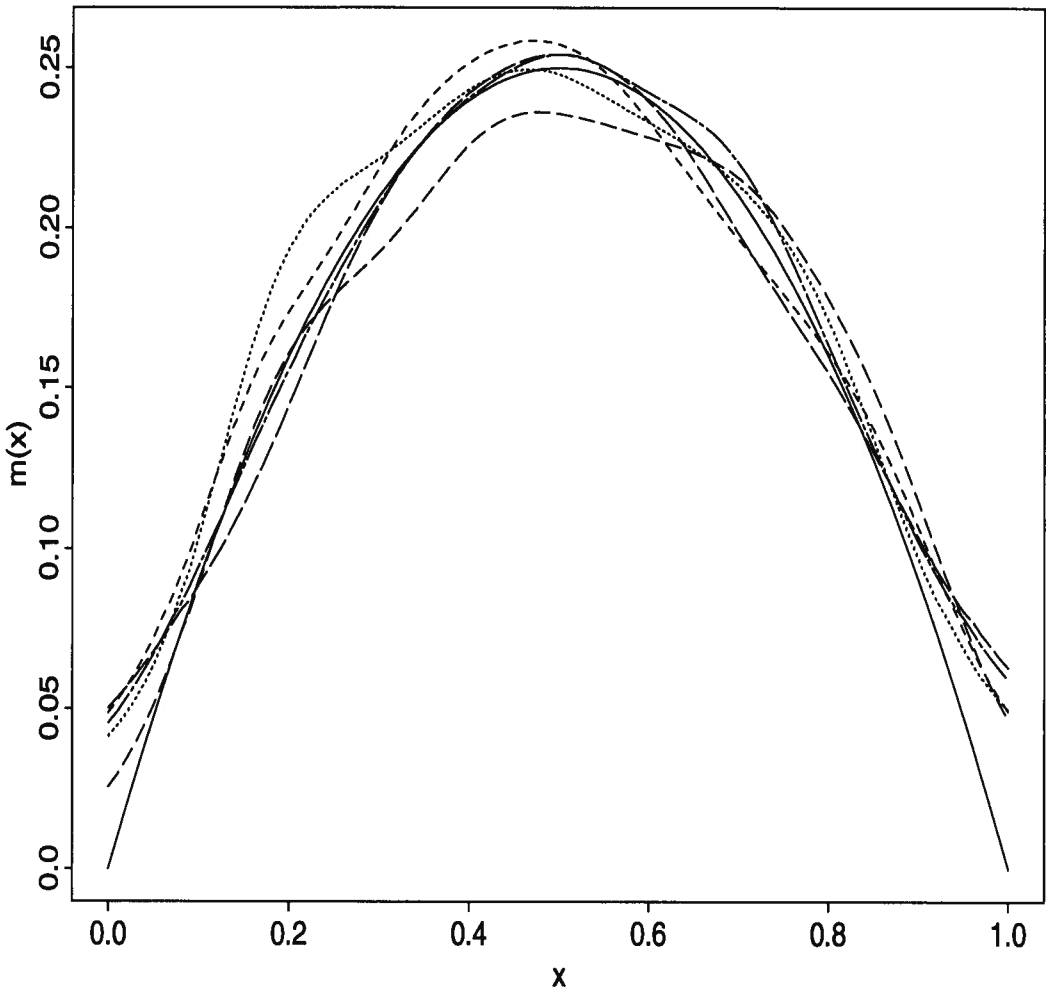

Figure 5.2: *The Nadaraya-Watson MREWLE based on five simulations from Model (5.21) with n=200.*

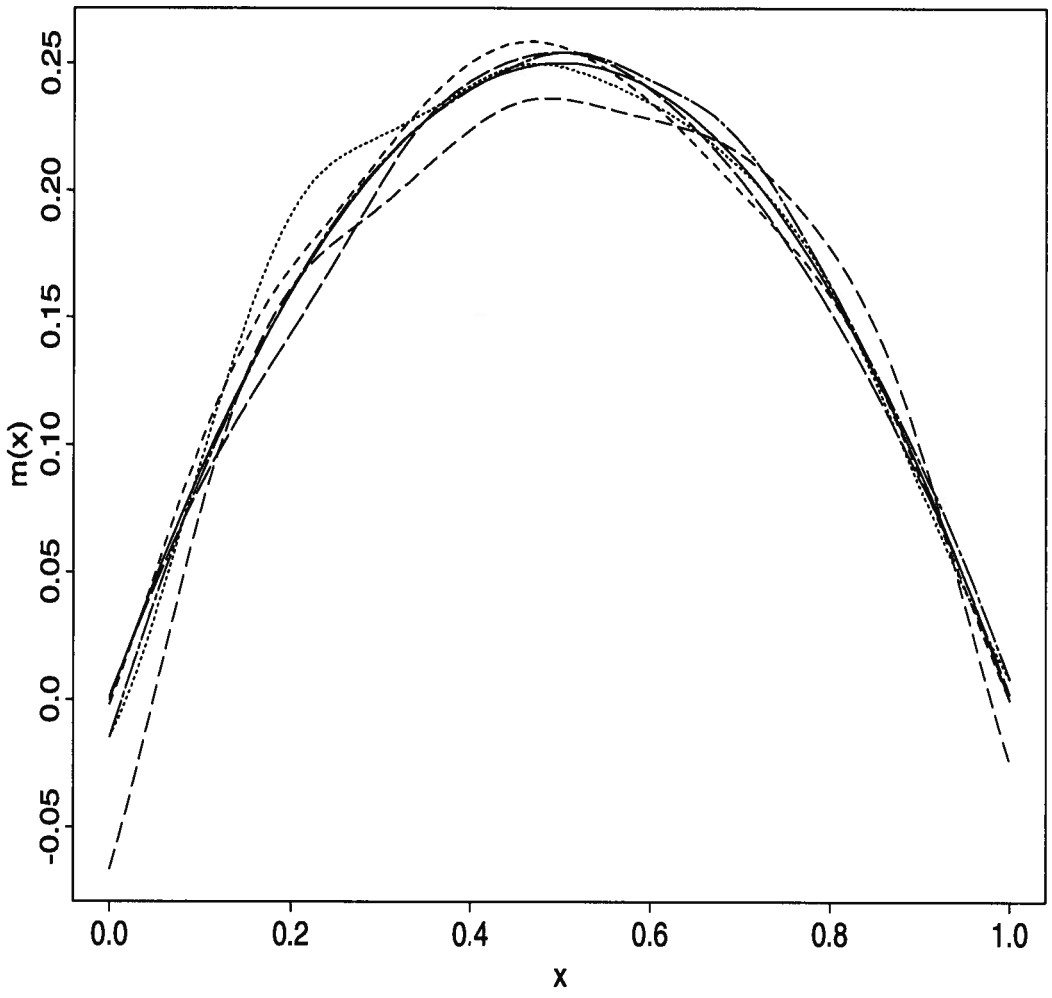

Figure 5.3: *The locally linear smoother MREWLE based on five simulations from Model (5.21) with n=200.*

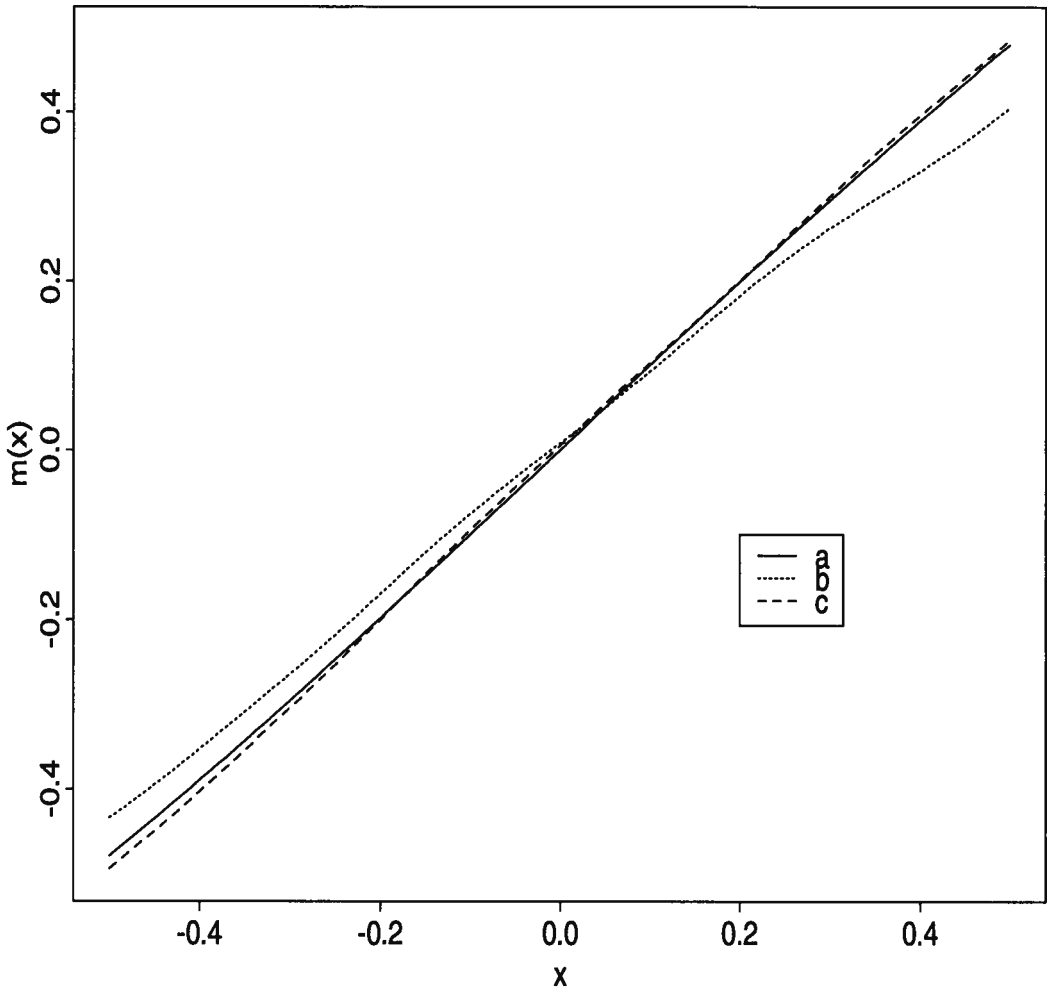

Figure 5.4: *A comparison of the Nadaraya-Watson MREWLE with the locally linear smoother MREWLE from Model (5.22) with* $n = 200$. *The true curve is a, the Nadaraya-Watson MREWLE, b, and the locally linear smoother MREWLE, c.*

lowing model:

$$f_{Y|X=x}(y) \sim \frac{\beta(x)}{\alpha} y^{\beta(x)-1} \exp(\frac{-y^{\beta(x)}}{\alpha}) \tag{5.23}$$

with $\alpha = 1$, $\beta(x) = 2 + 4x(1-x)$ and $X \sim$ Uniform(0,1). A typical realization with $n = 200$ is shown in Figure 5.5. The bandwidth used here is $h = 0.25$. The Nadaraya-Watson MREWLE has a large bias, while the locally linear smoother MREWLE performs reasonably well in tails.

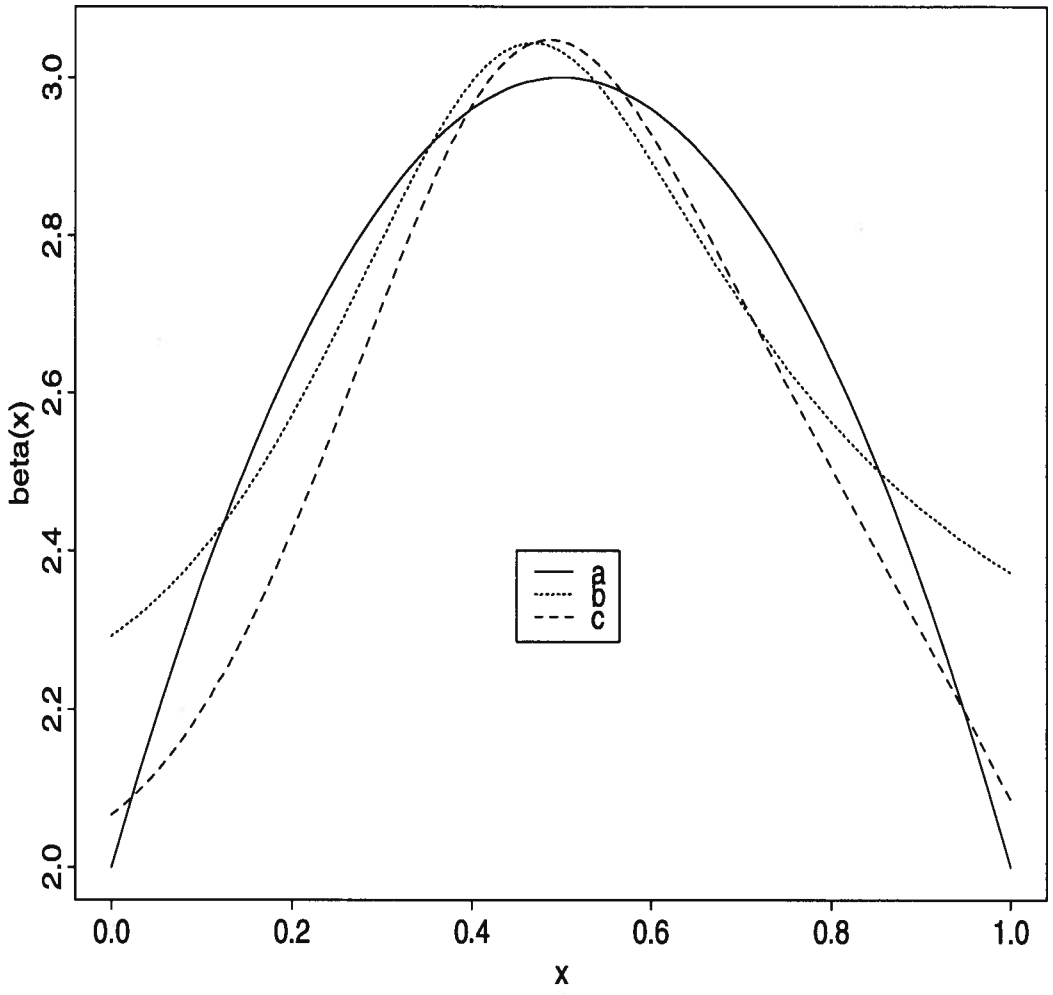

Figure 5.5: *A comparison of the Nadaraya-Watson MREWLE with the locally linear smoother MREWLE from Model (5.23) with* $n = 200$. *The true curve is a, the Nadaraya-Watson MREWLE, b, and the locally linear smoother MREWLE, c.*

Chapter 6

A Relevance Weighted Nonparametric Quantile Estimator

6.1 Introduction

This Chapter concerns situations in which a sample $X_1 = x_1, \ldots, X_n = x_n$ of independent observations are drawn from populations with different CDF's $F_1, \ldots, F_n$, respectively. Inference is about an attribute θ_0 of another population with CDF F_0; an observation may be available from the latter population as well. In this Chapter θ_0 will be a quantile of F_0; elsewhere we address other problems of interest within the same general framework. The $\{F_i\}$ are unknown, so that we are in the nonparametric case.

The special character of the problems investigated in this problem derives from the belief that there is relevant information in the X_i, $i = 1, \ldots, n$ about θ_0. However this information is deemed to be "inexact". By this we mean it cannot be translated into a prior distribution from which a marginal posterior distribution for θ_0 could be constructed. And we mean there are no known structural constraints among the attributes of the various populations to force the $x_1, \ldots, x_n$ into inferences about θ_0.

The example given below illustrates the problem. That example reflects the situation underlying nonparametric regression. In fact, our approach may be thought of as generalized smoothing. In nonparametric smoothing it leads to locally weighted regression quantile estimators $\hat{\theta}_0(t), -\infty < t < \infty$, for even rough regression quantile functions $\theta_0(t), \ -\infty < t < \infty$, if the relevance of the $x(t_i)$'s corresponding to t_i's remote from t can be ascertained. Our method bears on other problems like those of meta analysis where there is no well defined underlying mathematical structure. In Section 6.8, we briefly discuss linkages with standard statistical methods.

As we have noted in Chapter 2, there are problems where relevant sample information about θ may be used to advantage. Yet there does not seem to be a general theory underlying such problems. How to use the relevant sample information like that encountered in metaanalysis, thus becomes an important topic which seems to have been addressed largely on a piecemeal basis.

In Chapter 3, we introduced the "relevance weighted likelihood" (REWL) as a general device for using relevant sample information in parameter estimation. However, when we cannot specify the population CDF's parametric form the REWL based theory is of no avail. To cope with such problems we introduce in this paper a nonparametric but general theory based on an extension of the empirical distribution function which we call the REWED ("Relevance Weighted Empirical Distribution"). We then tackle the problem of quantile estimation within that framework which is examplified by the following example.

Example 6.1 Distribution Smoothing. *Let $X(t_i)$ have distribution function F_{t_i}, $0 \leq t_i \leq 1$, $i = 1, \ldots, n$. Assume that $X(t_1)$, ... , $X(t_n)$ are independent and that F_t changes smoothly with t, "smooth" meaning $\sup_x |F_t(x) - F_{t+\triangle(t)}(x)| \to 0$ as $\triangle(t) \to 0$.*

We seek to estimate F_t for a fixed t.

In general, let F_0 be an unknown distribution function describing the population of interest. The classical paradigm would assume independent and identically distributed (hereafter iid) observations from F_0. Here instead, only observations from other populations described by CDF's F_i, $i = 1, 2, ..., n$ are available. If we believe the $\{F_i\}$, $i = 1, 2, ..., n$ are related to F_0, $x_1, x_2, \ ... \ , x_n$ may be used for inference about attributes of F_0. The question is how.

In this Chapter, our answer to this question uses the REWED, defined in Section 6.2. From the REWED we can construct moment estimators for parameters defined in terms of the moments of F_0 [this idea will be discussed elsewhere]. But here we consider only the estimation of the quantiles of F_0.

In Section 6.2, the REW quantile estimator will be defined in addition to REWED for the problem identified by the last example. And we will offer generalizations along with some examples.

Strong consistency of the REWED and the quantile estimators are stated in Section 6.3 and proved in Section 6.9 under mild conditions. These results generalize the results for iid sampling. Some other asymptotic properties are given in Section 6.3.

R.R. Bahadur (1966) gives a useful asymptotic representation of the sample quantile as a simple sum of random variables by using the empirical distribution function. We give a generalization of Bahadur's results for general weights $\{p_{ni}\}$ in the non-iid case. This is the subject of Section 6.4.

We discuss the asymptotic normality of the REW quantile estimator in Section 6.5. In

Section 6.6, we apply the theory of this paper. We get reasonable estimators for location parameters for several distributions. By comparing them with the weighted sample mean, we find their asymptotic relative efficiency (ARE) in the iid case to be fairly high. Section 6.7 presents the results of a simulation study, using the REW quantile estimator for the nonparametric smoothing model. The proofs of our Theorems appear in Section 6.9.

6.2 The REWED and REW Quantile Estimation

Let us reconsider Example 6.1. Because $t \to F_t$ changes smoothly, we could hypothetically use $\sum_{i=1}^n p_i F_{t_i}(x)$ to approximate $F_t(x)$. The choice of the weights p_i would depend on the perceived relationship between F_{t_i} and $F_t(x)$. [The $\{p_i\}$ might plausibly be generated from a kernel.] But the $\{F_{t_i}\}_1^n$ are unknown. So instead we must use the data, $X(t_i)$, $i = 1, \ldots, n$ to estimate $F_t(x)$, say by $\sum_{i=1}^n p_i I(X(t_i) \leq x)$, $I(\cdot)$ being an indicator function. This empirical distribution we will call the *relevance weighted empirical distribution (REWED).*

Estimating $F_t(x)$ by the REWED results in two errors from: (i) using $\sum_{i=1}^n p_i F_{t_i}(x)$ to approximate $F_t(x)$; (ii) using $\sum_{i=1}^n p_i I(X(t_i) \leq x)$ to estimate $\sum_{i=1}^n p_i F_{t_i}(x)$. Much of this Chapter will be concerned with (ii).

To generalize the ideas in the above example, let

$$\mathbf{X}_n \stackrel{def}{=} [X_{n1}, \ldots, X_{nn}];\ n \geq 1$$

be a triangular array of row-independent random variables with associated array of distribution functions, $\mathbf{F}_n \stackrel{def}{=} [F_{n1}, \ldots, F_{nn}];\ n \geq 1$ and nonnegative constants

$$\mathbf{p}_n \stackrel{def}{=} [p_{n1}, \ldots, p_{nn}];\ n \geq 1$$

satisfying $\sum p_{ni} = 1$. Define:

- the *relevance weighted empirical distribution function* (REWED) by

$$F_n(x) = \sum_{i=1}^{n} p_{ni} I(X_{ni} \leq x);$$

- the *relevance weighted average distribution function* (REWADF) for $-\infty < x < \infty$ by

$$\bar{F}_n(x) = \sum_{i=1}^{n} p_{ni} F_{ni}(x);$$

- the pth quantile of $\bar{F}_n$ by

$$\xi_{p(n)} = inf\{x : \bar{F}_n(x) \geq p\} \quad 0 < p < 1;$$

- the pth *relevance weighted quantile* (REWQ) estimator by

$$\hat{\xi}_{np} = inf\{x : F_n(x) \geq p\}$$

for a sample $\{X_{n1}, ..., X_{nn}\}$.

To illustrate the use of these REW quantile estimators, we offer the following example.

Example 6.2 Nonparametric Regression. *Let*

$$Y_i = f(x_i) + \epsilon_i \qquad x_i \in [a, b] \quad i = 1,\ 2,\ \ldots\ ,\ n;$$

here $\epsilon_1,\ \epsilon_2,\ \ldots\ ,\ \epsilon_n$ *are iid, symmetric,* $E(\epsilon_i) = 0$ *for all* i*, and* $f(x)$ *is a smooth function. To estimate* $f(x)$ *we may use the median of the REWED. This kind of estimator is usually robust and it often quite efficient.*

6.3 Strong Consistency of REW Quantile Estimators

In this section, we describe strong and uniform consistency properties of the REWED. Then we describe the strong consistency of the quantile estimators derived from the REWED. In the following discussion, we assume that F is the CDF of interest and ξ_p its pth quantile.

Theorem 6.1 *(Strong Consistency of $F_n(x)$). a) Suppose $\sum_{n=1}^{\infty} \exp(-\epsilon^2 K_n) < \infty$ for all $\epsilon > 0$, where $K_n = (\sum_{i=1}^{n} p_{ni}^2)^{-1}$. Then $|F_n(x) - \bar{F}_n(x)| \to 0$ a.s. for all x.*

b) Further, if $|F(x) - \bar{F}_n(x)| \to 0$ for all x, then $|F_n(x) - F(x)| \to 0$ a.s. for all x.

Corollary 1. If $\log(n)/K_n = o(1)$, then

$$|F_n(x) - \bar{F}_n(x)| \to 0 \quad a.s. \quad for \ every \ x. \ \square$$

The hypothesis of the theorem is easily satisfied. If, for example, $\max_i\{p_{ni}\} = o[\{\log(n)\}^{-1}]$, then the hypothesis is satisfied. The assumption $|F(x) - \bar{F}_n(x)| \to 0$ for all x is essential; without this, we cannot get a consistent estimator of the CDF $F(x)$. Qualitatively this condition is the one which gives operational meaning to the notion of "relevance weights".

Theorem 6.2 *(Uniformly Strong Consistency of $F_n(x)$) a) Under the hypothesis of Theorem 6.1 a), and the further assumptions that (i) $sup_{x,n} \bar{f}_n(x)$ is bounded and, (ii) $\limsup_{M\to\infty} \sup_n\{(1 - \bar{F}_n(M)), \bar{F}_n(-M)\} \to 0$, where $\bar{f}_n(x)$ is the derivative of $\bar{F}_n(x)$, then*

$$\sup_x |F_n(x) - \bar{F}_n(x)| \to 0, \quad a.s..$$

b) Further, if $\sup_x |F(x) - \bar{F}_n(x)| \to 0$, *then*

$$\sup_x |F_n(x) - F(x)| \to 0, \quad a.s..$$

When the distributions underlying our investigation derive from the same family, the conditions of the last theorem are usually satisfied.

Theorem 6.3 *(Strong Consistency of* $\hat{\xi}_{np}$*) Under the conditions of Theorem 6.2 a), suppose* $x = \xi_{p(n)}$ *solves uniquely the inequalities* $\bar{F}_n(x-) \leq p \leq \bar{F}_n(x)$. *Then*

$$\hat{\xi}_{np} - \xi_{p(n)} \to 0 \quad a.s. \; for \; n \to \infty.$$

b) Further, if $\sup_x |F(x) - \bar{F}_n(x)| \to 0$, *then*

$$\hat{\xi}_{np} - \xi_p \to 0 \quad a.s. \; for \; n \to \infty.$$

The uniqueness condition on $\xi_{p(n)}$ imposed in the last theorem cannot be dropped.

We finish this section with the following theorem giving a probabilistic inequality for quantile estimators. This theorem will be used in next section.

Theorem 6.4 *Suppose* $x = \xi_{p(n)}$ *solves uniquely the inequalities* $\bar{F}_n(x-) \leq p \leq \bar{F}_n(x)$ *for any given* $p \in (0,1)$. *Then*

$$P(|\hat{\xi}_{np} - \xi_{p(n)}| > \epsilon) \leq 2\exp\{-2\delta_\epsilon^2(n)K_n\}$$

for every $\epsilon > 0$ *and* n, *where* $\delta_\epsilon(n) = \min\{\bar{F}_n(\xi_{p(n)} + \epsilon) - p, p - \bar{F}_n(\xi_{p(n)} - \epsilon)\}$.

The last theorem shows $P(|\hat{\xi}_{np} - \xi_{p(n)}| > \epsilon)$ converges to 0 exponentially fast. The value of ϵ (> 0) may depend upon K_n if desired. These bounds hold for each $n = 1, 2, \ldots$ and so may be applied for any fixed n as well as for asymptotic analysis.

6.4 Asymptotic Representation Theory.

For the case of iid data and $p_{ni} = 1/n,\ i = 1, \ldots,\ n$, Bahadur (1966) expresses sample quantiles asymptotically as sums of independent random variables by representing them as a linear transform of the sample distribution function evaluated at the relevant quantile. From these representations, a number of important properties ensue. (see Bahadur 1966 and Serfling 1980 for details). We now generalize this asymptotic representation to the cases of non-iid observations and general p_{ni}.

Theorem 6.5 *Let* $0 < p < 0$ *and* $m_n = \max_{1 \le i \le n}\{p_{ni}\}$. *Suppose:*

1. $\bar{F}_n$ *has bounded second derivative in the neighbourhood of* $\xi_{p(n)}$ *with* $\bar{F}_n'(\xi_{p(n)}) = \bar{f}_n(\xi_{p(n)})$;
2. *there exists* $c > 0$, *such that* $\inf_n \bar{f}_n(\xi_{p(n)}) > c$;
3. *there exists* $c^* > 0$, *such that* $\sum_{n=1}^{\infty} K_n^{-c^*} \le \infty$;
4. F_{ni} *has a uniformly bounded first derivative in the neighbourhood of* $\xi_{p(n)}$;
5. $m_n = o\{K_n^{-3/4}(\log K_n)^{-1/4}\}$.

Then

$$\hat{\xi}_{np} = \xi_{p(n)} + \frac{p - F_n(\xi_{p(n)})}{\bar{f}_n(\xi_{p(n)})} + R_n.$$

where

$$R_n = O\{K_n^{-3/4}(\log K_n)^{3/4}\}, \quad n \to \infty, \quad \textit{with probability } 1.$$

The Bahadur representation is a special case of this theorem suggesting the result of our theorem may be fairly accurate, that is hard to improve upon.

The distribution of REW sample quantile is usually hard to find, but the REW sample distribution relatively easy. By this theorem, we can use the REW sample distribution function evaluated at the relevant quantile to study the REW sample quantile asymptotically. A simple example is that we can use this representation to prove quite easily the asymptotic normality of the REW sample quantile.

6.5 Asymptotic Normality of $\hat{\xi}_{np}$

Except for the case of iid random variables, we cannot always find the exact distribution of $\hat{\xi}_{np}$. The asymptotic distribution of $\hat{\xi}_{np}$ given in the following theorem may therefore be useful.

Theorem 6.6 *Let* $0 < p < 1$ *and* $V_n = \sum_{i=1}^{n} p_{ni}^2 F_{ni}(\xi_{p(n)})(1 - F_{ni}(\xi_{p(n)}))$. *Assume* $\bar{F}_n$ *is differentiable at* $\xi_{p(n)}$, $\inf_n \bar{F}_n'(\xi_{p(n)}) > c > 0$ *and* $\max_{1 \le i \le n}(p_{ni} V_n^{-1/2}) \to 0$ *as* $n \to \infty$. *Then*

$$\lim_{n \to \infty} P\{\bar{f}_n(\xi_{p(n)})(\hat{\xi}_{np} - \xi_{p(n)}) V_n^{-1/2} \le t\} = \Phi(t)$$

where $\Phi(t)$ *is the distribution function of* $N(0,1)$.

6.6 Applications

In this section, we use REW sample quantiles to estimate location parameters, and compare these estimators with the weighted sample mean estimators.

Example 6.3 *Let* $\{X_i\}$ *be an independent sample with* $X_i \sim N(\mu, \sigma_i^2)$ $i = 1, \ldots, n$, *the* σ_i^2 *being known and* μ *unknown. An estimate of* μ *is required.*

Analysis of Example 6.3 Using the weighted sample mean to estimate μ seems natural:

$$\hat{\mu} = \sum_{i=1}^{n} c_i X_i, \quad \sum_{i=1}^{n} c_i = 1 \text{ and } c_i \geq 0.$$

We easily deduce that $c_i = [\sigma_i^{-2}]/[\sum_{j=1}^{n} \sigma_j^{-2}]$ minimizes the mean squared error. Then

$$\hat{\mu} \sim AN\{\mu, (\sum_{i=1}^{n} 1/\sigma_i^2)^{-1}\}.$$

Now let us try using the median to estimate μ. Let F_{ni} be the distribution of X_i and $\bar{F}_n = \sum_{i=1}^{n} p_{ni} F_{ni}$. The median of $\bar{F}_n$ is μ and we use the sample median $\hat{\xi}_{med}$ to estimate μ. By the results of Section 6.5, we get

$$\hat{\xi}_{med} \sim AN\{\mu, \frac{\sum_{i=1}^{n} p_{ni}^2}{4(\sum_{i=1}^{n} p_{ni} f_{ni}(\mu))^2}\};$$

here $f_{ni}(\mu) = (\sqrt{2\pi}\sigma_i)^{-1}$.

We want to minimize the variance of the asymptotic distribution subject to $\sum_{i=1}^{n} p_{ni} = 1$. We easily obtain $p_{ni} = \frac{1/\sigma_i}{\sum_{j=1}^{n} 1/\sigma_j}$.

The asymptotic relative efficiency of these two estimators is

$$ARE(\hat{\mu}, \hat{\xi}_{med}) = \frac{2}{\pi}$$

Remarks 1

1. *For the iid normal case, the ARE of the sample mean estimator relative to the sample median estimator is $\frac{2}{\pi}$. Here we have proved that when the samples are from normal distributions with the same mean, but different variances, the ARE of the weighted sample mean estimator relative to the weighted sample median estimator yields the same value $\frac{2}{\pi}$.*

2. *The weights used in the sample mean are different from the weights used in the sample median. We only compare the two best estimators here. If we use the same weights, the ARE can be larger or smaller than $\frac{2}{\pi}$.*

3. *The weighted sample median should be more robust than the weighted sample mean.*

Example 6.4 *Consider the double exponential family. Assume the density of* X_i *to be* $1/2r_i \exp(-|x-\mu|/r_i)$*; the* r_i *are known while* μ *is unknown* $i = 1, \ldots, n$*. We again use the weighted sample mean and weighted sample median of Example 6.3 to estimate* μ*.*

Analysis of Example 6.4. Choose $c_i = (r_i^{-2})/(\sum_{j=1}^n r_j^{-2})$ to minimize the mean squared error. Then

$$\hat{\mu} \sim AN\{\mu, 2/(\sum_{i=1}^n 1/r_i^2)\}.$$

By choosing $p_{ni} = (1/r_i)/(\sum_{j=1}^n 1/r_j)$, we get the weighted sample median $\hat{\xi}_{med}$. From the results of Section 6.5,

$$\hat{\xi}_{med} \sim AN(\mu, \frac{1}{\sum_{i=1}^n 1/r_i^2}).$$

The asymptotic relative efficiency of these two estimators is

$$ARE(\hat{\mu}, \hat{\xi}_{med}) = 2.$$

Remarks 2 1. *The ARE of the best sample mean estimator relative to the best sample median estimator does not depend on the* $\{r_i\}$*'s. The median is a more efficient estimator.*

2. *As in the normal case, the weights used in the sample mean do not equal the weights used in the sample median.*

3. *The weighted sample median should be more robust.*

6.7 Simulation Study

We have shown via asymptotics that the REW quantile estimator possesses a number of desirable asymptotic properties. In this section we use two simulated examples to obtain insight into its performance with a finite sample.

The model in Example 6.2 is used in this simulation where we compare the REW quantile estimator with the Nadaraya-Watson estimate. The Gaussian kernel function is used to generate the relevance weights.

Simulation Study 1. A random sample of size n is simulated from the model

$$Y = X(1 - X) + \epsilon,$$

with $\epsilon \sim N(0, 0.5)$ independent of $X \sim U(0, 1)$. A typical realization when $n = 1000$ is shown in Figure 6.1. The bandwidth used here and in all subsequences is $h = 0.1$ (choosed by eye). Let us next add 50 outliers from $N(2, 0.5)$ to the simulation experiment just described. The result is shown in Figure 6.2. We have not used the boundary corrections.

Simulation Study 2. In the model of Simulation 1, instead of using the normal error, we get the ϵ from a double exponential distribution with $r = 0.1$. Figure 6.3 shows the results of a curve fit based on 100 simulated observations. The simulation results with 10 outliers from $N(-.5, 0.25)$ are shown in Figure 6.4.

For the data of Simulation 1 without outliers, the quantile curve estimate obtained by using the REW quantile estimator is shown in Figure 6.5 with 0.25 quantile.

The results of the simulation can be summarized as follows:

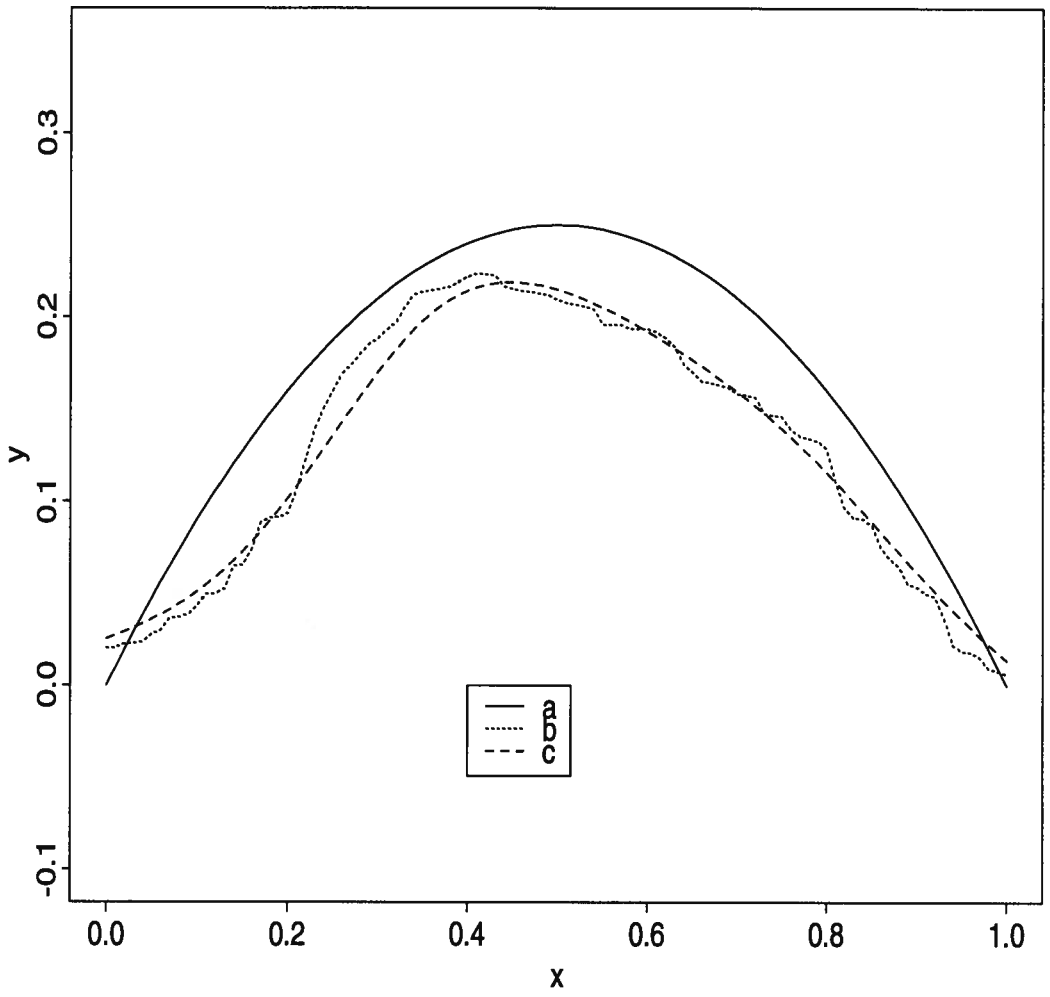

Figure 6.1: *A comparison of the Nadaraya-Watson estimate with REW quantile estimator. The model is* $Y = X * (1 - X) + \epsilon$, *where* X *is uniform (0,1) and* ϵ *is* $N(0, 0.5)$. *The sample size* $n = 1000$ *and the bandwidth,* $h = 0.1$. *The true curve is a, the REW quantile estimator, b, and the Nadaraya-Watson, c.*

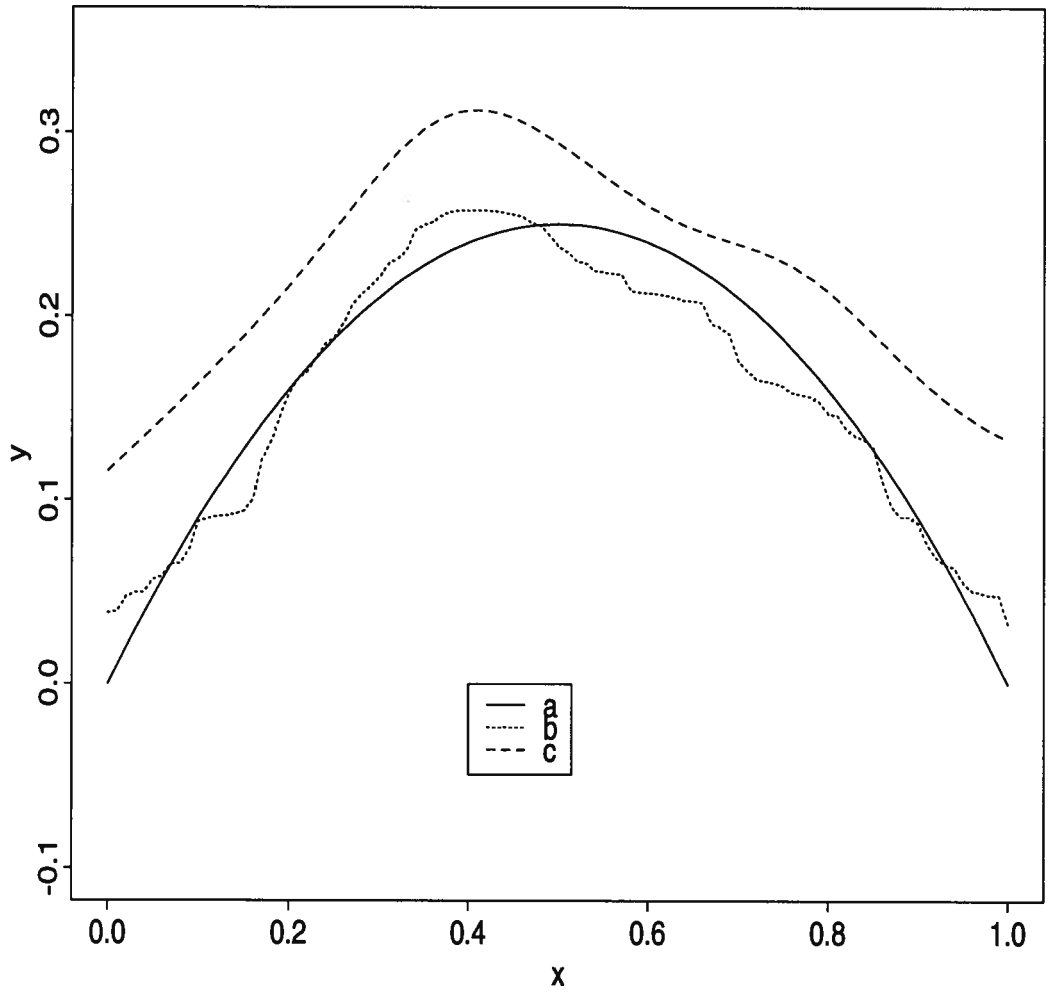

Figure 6.2: *A comparison of the Nadaraya-Watson estimate with REW quantile estimator with outliers. To the data depicted in Figure 6.1, we add* 50 ϵ*-outliers from* $N(2, 0.5)$. *The true curve is a, the REW quantile estimator, b, and the Nadaraya-Watson, c.*

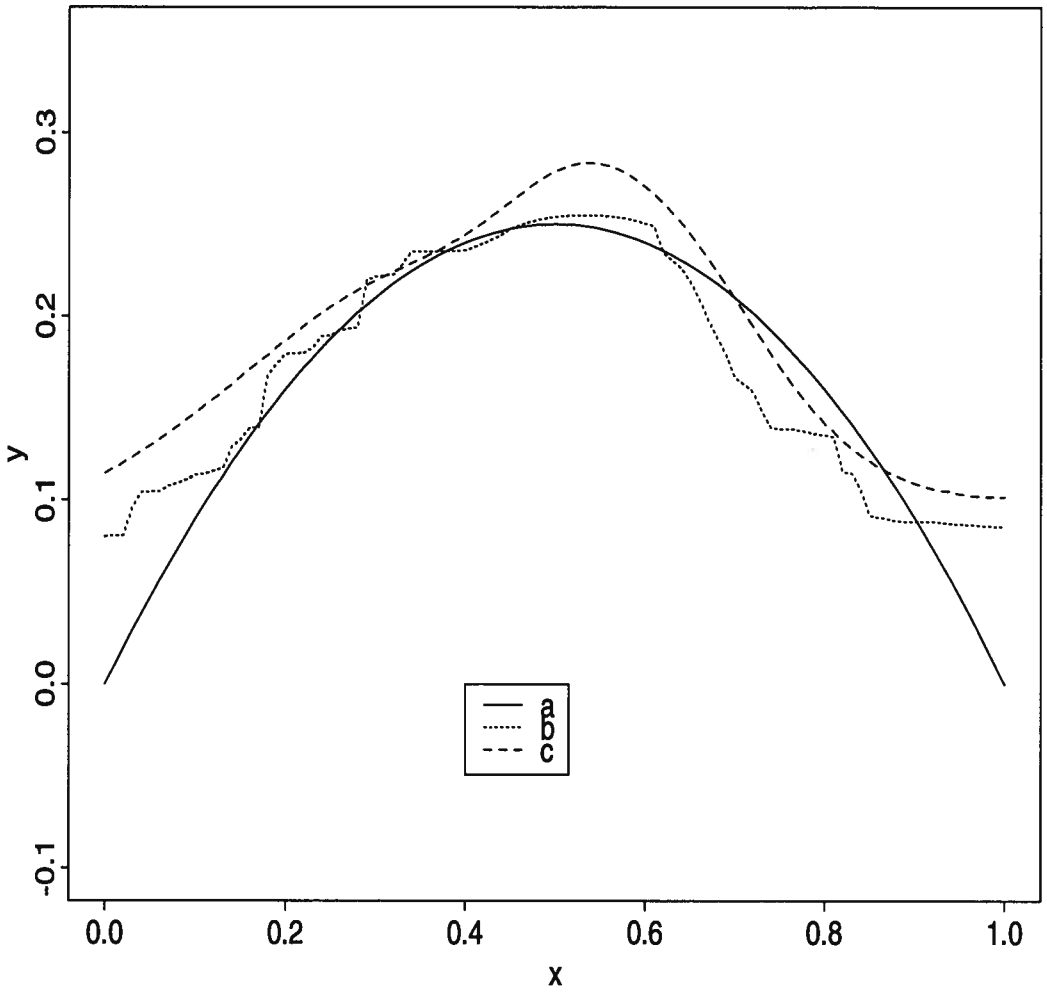

Figure 6.3: *A comparison of the Nadaraya-Watson estimate with REW quantile estimator. The model is* $Y = X*(1-X)+\epsilon$, *where* X *is from uniform (0,1) and* ϵ *from a double exponential distribution with* $r = 0.1$. *The sample size is* $n = 100$ *and the bandwidth,* $h = 0.1$. *The true curve is a, the REW quantile estimator, b, and the Nadaraya-Watson, c.*

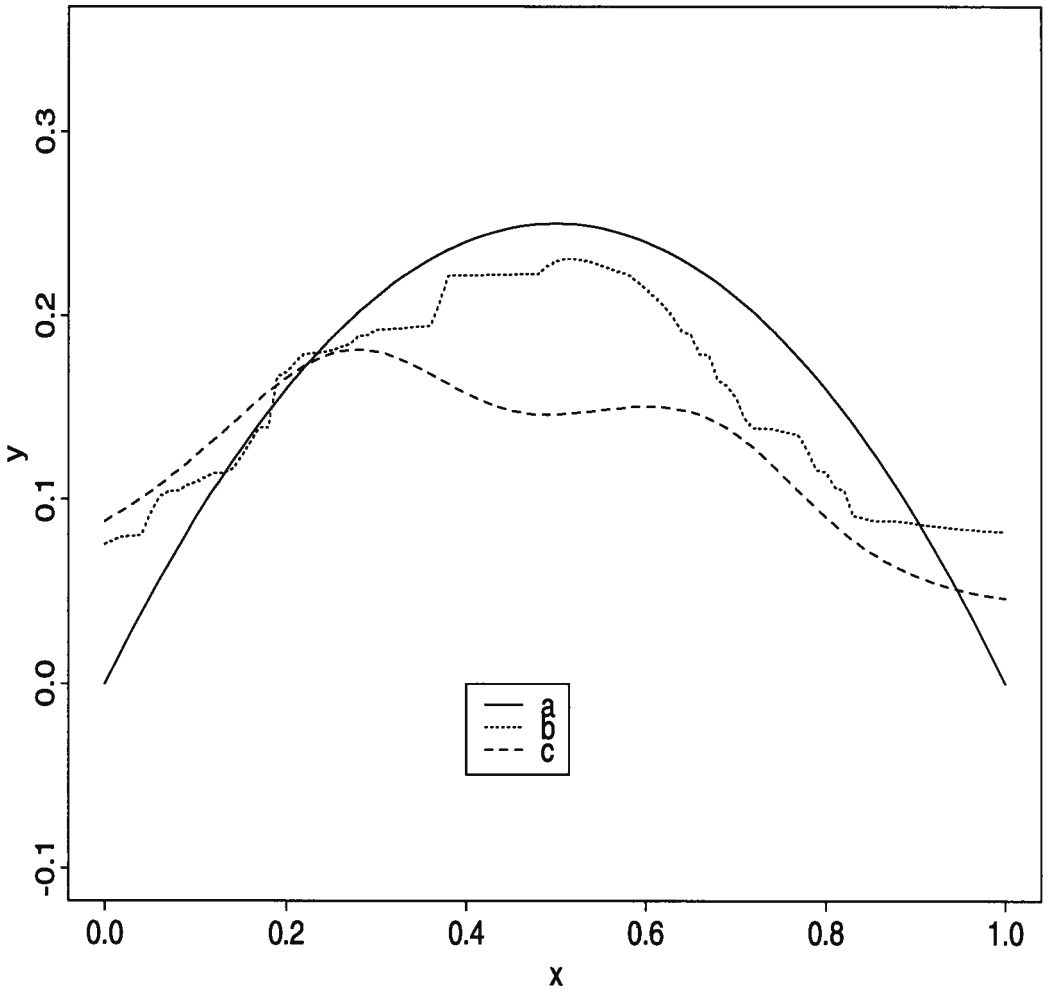

Figure 6.4: *A comparison of the Nadaraya-Watson estimate with REW quantile estimator with outliers. To the data depicted in Figure 6.3, we add* 10 ϵ*-outliers from* $N(-.5, .25)$. *The true curve is a, the REW quantile estimator, b, and the Nadaraya-Watson, c.*

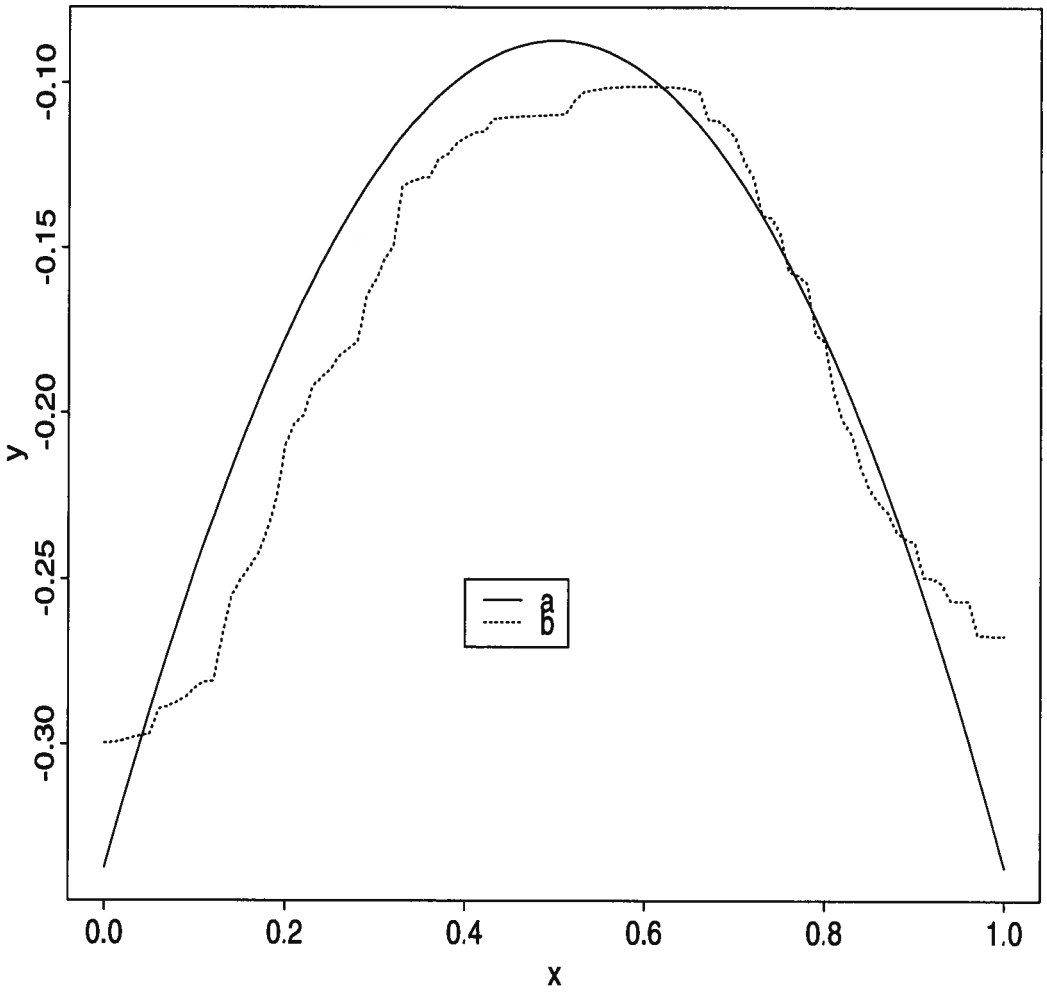

Figure 6.5: *A REW quantile estimator of a quantile curve. The .25 quantile curve is estimated for the data depicted in Figure 1. The true quantile curve is a, and the REW quantile estimator is b.*

1. In the model of Example 6.2, when the error has the double exponential distribution, the REW quantile estimator performs a little better than Nadaraya-Watson estimate, see Figure 6.3. Even when the error is normal, the REW quantile estimator performs about as well as the Nadaraya-Watson estimate (see Figure 6.1).

2. When the data have a small fraction of outliers, say about 5 or 10 percent, the REW quantile is robust (see Figures 6.2 and 6.4). By contrast, the Nadaraya-Watson estimator fails. This observation suggests we use the REW quantile estimator and Nadaraya-Watson estimate together to diagnose the model and determine if there are outliers in the data set. If the REW quantile estimator and Nadaraya-Watson estimate disagree, then we should reconsider the model and the outliers.

3. The REW quantile estimator seems promising judging from these simulation studies.

4. Computing the REW quantile curve estimator took about one minute in Simulation Study 1 using Splus in a Sun workstation.

6.8 Discussion

We have presented a general method for estimating a population quantile based on independent observations drawn from other related but not identical populations. We have shown the estimator to be strongly consistent and asymptotically normal under mild assumptions. Our method derives from a generalization of the empirical distribution (the REWED), and we have shown that the latter is also strongly consistent under certain conditions.

The context of our method includes that of nonparametric regression and smoothing. Thus our estimator may be viewed as a generalized smoothing quantile estimator. In the special case of Example 6.2, we obtain a nonparametric-nonparametric quantile estimator in as much as nothing is assumed about the form of the population distributions involved. In particular, as the Examples of Section 6.6 show, heteroscedasticity is allowed in the smoothing context.

Our theory depends on the relevance weights, $\{p_{ni}\}$ used to construct the REWED. These weights express the statistician's perceived relationships among the populations and would usually be chosen on intuitive grounds. Making $\bar{F}_n$ approximate F_t well is a primary objective in this choice. Additional restrictions on the $\{p_{ni}\}$ stem from the large sample theory developed in this paper. Theorems 6.1, 6.2, and 6.3 on consistency, for example, require that $\sum \exp(-\epsilon K_n) < \infty$ for all $\epsilon > 0$ where $K_n = (\sum_i p_{ni}^2)^{-1}$. This imposes a requirement that the $\{p_{ni}\} \to 0$ fairly rapidly as $n \to \infty$, say faster than $1/\log(n)$. And for asymptotic normality, we see in Theorem 6.6 the requirement that $\max_{1 \le i \le n}(p_{ni} V_n^{-1/2}) \to 0$ as $n \to \infty$ where $V_n = \sum_i p_{ni}^2 F_{ni}(\xi_{p(n)})(1 - F_{ni}(\xi_{p(n)}))$. We believe these conditions offer some guidance on the choice of the relevance weights without unduly restricting it. In the smoothing model, we can usually use the kernel weights as the relevance weights like we did in the simulation study. The kernel weights usually satisfy the above conditions. $F_n(x)$ also arises as a population distribution estimator in finite population sampling theory (see Sarndal, Swenson and Wretman 1992, p199) where F_{ni} may be regarded as the distribution function of the subpopulation from which x_i is drawn; here $p_i = \pi_i^{-1} / \sum \pi_j^{-1}$, π_i being x_i's selection probability, $i = 1, \ldots, n$..

We would note a Bayesian connection with our theory. If the $\{p_{ni}\}$ are thought of as prior weights, then $\bar{F}_n$ is just the marginal CDF of the independent observations obtained by mixing the conditional models $\{F_{t_i}\}$. Viewed from this perspective, the weights should

be chosen to make the CDF for the population of interest, F_t, that marginal CDF. We would note that incidentally this paper does provide a large sample theory for the Bayesian marginal mixture distribution, in particular for the quantiles of that mixture distribution.

6.9 Proofs of the Theorems

Lemma 6.1 *(Marcus and Zinn, 1984). Let* $\{c_n\}, n = 1, ..., \infty$, *be a sequence of real numbers and* $\{X_n\}, n = 1, ..., \infty$, *a sequence of independent random variables. Define* $U_n(t)$ *by*

$$U_n(t) = \sum_{i=1}^{n} c_i[I(X_i \leq t) - P(X_i \leq t)].$$

Then

$$P(|U_n(t)|(\sum_{i=1}^{n} {c_i}^2)^{-1/2} > \lambda) \leq \exp(-\lambda^2/8)(1 + 2\sqrt{2\pi\lambda})$$

for all $\lambda > 0$. □

Proof of Theorem 6.1.

$$|F_n(x) - \bar{F}_n(x)| = |\sum_{i=1}^{n} p_{ni}[I(X_{ni} \leq x) - F_{ni}(x)]| \stackrel{def}{=} |V_n(x)|,$$

say. So on applying Lemma 6.1 with $\lambda = \epsilon K_n^{1/2}$,

$$\begin{aligned} & P(|F_n(x) - \bar{F}_n(x)| > \epsilon) \\ = \; & P(|V_n(x)|(\sum_{i=1}^{n} p_{ni}^2)^{-1/2} > \epsilon(\sum_{i=1}^{n} p_{ni}^2)^{-1/2}) \\ \leq \; & (1 + 2\sqrt{2\pi\epsilon K_n^{1/2}}) \exp(-\epsilon^2 K_n/8) \end{aligned}$$

for every $\epsilon > 0$. The assumption, $\sum_{n=1}^{\infty} \exp(-\epsilon^2 K_n) < \infty$, implies that $K_n \to \infty$ when $n \to \infty$. It follows that for every $\epsilon > 0$, there exists N, such that for every $n > N$

$$(1 + 2\sqrt{2\pi\epsilon K_n^{1/2}}) < \exp(\frac{\epsilon^2 K_n}{16}).$$

Consequently

$$(1+2\sqrt{2\pi\epsilon K_n^{1/2}})\exp(-\frac{\epsilon^2 K_n}{8}) \leq \exp(-\frac{\epsilon^2 K_n}{16}).$$

But $\sum_{n=1}^{\infty}\exp(-\epsilon^2 K_n) < \infty$ for all $\epsilon > 0$. So $\sum_{n=1}^{\infty}\exp(-\frac{\epsilon^2 K_n}{16}) < \infty$ for every $\epsilon > 0$.

Hence

$$\sum_{n=1}^{\infty} P(|F_n(x) - \bar{F}_n(x)| > \epsilon) < \infty \quad \textit{for all } \epsilon > 0.$$

The Borel-Cantelli Lemma then implies

$$|F_n(x) - \bar{F}_n(x)| \to 0 \quad a.s. \quad \textit{for every } x. \quad \square$$

Proof of Theorem 6.2. Let M be a large positive integer and

$$u_n = \max_{-M^2 \leq i \leq M^2} |F_n(i/M) - \bar{F}_n(i/M)|.$$

By Theorem 1, $u_n \to 0$ $a.s.$. Also monotonicity implies that for $(i-1)/M < t \leq i/M$

$$\begin{aligned} & F_n(t) - \bar{F}_n(t) \leq F_n[i/M] - \bar{F}_n[(i-1)/M] \\ = \; & [F_n[i/M] - \bar{F}_n[i/M]] + [\bar{F}_n[i/M] - \bar{F}_n[(i-1)/M]]. \end{aligned}$$

By similar reasoning,

$$F_n(t) - \bar{F}_n(t) \geq [F_n[(i-1)/M] - \bar{F}_n[(i-1)/M]] - [\bar{F}_n[i/M] - \bar{F}_n[(i-1)/M]].$$

So

$$\begin{aligned} & \limsup_{n\to\infty} |F_n(x) - \bar{F}_n(x)| \\ \leq \; & \limsup_{n\to\infty} u_n + \limsup_{n\to\infty} \max_{-M^2 \leq i \leq M^2} \{\bar{F}_n(\frac{i}{M}) - \bar{F}_n(\frac{i-1}{M}), 1 - \bar{F}_n(M), \bar{F}_n(-M)\} \\ \leq \; & \limsup_{n\to\infty} u_n + \limsup_{n\to\infty} M^{-1} sup_{x,n} \bar{f}_n(x) + \limsup_{n\to\infty} \{(1 - \bar{F}_n(M)), \bar{F}_n(-M)\} \end{aligned}$$

under the assumptions. Since M is arbitrary, the result follows. $\square$

Proof of Theorem 6.3. Let $\epsilon > 0$. By the uniqueness condition and the definition of $\xi_{p(n)}$,

$$\bar{F}_n(\xi_{p(n)} - \epsilon) < p < \bar{F}_n(\xi_{p(n)} + \epsilon).$$

By Theorem 6.2, $F_n(\xi_{p(n)} - \epsilon) - \bar{F}_n(\xi_{p(n)} - \epsilon) \to 0$ *a.s.* and $F_n(\xi_{p(n)} + \epsilon) - \bar{F}_n(\xi_{p(n)} + \epsilon) \to 0$ *a.s.*. Hence $P[F_m(\xi_{p(m)} - \epsilon) < p < F_m(\xi_{p(m)} + \epsilon),\ for\ all\ m \geq n] \to 1$ as $n \to \infty$. That is, $P(\sup_{m>n} |\hat{\xi}_{pm} - \xi_{p(m)}| > \epsilon) \to 0$ *as* $n \to \infty$. This completes the proof. □

To prove the Theorem 6.4, we need the following useful result of Hoeffding (1963).

Lemma 6.2 *Let $Y_1, ..., Y_n$ be independent random variables satisfying $P(a_i \leq Y_i \leq b_i) = 1$ for each i, where $a_i < b_i$. Then for $t > 0$,*

$$P(\sum_{i=1}^{n}[Y_i - E(Y_i)] \geq t) \leq \exp(-2t^2/\sum_{i=1}^{n}(b_i - a_i)^2). \quad \square$$

Proof of Theorem 6.4. Fix $\epsilon > 0$. Then

$$P\{|\hat{\xi}_{np} - \xi_{p(n)}| > \epsilon\} \leq P\{\hat{\xi}_{np} \geq \xi_{p(n)} + \epsilon\} + P\{\hat{\xi}_{np} \leq \xi_{p(n)} - \epsilon\}.$$

But with $Y_i = p_{ni} I(X_{ni} > \xi_{p(n)} + \epsilon)$,

$$\begin{aligned}
& P\{\hat{\xi}_{np} \geq \xi_{p(n)} + \epsilon\} \\
= \ & P\{p > F_n(\xi_{p(n)} + \epsilon)\} \\
= \ & P\{\sum_{i=1}^{n} p_{ni} I(X_{ni} > \xi_{p(n)} + \epsilon) > 1 - p\} \\
= \ & P\{\sum_{i=1}^{n}(Y_i - E(Y_i)) > 1 - p - \sum_{i=1}^{n} p_{ni}(1 - F_{ni}(\xi_{p(n)} + \epsilon))\} \\
= \ & P\{\sum_{i=1}^{n}(Y_i - E(Y_i)) > \bar{F}_n(\xi_{p(n)} + \epsilon) - p\}.
\end{aligned}$$

Because $P(0 \leq Y_i \leq p_{ni}) = 1$ for each i, by Lemma 6.2, we have

$$P(\hat{\xi}_{np} \geq \xi_{p(n)} + \epsilon) \leq \exp(-2\delta_1^2/\sum_{i=1}^{n} p_{ni}^2) = \exp(-2\delta_1^2 K_n);$$

here $\delta_1 = \bar{F}_n(\xi_{p(n)} + \epsilon) - p$. Similarly,

$$P(\hat{\xi}_{np} \leq \xi_{p(n)} - \epsilon) \leq \exp(-2\delta_2^2 / \sum_{i=1}^{n} p_{ni}^2) = \exp(-2\delta_2^2 K_n)$$

where $\delta_2 = p - \bar{F}_n(\xi_{p(n)} - \epsilon)\}$.

Putting $\delta_\epsilon(n) = \min\{\delta_1, \delta_2\}$, completes the proof. □

To prove Theorem 6.5, we need the following results (see Shorack and Wellner, 1986, page 855)

Lemma 6.3 *(Bernstein) Let $Y_1, Y_2, ..., Y_n$ be independent random variables satisfying $P(|Y_i - E(Y_i)| \leq m) = 1$, for each i, where $m < \infty$. Then, for $\epsilon > 0$,*

$$P[|\sum_{i=1}^{n}(Y_i - E(Y_i))| \geq \epsilon] \leq 2\exp[-\frac{\epsilon^2}{2\sum_{i=1}^{n} Var(Y_i) + \frac{2}{3}m\epsilon}]$$

for all n = 1, 2, ...

Lemma 6.4 *Let $0 < p < 1$. Suppose conditions 1-3 of Theorem 6.5 hold. Then with probability 1 (hereafter wp1)*

$$|\hat{\xi}_{np} - \xi_{p(n)}| \leq \frac{(\sqrt{c^*/2} + 1)K_n^{-1/2}(\log K_n)^{1/2}}{\bar{f}_n(\xi_{p(n)})}$$

for all sufficiently large n.

Proof. Since $\bar{F}_n$ is continuous at $\xi_{p(n)}$ with $\bar{F}_n'(\xi_{p(n)}) > 0$, $\xi_{p(n)}$ solves uniquely $\bar{F}_n(x-) \leq p \leq \bar{F}_n(x)$ and $p = \bar{F}_n(\xi_{p(n)})$. Put

$$\epsilon_n = (\sqrt{c^*/2} + 1)K_n^{-1/2}(\log K_n)^{1/2} / \bar{f}_n(\xi_{p(n)}).$$

We then have

$$\begin{aligned} & \bar{F}_n(\xi_{p(n)} + \epsilon_n) - p = \bar{F}_n(\xi_{p(n)} + \epsilon_n) - \bar{F}_n(\xi_{p(n)}) \\ = \; & \bar{f}_n(\xi_{p(n)})\epsilon_n + o(\epsilon_n) \\ \geq \; & \sqrt{c^*/2}(\log K_n)^{1/2}/K_n^{1/2} \end{aligned}$$

for all sufficiently large n.

Likewise we may show that $p - \bar{F}_n(\xi_{p(n)} - \epsilon_n)$ satisfies a similar inequality. Thus, with $\delta_\epsilon(n)$ as defined in Theorem 6.4, we have

$$2K_n\delta_\epsilon(n)^2 \geq c^* \log K_n$$

for all n sufficiently large. Hence by Theorem 6.4,

$$P(|\hat{\xi}_{np} - \xi_{p(n)}| > \epsilon_n) \leq \frac{2}{K_n^{c^*}}$$

for all sufficiently large n.

This last result, hypothesis 3 of this theorem and the Borel-Cantelli Lemma imply that wp1 $|\hat{\xi}_{np} - \xi_{p(n)}| > \epsilon_n$ holds for only finitely many n. This completes the proof. □

Lemma 6.5 *Let $0 < p < 1$ and T_n be any estimator of $\xi_{p(n)}$ for which $T_n - \xi_{p(n)} \to 0$ wp1. Suppose $\bar{F}_n$ has a bounded second derivative in the neighborhood of $\xi_{p(n)}$. Then wp1*

$$\bar{F}_n(T_n) - \bar{F}_n(\xi_{p(n)}) = \bar{F}_n'(\xi_{p(n)})(T_n - \xi_{p(n)}) + O((T_n - \xi_{p(n)})^2)$$

as $n \to \infty$.

Proof. The proof is an immediate consequence of the Taylor expansion. □

For convenience in presenting the next result, we set

$$D_n(x) = [F_n(\xi_{p(n)} + x) - F_n(\xi_{p(n)})] - [\bar{F}_n(\xi_{p(n)} + x) - \bar{F}_n(\xi_{p(n)})].$$

Lemma 6.6 *Let $\{a_n\}$ be a sequence of positive constants such that*

$$a_n \sim c_0 K_n^{-1/2} (\log K_n)^q$$

as $n \to \infty$, for some constants $c_0 > 0$ and $q \geq 1/2$. Let $m_n = \max_{1 \leq i \leq n}\{p_{ni}\}$ and

$$H_{pn} = \sup_{|x| \leq a_n} |D_n(x)|.$$

If $m_n = o(K_n^{-3/4}(\log K_n)^{(q-1)/2})$, then under the hypothesis of Theorem 6.5, wp1

$$H_{pn} = O(K_n^{-3/4}(\log K_n)^{\frac{1}{2}(q+1)}).$$

Proof. Let $\{b_n\}$ be any sequence of positive integers such that $b_n \sim c_0 K_n^{1/4}(\log K_n)^q$ as $n \to \infty$. For successive integers $r = -b_n, ..., b_n$, put $\eta_{r,n} = a_n b_n^{-1} r$ and $\alpha_{r,n} = \bar{F}_n(\xi_{p(n)} + \eta_{r+1,n}) - \bar{F}_n(\xi_{p(n)} + \eta_{r,n})$. The monotonicity of F_n and $\bar{F}_n$ implies that for $\eta_{r,n} \leq x \leq \eta_{r+1,n}$,

$$\begin{aligned} D_n(x) &\leq [F_n(\xi_{p(n)} + \eta_{r+1,n}) - F_n(\xi_{p(n)})] - [\bar{F}_n(\xi_{p(n)} + \eta_{r,n}) - \bar{F}_n(\xi_{p(n)})] \\ &\leq D_n(\eta_{r+1,n}) + [\bar{F}_n(\xi_{p(n)} + \eta_{r+1,n}) - \bar{F}_n(\xi_{p(n)} + \eta_{r,n})]. \end{aligned}$$

Similarly,

$$D_n(x) \geq D_n(\eta_{r,n}) - [\bar{F}_n(\xi_{p(n)} + \eta_{r+1,n}) - \bar{F}_n(\xi_{p(n)} + \eta_{r,n})].$$

So

$$H_{pn} \leq A_n + \beta_n,$$

where $A_n = \max\{|D_n(\eta_{r,n})| : -b_n \leq r \leq b_n\}$ and $\beta_n = \max\{\alpha_{r,n} : -b_n \leq r \leq b_n - 1\}$. Since $\eta_{r+1,n} - \eta_{r,n} = a_n b_n^{-1} \sim K_n^{-3/4}$, $-b_n \leq r \leq b_n - 1$, we have by the Mean Value

Theorem that

$$\alpha_{r,n} \leq [\sup_{|x|\leq a_n} \bar{F}_n'(\xi_{p(n)}+x)](\eta_{r+1,n}-\eta_{r,n}) \sim [\sup_{|x|\leq a_n} \bar{F}_n'(\xi_{p(n)}+x)]K_n^{-3/4},$$

$-b_n \leq r \leq b_n - 1$. Thus

$$\beta_n = O(K_n^{-3/4}), \quad n \to \infty.$$

We now establish that wp1

$$A_n = O(K_n^{-3/4}(\log K_n)^{\frac{1}{2}(q+1)}) \ \ as\ n \to \infty.$$

By the Borel-Cantelli Lemma it suffices to show that

$$\sum_{n=1}^{\infty} P(A_n \geq \gamma_n) < \infty$$

where $\gamma_n = c_1 K_n^{-3/4}(\log K_n)^{\frac{1}{2}(q+1)}$ for some constant $c_1 > 0$. Now

$$P(A_n \geq \gamma_n) \leq \sum_{r=-b_n}^{b_n} P(|D_n(\eta_{r,n})| \geq \gamma_n).$$

And

$$|D_n(\eta_{r,n})| = |\sum_{i=1}^{n} p_{ni}(I(X_{ni} \in (\xi_{p(n)}, \xi_{p(n)}+\eta_{r,n})) - E(I(X_{ni} \in (\xi_{p(n)}, \xi_{p(n)}+\eta_{r,n}))))|$$

by definition. With $Y_i = p_{ni} I(X_{ni} \in (\xi_{p(n)}, \xi_{p(n)}+\eta_{r,n}))$, Bernstein's Lemma (see Lemma 6.3) implies

$$P(|D_n(\eta_{r,n})| \geq \gamma_n) \leq 2\exp(-\gamma_n^2/D_n)$$

where $D_n = 2\sum_{i=1}^{n} Var(Y_i) + 2/3 m_n \gamma_n$.

Choose $c_2 > \sup_{n,i} f_{ni}(\xi_{p(n)})$. Then there exists an integer N such that

$$F_{ni}(\xi_{p(n)}+a_n) - F_{ni}(\xi_{p(n)}) < c_2 a_n$$

and

$$F_{ni}(\xi_{p(n)}) - F_{ni}(\xi_{p(n)} - a_n) < c_2 a_n$$

both of the above inequalities being for all $n > N$ and $i = 1, \ldots, n$. Then

$$\sum_{i=1}^{n} Var(Y_i) \leq \sum_{i=1}^{n} p_{ni}^2 c_2 a_n = K_n^{-1} c_2 a_n.$$

Hence

$$\gamma_n^2 / D_n \geq \gamma_n^2 / \{2K_n^{-1} c_2 a_n + 2/3 m_n \gamma_n\} \geq c_1^2 \log K_n / (4 c_2 c_0)$$

for all sufficiently large n. The last result obtains because of the condition $m_n = o(K_n^{-3/4} (\log K_n)^{(q-1)/2})$.

Given c_0 and c_2, we may choose c_1 large enough that $c_1^2 (4 c_2 c_0)^{-1} > c^* + 1$. It then follows that there exists N^* such that

$$P(|D_n(\eta_{r,n})| \geq \gamma_n) \leq 2K_n^{-(c^*+1)}$$

for all $|r| \leq b_n$ and $n > N^*$. Consequently, for $n > N^*$

$$P(A_n \geq \gamma_n) \leq 8 b_n K_n^{-(c^*+1)}.$$

In turn this implies

$$P(A_n \geq \gamma_n) \leq 8 K_n^{-c^*}.$$

Hence $\sum_{n=1}^{\infty} P(A_n \geq \gamma_n) < \infty$, and the proof is complete. □

Proof of Theorem 6.5. Under the conditions of Theorem, we may apply Lemma 6.4. This means Lemma 6.5 becomes applicable with $T_n = \hat{\xi}_{np}$ and we have wp1,

$$\bar{F}_n(\hat{\xi}_{np}) - \bar{F}_n(\xi_{p(n)}) = \bar{f}_n(\xi_{p(n)})(\hat{\xi}_{np} - \xi_{p(n)}) + O(K_n^{-1} \log K_n), \quad as\ n \to \infty.$$

Now using Lemma 6.6 with $q = 1/2$, and appealing to Lemma 6.4 again, we may pass from the last conclusion to: wp1

$$F_n(\hat{\xi}_{np}) - F_n(\xi_{p(n)}) = \bar{f}_n(\xi_{p(n)})(\hat{\xi}_{np} - \xi_{p(n)}) + O(K_n^{-3/4} (\log K_n)^{3/4}), \quad as\ n \to \infty.$$

Finally, since wp1: $F_n(\hat{\xi}_{np}) = p + O(m_n)$, as $n \to \infty$, we have wp1

$$p - F_n(\xi_{p(n)}) = \bar{f}_n(\xi_{p(n)})(\hat{\xi}_{np} - \xi_{p(n)}) + O(K_n^{-3/4}(\log K_n)^{3/4}), \quad as\ n \to \infty.$$

This completes the proof. □

Proof of Theorem 6.6. Fix t, and put

$$G_n(t) = P[\bar{f}_n(\xi_{p(n)})(\hat{\xi}_{np} - \xi_{p(n)})V_n^{-1/2} \le t] = P[\hat{\xi}_{np} \le a_n],$$

where $a_n = \xi_{p(n)} + tV_n^{1/2}/\bar{f}_n(\xi_{p(n)})$ Then by the definition of $\hat{\xi}_{np}$

$$G_n(t) = P(F_n(a_n) \ge p).$$

Thus

$$\begin{aligned} & G_n(t) = P(\sum_{i=1}^{n} p_{ni} I(X_{ni} \le a_n) \ge p) \\ = \; & P(V_n^{-1/2} \sum_{i=1}^{n} p_{ni}(I(X_{ni} \le a_n) - E(I(X_{ni} \le a_n))) \ge V_n^{-1/2}[p - \sum_{i=1}^{n} p_{ni} E(I(X_{ni} \le a_n))]) \\ = \; & P(Z_n \ge c_n); \end{aligned}$$

here

$$Z_n = V_n^{-1/2} \sum_{i=1}^{n} p_{ni}(I(X_{ni} \le a_n) - E(I(X_{ni} \le a_n)))$$

and

$$c_n = V_n^{-1/2}.[p - \sum_{i=1}^{n} p_{ni} E(I(X_{ni} \le a_n))]$$

We first prove $Z_n \to N(0,1)$ in distribution and then that $c_n \to -t$ as $n \to \infty$ to complete the proof . To this end

$$Z_n = \sum_{i=1}^{n} p_{ni} V_n^{-1/2}(I(X_{ni} \le a_n) - F_{ni}(a_n)) = \sum_{i=1}^{n} \eta_{ni}$$

where $\eta_{ni} = p_{ni} V_n^{-1/2}(I(X_{ni} \le a_n) - F_{ni}(a_n))$.

From the condition $\max_{1 \le i \le n}(p_{ni} V_n^{-1/2}) \to 0$, we get $\max_{1 \le i \le n} p_{ni} \to 0$ and $V_n \to 0$. We then easily obtain for every $\epsilon > 0$ and $\tau > 0$:

1. $\sum_{i=1}^{n} P(|\eta_{ni}| \geq \epsilon) \to 0$. (since $|\eta_{ni}| \leq 2\max_{1\leq i\leq n}(p_{ni}V_n^{-1/2}) \to 0$);

2. $\sum_{i=1}^{n}[E(\eta_{ni}^2 I(|\eta_{ni}| < \tau)) - (since E(\eta_{ni}I(|\eta_{ni}| < \tau)))^2] \to 1$. (For n large enough, $I(|\eta_{ni}| < \tau)) = 1$);

3. $\sum_{i=1}^{n} E(\eta_{ni}I(|\eta_{ni}| < \tau)) \to 0$.

So $Z_n \to N(0,1)$ in distribution by the CLT (see Chung, 1968, page 191) for triangular independent random variables.

Next we prove $c_n \to -t$.

$$\begin{aligned} & c_n = V_n^{-1/2}[p - \sum_{i=1}^{n} p_{ni}E(I(X_{ni} \leq a_n))] \\ = \; & V_n^{-1/2}\sum_{i=1}^{n} p_{ni}(F_{ni}(\xi_{p(n)}) - F_{ni}(a_n)) \\ = \; & -V_n^{-1/2}\sum_{i=1}^{n} p_{ni}(f_{ni}(\xi_{p(n)})(a_n - \xi_{p(n)}) + o(a_n - \xi_{p(n)})) \\ \to \; & -t \;\; as \; n \to \infty. \end{aligned}$$

The proof is now complete. □

Chapter 7

Some Further Results

In this Chapter we discuss some further results about the use of relevant sample information. In Section 7.1, we consider the locally polynomial MREWLE for generalized smoothing models. We show that the locally linear MREWLE has many desirable properties.

In Chapter 3, we have considered the full parametric approach to relevant sample analysis. Chapter 6 proposes a kind of nonparametric approach. In Section 7.2, we outline a semiparametric approach by using estimating functions.

7.1 Locally polynomial maximum relevance weighted likelihood estimation

Locally regression is a popular form of nonparametric regression, combining excellent theoretical properties with conceptual simplicity and flexibility to find structure in many datasets. The method was introduced by Stone (1977) using a rectangular window. Cleveland (1979) introduced his lowess procedure, which is based on locally polynomial

fitting and incorporates many important features, such as design adaptiveness , kernel weighting and robustness. Recently, Fan (1992, 1993) has studied minimax properties of locally linear regression. A detailed summary of the advantages of locally regression compared to kernel fitting may be found in Hastie and Loader (1993). Fan, Heckman and Wand (1993) have studied locally polynomial kernel regression for generalized linear models and quasi-likelihood functions. Weerahandi and Zidek (1988) present locally smooth processes from Bayesian viewpoint.

In this section we investigate the generalization of locally polynomial fitting with kernel methods for the MREWLE in the case of generalized smoothing models. We are motivated by the fact that locally regressions are both intuitively and mathematically simple which allows us to achieve a deeper understanding of their performance.

In the generalized smoothing model of Chapter 5, when we suppose that θ has q continuous derivatives, then in a small neighborhood of a point x,

$$\begin{aligned}\theta(z) &= \theta(x) + \theta'(x)(z-x) + \cdots + \frac{\theta^{(q)}(x)}{q!}(z-x)^q + o\{(z-x)^q\} \\ &= \beta_0 + \beta_1(z-x) + \cdots + \beta_q(z-x)^q + o\{(z-x)^q\}. \qquad (7.1)\end{aligned}$$

To estimate $\theta(x)$, we want to use the information from $\{(Y_i, X_i)\}$. As with the relevance weighted likelihood in Chapter 5, we define the locally polynomial relevance weighted likelihood at the point x to be

$$\prod_{i=1}^{n} f^{p_i}\{Y_i, \beta_0 + \beta_1(X_i - x) + \cdots + \beta_q(X_i - x)^q\}. \qquad (7.2)$$

Here $\{p_i\}$ are the relevance weights.

One would then estimate $(\beta_0, \ldots, \beta_q)$ by maximizing the local polynomial relevance weighted likelihood (7.2). This is locally polynomial MREWLE. Because $\beta_0 + \beta_1(z - x) + \cdots + \beta_q(z-x)^q$ is a better approximation to $\theta(z)$ than $\theta(x)$, the locally polynomial

REWL in (7.2) is better than the REWL in Chapter 5. We would expect the locally polynomial MREWLE to perform better than its locally constant counterpart.

There are many possible strategies for weighting the likelihood (See Chapter 5). Because of their mathematical simplicity we will use kernel weights, the Nadaraya-Watson weights in Chapter 5.

The locally polynomial kernel MREWLE of $\theta(x)$ is then given by

$$\hat{\theta}_0(x; q, h_n) = \hat{\beta}_0$$

where $(\hat{\beta}_0, \ldots, \hat{\beta}_q)$ maximizes

$$\sum K_h(x - X_i) \log f\{Y_i, \beta_0 + \cdots \beta_q (X_i - x)^q\}.$$

Note that in the case $q = 0$, the estimator of $\theta(x)$ is the Nadaraya-Watson MREWLE in Chapter 5.

Locally polynomial fitting also provides consistent estimates of higher order derivatives of $\theta(x)$ through the coefficients of the higher order terms in the polynomial fit. Thus, for $0 \leq r < q$, we define locally polynomial MREWLE of $\theta^{(r)}(x)$ to be

$$\hat{\theta}_r(x; q, h) = r! \hat{\beta}_r.$$

We leave the investigation of the asymptotic properties of $\hat{\theta}_r(x; q, h)$ to future work. In most applications, we are interested in a low-degree polynomial ($p = 1, 2,$ or 3). For brevity, we consider the locally linear MREWLE in this section. The locally linear MREWLE maximizes

$$\sum K_{h_n}(x - X_i) \log f\{Y_i, \beta_0 + \beta_1 (x - X_i)\}.$$

The corresponding locally linear REWL equations are

$$\sum K_{h_n}(x - X_i)\frac{\partial \log f\{Y_i, \beta_0 + \beta_1(x - X_i)\}}{\partial \beta_0} = 0$$

and

$$\sum K_{h_n}(x - X_i)\frac{\partial \log f\{Y_i, \beta_0 + \beta_1(x - X_i)\}}{\partial \beta_0}(X_i - x) = 0.$$

We state the following asymptotic properties of the locally linear MREWLE as conjectures. Proof will be left to future work.

Assume that the second derivative of θ is continuous. If $h_n \to 0$ and $nh_n \to \infty$, then under Assumption 5.1, for $x \in (a_o, b_o)$, the locally linear REWL equations admit a sequence of solutions $\{\hat{\theta}_n(x, 1, h_n)\}$ satisfying:

$$\begin{aligned}[\frac{nh_n g(x) I\{\theta(x)\}}{d_K}]^{1/2}[\hat{\theta}_n(x, 1, h_n) - \theta(x) \quad &- \quad \frac{\theta''(x) c_K h_n^2}{2}\{1 + O(h_n)\}] \\ &\to \quad N(0, 1).\square \end{aligned} \tag{7.3}$$

So the asymptotic bias and variance are $\theta''(x)c_K h_n^2/2$ and $d_K/[nh_n g(x) I\{\theta(x)\}]$ respectively.

Now we compare the result of above statement with Theorem 1 of Fan (1992) for locally linear regression. Fan (1992) proved that under certain conditions, the locally linear regression estimator $\hat{\theta}$ has MSE

$$E\{\hat{\theta}(x) - \theta(x)\}^2 = \{\frac{\theta''(x)}{2}\}^2 c_K^2 h_n^4 + \frac{d_K \sigma^2(x)}{nh_n g(x)} + o_p(h_n^4 + \frac{1}{h_n}). \tag{7.4}$$

From the bias and variance comparison, we find the locally linear MREWLE and the locally linear regression estimator have the same bias. But the locally linear MREWLE always has smaller variance which depends on Fisher information. Also the locally linear regression estimator only works for a location parameter, while the locally linear MREWLE works for any parameter.

Fan (1992) also show that locally linear regression estimator is the best among all the linear estimators of $\theta(x)$. The above result of the locally linear MREWLE tells us that we usually can find a better estimator than the locally linear regression when the density $f\{y_i, \theta(x)\}$ is known.

This result supports the argument in Section 5.2, that when we know the distribution family, the linear smoother is asymptotic inadmissible. In the next section, we will show that the locally linear MREWLE is the best among all the locally linear estimators for estimating functions.

This locally linear MREWLE has all the desirable properties discussed in Fan (1992). Now we use a simple simulation to illustrate its finite sample behavior.

We use the model given in (5.21). The simulation sample size $n = 200$, bandwidth $h = 0.1$, and Gaussian kernel function are used. Three methods are used: the Nadaraya-Watson MREWLE, the locally linear smoother MREWLE (defined in Chapter 5) and the locally linear MREWLE. Figure 7.1 shows a comparison among these three methods. The locally linear MREWLE based on five simulations is shown in Figure 7.2.

From Figure 7.1, both the locally linear smoother MREWLE and the locally linear MREWLE perform very well. The Nadaraya-Watson MREWLE has large boundary effects. The locally linear MREWLE works excellent as showing in Figure 7.2. From the results of Simulation 1 of Chapter 5 and Figure 7.2, We can conclude that the locally linear MREWLE is the best.

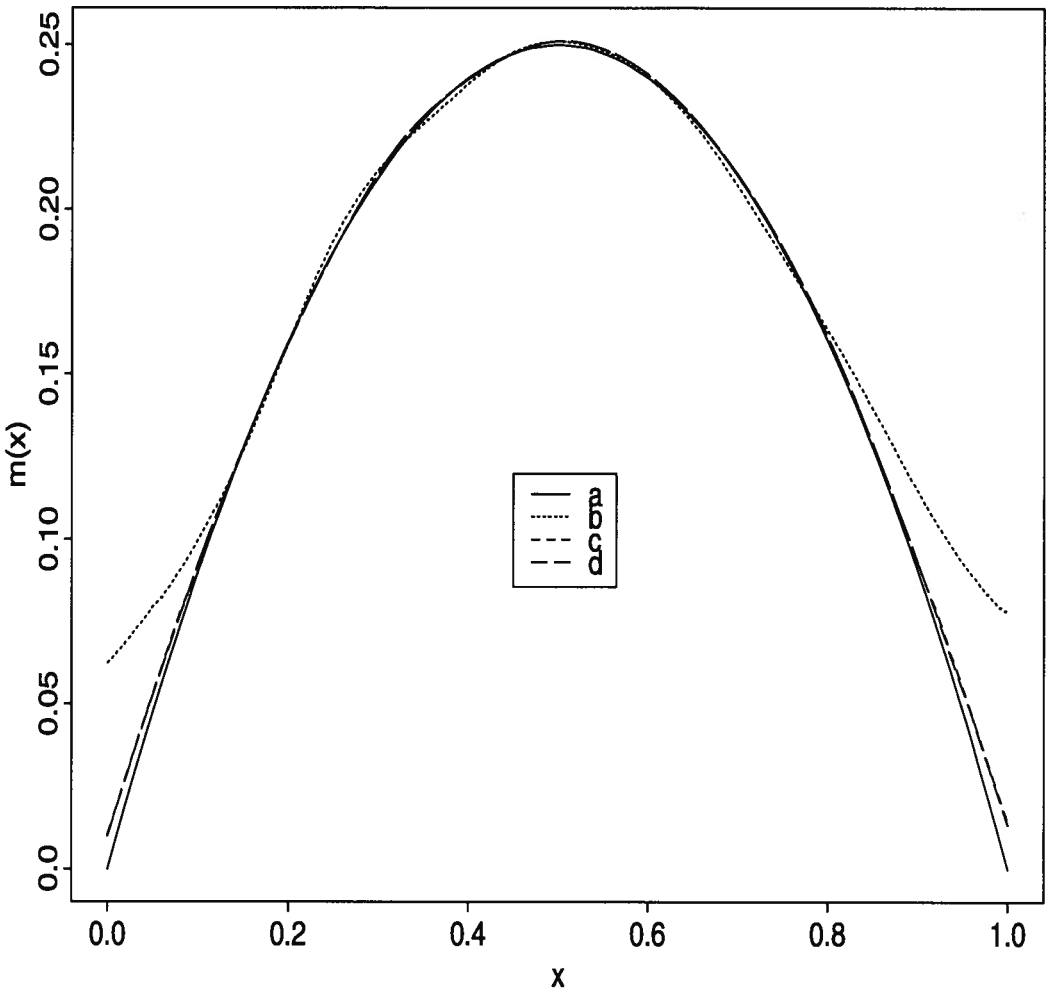

Figure 7.1: *A comparison of the Nadaraya-Watson MREWLE, the locally linear smoother MREWLE and the locally linear MREWLE from Model (5.21) with n=200. The true curve is a, the Nadaraya-Watson MREWLE, b, the locally linear smoother MREWLE, c, and the locally linear MREWLE, d.*

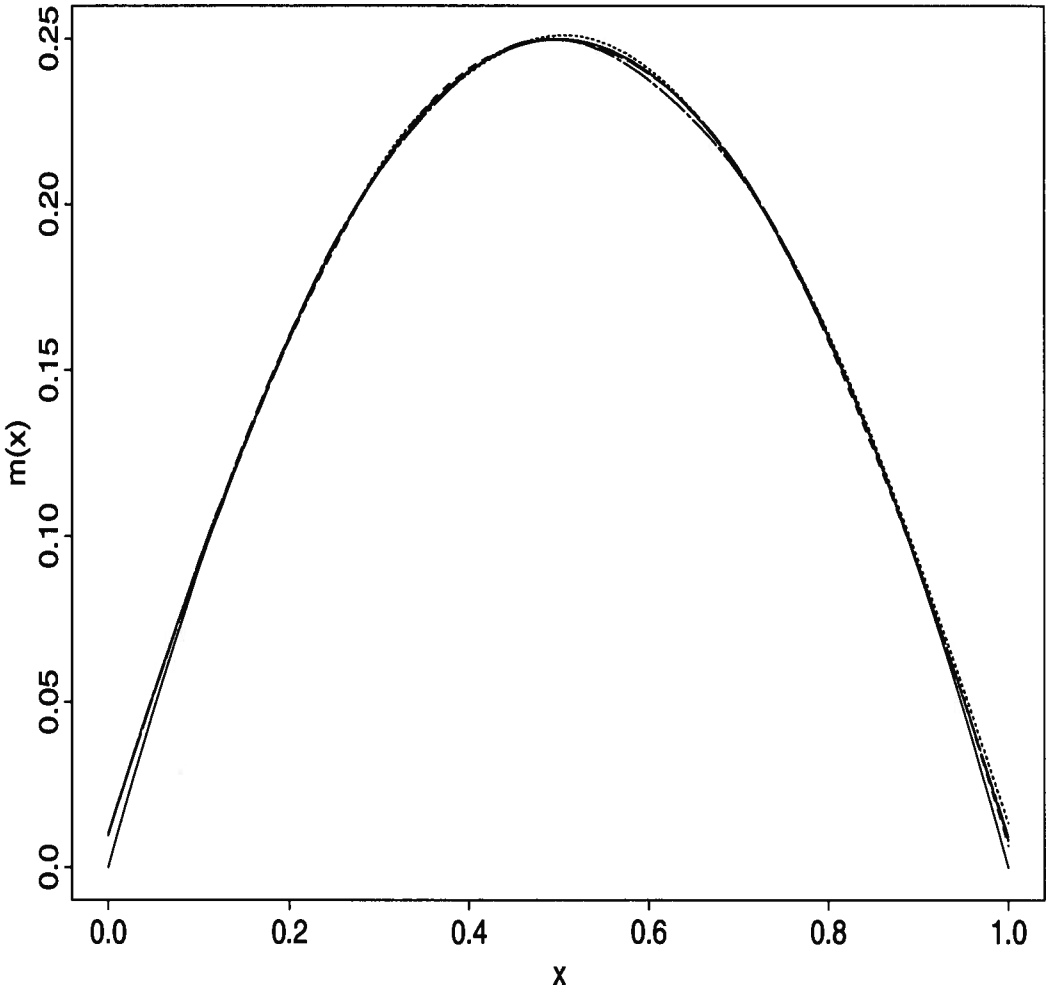

Figure 7.2: *The locally linear MREWLE based on five simulations from Model (5.21) with n=200.*

7.2 Relevance weighted estimating functions

For the exact sample case, it is well known that in many practical circumstances, where even though the full likelihood is unknown, one can specify some relationship about the parameters. For example, if we can specify the relationship between the mean and variance, then we use a quasi-likelihood function approach. More generally, the theory of estimating functions (which include quasi-likelihood as a special case) have been developed in the series papers of Godambe (e.g. Godambe (1960), Godambe and Thompson (1974), Godambe (1976)). This theory focused initially on the elementary fact that any point estimator may regard as the solution of an equation

$$\xi(y, \theta) = 0. \tag{7.5}$$

for data vector y and parameters θ. Unlike traditional approachs, however, which impose conditions on estimator as unbiasedness, invariance etc., the approach here focuses on properties of the estimating function itself, rather than the estimator derived from it. Thus, instead of dealing with linear estimators, unbiased estimators and so on, we deal with linear estimating functions, unbiased estimating functions and so on.

The simplest condition one might impose on an estimating function is that of unbiasedness, i.e.

$$E_\theta \xi(Y; \theta) = 0.$$

Note that this does not necessarily imply unbiasedness of the corresponding estimator unless ξ is linear in θ. However, under suitable regularity conditions it does imply consistency of the estimator.

Clearly, there will be a great variety of competing estimation functions. Godambe (1960) suggested that an optimal estimating function (OEF) would be one for which

the efficiency index

$$Eff(\xi) = \frac{E_\theta(\xi^2)}{\{E(\partial\xi/\partial\theta)\}^2} \tag{7.6}$$

takes its smallest possible value among all unbiased estimation functions.

Under the type of regularity conditions used in the Cramer-Rao argument, Godambe (1960) showed that the least possible value of (7.6) is attained, in the one parameter case by the score function

$$\xi^* = \frac{\partial \log f(y,\theta)}{\partial\theta}.$$

Moreover, the smallest possible value is the reciprocal of Fisher information function. We thus have a justification for maximum likelihood estimation which unlike more traditional arguments, does not rely explicitly on asymptotics.

This fairly general situation includes many models important in application. Examples are:

- nuisance parameter model e.g. Neyman and Scott (1948);
- location model i.e. $\mathcal{F} = \{f(x-\theta) = f \text{ unknown }\}$;
- scale model, $\mathcal{F} = \{f(x/\theta) = f \text{ unknown }\}$;
- semi-parametric models such as Cox's regression model, Cox (1972);
- generalized linear models with unspecified error distributions (for example, quasi-likelihood models).

In Chapter 2, we have classified different sources of information. For relevant sample information, the full likelihood approach has been proposed in Chapter 3. Because of its generality, in this section we investigate the generalization of estimating functions for the relevant sample, which we call *relevance weighted estimating functions.*

Let $Y_1, \cdots, Y_n$ be random variables on a sample space with probability density $f(y, \theta_i)$, $i = 1, \cdots, n$. Here θ_i is a real- or vector-valued parameter which is assume to be unknown. Interest lies in the study population corresponding to parameter θ. We know that there is some relation (see Chapter 2) between θ_i and θ. Our object is to estimate θ on the basis of observed values $Y_1, \cdots, Y_n$. If we know the distribution $f(y, \theta)$, we can use the REWL method in Chapter 3. When $f(y, \theta)$ is unknown, but we can specify some relationship about the parameters (for example, $E_\theta \xi(Y, \theta) = 0$), then we should use the following method of relevance weighted estimating functions instead.

As with estimating equation (7.5) for an exact sample, we define the relevance estimating equation:

$$\sum_{i=1}^{n} p_i \xi(Y_i, \theta) = 0 \tag{7.7}$$

to estimate the parameter θ from data $Y_1, \cdots, Y_n$. Here $\{p_i\}$ are the corresponding relevance weights of Y_i.

Note that if choose $\xi(y, \theta) = \partial \log f(y, \theta)/\partial\theta$, we get the relevance weighted likelihood equation in Chapter 3. Also if let $\xi(y, \theta) = y - \theta$, the relevance weighted estimating equation become the relevance weighted least square equation.

Since the estimating function $\xi(y, \theta)$ is initially used to obtain an estimator by solving the equation $\xi(y, \theta) = 0$, the unbiasedness condition

$$E_\theta(\xi(Y, \theta)) = 0 \tag{7.8}$$

becomes natural when the sample Y is exact. Now for relevant observations, Y_i, from $f(y, \theta_i)$, usually $E_{\theta_i} \xi(Y_i, \theta) \neq 0$. This means that relevant samples are usually biased. This makes us define the bias function for relevant samples as

$$b_i = E_{\theta_i} \xi(Y_i, \theta). \tag{7.9}$$

Both the bias function in Chapter 4 and 5 are special cases of (7.9).

For competing estimating functions, we define the following efficiency index for estimating function (7.7) as

$$Eff(\xi) = \frac{E(\sum p_i \xi(Y_i, \theta))^2}{[E \sum p_i \partial \xi(Y_i, \theta)/\partial\theta]^2}. \tag{7.10}$$

An optimal estimating function would be one for which the above efficiency index takes its smallest possible value among all ξ functions which satisfy $E_\theta \xi(Y, \theta) = 0$.

Under certain conditions, we can generalize the results of Chapter 4 to the case of relevance weighted estimating functions.

We now apply the relevance weighted estimating functions to generalized smoothing models. Consider a model for the relationship between a dependent variable Y and an independent variable X. Suppose that Y given $X = x$ satisfy

$$E\xi(Y, \theta(x)) = 0, \tag{7.11}$$

where θ is an unknown smooth function. The NR (Hardle 1990), semiparametric smoothing models (Severini and Staniswalis 1994), generalized linear smoothing model (Fan, Heckman and Wand 1993) and generalized smoothing model in Chapter 3 are special cases of (7.11).

To estimate $\theta(x)$, we construct the relevance weighted (REW) estimating equation as

$$\sum_{i=1}^{n} p_{ni}(x) \xi(Y_i, \theta(x)) = 0.$$

As in Chapter 5, we can prove several theorems about the properties of REW estimating equation. By using the efficiency index (7.10), we can compare the REW estimating functions.

Now we extend the idea of locally polynomial approximations to REW estimating equation. For simplicity, we only consider the locally linear REW estimating equations, which is defined as

$$\sum p_{ni}(x)\xi\{Y_i, \theta(x) + \theta'(x)(X_i - x)\} = 0 \tag{7.12}$$

and

$$\sum p_{ni}(x)\xi\{Y_i, \theta(x) + \theta'(x)(X_i - x)\}(X_i - x) = 0. \tag{7.13}$$

The above locally linear REW estimating equations are easily extended to locally polynomial REW estimating equations.

If we use the Nadaraya-Watson kernel weights, after some calculation, we get the efficiency index for $\theta(x)$ to be

$$\{\frac{\theta''(x)}{2}\}^2 c_K^2 h_n^4 + \frac{d_K}{nh_n g(x)} \frac{E[\xi\{Y, \theta(x)\}]^2}{[E\partial\xi\{Y, \theta(x)\}/\partial\theta(x)]^2} + o_p(h_n^4 + \frac{1}{nh_n}). \tag{7.14}$$

The minimum of (7.14) is attained by choosing

$$\xi(Y, \theta) = \frac{\partial \log f(Y, \theta)}{\partial\theta}.$$

This result tells us that *the locally linear MREWLE is the best among all locally linear REW estimating equations (with Kernel weights).* So when the likelihood function is available, we should use MREWLE to extract the relevant sample information from relevant samples.

Similarly, we can show that *the locally linear quasi-MREWLE (maximum relevance weighted quasi-likelihood estimator) is the best among all locally linear REW linear estimating equations.*

These last two statements ensure that both relevance weighted likelihood and relevance weighted quasi-likelihood play important roles in generalized smoothing models.

Chapter 8

An Approach of Bootstrapping Through Estimating Equations

8.1 Introduction

This Chapter presents a new bootstrap method. It has computational and theoretical advantages when the data obtain from non-identically distributed observables. And it differs from conventional bootstrap methods in that it resamples components of an estimating function rather than the data themselves.

Various authors explore bootstrap resampling procedures for estimating a sampling distribution in situations where the sampled observables are independent and identically distributed (*c.f.* Efron 1979, Bickel and Freedman 1981 and Singh 1981). However, the potential value of bootstrap methods lies in more complex situations like nonlinear regression analysis. In such situations standard inferential methods encounter serious difficulty. A number of authors (Efron 1979, Freedman 1981 and Wu 1986) propose the use of bootstrap methods instead.

The recent publications of Hall (1992) as well as Efron and Tibshirani (1993) survey the literature on bootstrapping. A recent contribution is that of Liu and Singh (1992); it provides a succinct overview of bootstrapping and assessment of methods for estimating mean square errors of statistics in heteroscedastic linear regression. Booth, Hall and Wood (1992) extend the bootstrap to estimate conditional distributions (obtaining confidence intervals and hypothesis tests in particular). Lahiri (1992) "fixed "the bootstrap M-estimator (we offer an alternative in Section 8.6).

However, we describe in Section 8.2 the three principal bootstrapping methods for regression; this is necessary to put our results in perspective. Two of these methods bootstrap residuals. They implicitly assume exchangeability of the residuals and hence fail to be robust against "heteroscedasticity" of high order moments (the third and fourth moments, in particular) . The third method resamples from the (y, x) pairs and also encounters major difficulties.

The new bootstrap in Section 8.3 can be used whenever estimating equations define the estimator. The method, unlike conventional bootstrapping, does not resample the data. Instead we start with the estimator of interest and generate random estimator to fit residuals rather than use model to fit residuals. The approximate sampling distribution for the estimator is then found by taking the empirical distribution of the resulting estimator plus bootstrapped residuals. A critical feature of our approach is its use of the components of the estimating function itself to transform the residuals to an appropriate scale.

For simplicity, we restrict ourselves in this Chapter to estimation for the linear model. Section 8.4 gives the asymptotic properties of our method, which yield covariance estimators having the desiderata listed by Wu (1986, Section 5). In particular, for linear

regression models (8.1), the proposed covariance estimators are almost unbiased and consistent for heteroscedastic errors. For homoscedastic errors, we show that one of these estimators is unbiased for estimating the covariance of the least squares estimator. The new method is robust against nonhomogeneity, not just of the second, but also of higher moments. Such robustness is needed in constructing Edgeworth approximations to the distribution of a bootstrap estimator where consistent third moment estimators are then essential. Finally, the asymptotic distribution of the new bootstrap estimator is normal for both random and nonrandom regression design matrices.

In Section 8.5, we give a general comparison of all the principal bootstrap methods for regression. This comparison shows on heuristic grounds that our new method is more natural than its competitors in heteroscedastic regression. The simulation reported in Section 8.5 also shows our bootstrap to advantage. In that simulation, we use the bootstrap distribution to estimate the true distribution. From the estimate, we can estimate confidence intervals among other things.

Because our new methods based on the estimating equations, it generalizes readily to nonlinear regression. The importance of this advantage derives from the growing importance in practice of nonlinear regression models. Three nonlinear situations are discussed in Section 8.6. Section 8.8 contains the proofs of our main results.

8.2 Problems with Common Bootstrap Regression Methods

Define a linear model by $Y_i = x_i^T\beta + e_i$, where x_i is a $k \times 1$ nonrandom vector. Here β denotes a $k \times 1$ parameter vector and the $\{e_i\}$, uncorrelated errors with means

zero and variances $\{\sigma_i^2\}$, respectively. Assume the model includes an intercept term so that the same designated coordinate of x_i is 1 for all i. With $Y = (Y_1, ..., Y_n)^T$, $e = (e_1, ..., e_n)^T$ and $X = (x_1, ..., x_n)^T$, reexpress the linear model as

$$Y = X\beta + e, \quad Cov(e) = \Omega = diag(\sigma_1^2, ..., \sigma_n^2). \tag{8.1}$$

To reduce clutter, we drop in (8.1) and in the sequel the subscript representing the fixed parameter in conditional distributions. Thus we use E, var and cov to represent respectively, the conditional expectation, variance and covariance induced by the model in equation (8.1) for fixed β and Ω. Assume X^TX is nonsingular so that we have the usual coefficient estimator, $\hat{\beta} = (X^TX)^{-1}X^TY$. The observed Y is $y = (y_1, ..., y_n)^T$.

Suppose a bootstrap distribution of some parameter $\hat{\psi} = \psi(\hat{\beta})$ is required. Efron (1979) suggests two methods (described in next subsection) based on either the residuals of the least squares fit or the complete observation vectors. These methods are studied by Freedman (1981), Wu (1986) and Hinkley (1988), among others.

8.2.1 Residual Resampling

a) **Efron's Method**. The most common method of bootstrapping an ordinary linear model, that of Efron (1979), exploits the assumed exchangeability of the error terms when $\sigma_i^2 = \sigma^2$ for all i. This method is the focus of work by Freedman and Peters (1984) , Wu (1986) and Hinkley (1988). For the model (8.1), e represents a vector of iid random variables with mean 0 and covariance $\sigma^2 I$.

The bootstrap uses the empirical distribution function, $\hat{F}_r$: mass $1/n$ at r_i, $i = 1, ..., n$; here the $\{r_i\}$ represent the elements of the residual vector $r = [I - X(X^TX)^{-1}X^T]y$

where $y = (y_1, \cdots, y_n)^T$ when the observed responses are $\{y_1, \cdots, y_n\}$. Since an intercept term is included in the linear model $\sum r_i = 0$.

The bootstrap (Monte Carlo) approximation obtains from an iid sample of n residuals from $\hat{F}_r$, say $\{e_1^*, \ldots, e_n^*\}$. Let $\mathbf{e}^* = (e_1^*, \ldots, e_n^*)$ and $\mathbf{y}^* = (y_1^*, \cdots, y_n^*)$ where $y_i^* = x_i^T\hat{\beta} + e_i^*$ for all i. Finally define the bootstrap LS Estimator (LSE) as

$$\hat{\beta}^* = (X^TX)^{-1}X^T\mathbf{y}^* = \hat{\beta} + (X^TX)^{-1}X^T\mathbf{e}^*.$$

Efron (1979) shows that $E_*\hat{\beta}^* = \hat{\beta}$ and that $v_b = E_*(\hat{\beta}^* - \hat{\beta})(\hat{\beta}^* - \hat{\beta})^T = \hat{v} = \hat{\sigma}^2(X^TX)^{-1}$, where $\hat{\sigma}^2 = \frac{1}{n}\sum r_i^2$.

We may use $\hat{\psi}^* = \psi(\hat{\beta}^*)$ for estimating ψ. If ψ is smooth, we may develop asymptotic theory for $\hat{\psi}^*$ directly from that of $\hat{\beta}^*$. For brevity we will not consider such an extension of our results here.

Notice that $\hat{v}$ is the usual maximum likelihood estimator of the covariance of $\hat{\beta}$; it is biased since $E_Y(v_b) = n^{-1}(n-k)\sigma^2(X^TX)^{-1}$. The bias results from the fact that $Var_Y(r_i) = \sigma^2(1-h_i)$, where $h_i = x_i^T(X^TX)^{-1}x_i$ is the ith diagonal element of $X(X^TX)^{-1}X^T$. A bootstrap method which uses standardized residuals, $r_i(1-h_i)^{-1/2}$ or $r_i(1-kn^{-1})^{-1/2}$ eliminates this bias.

b) **Wu's Method.** For unequal σ_i^2's in (8.1), v_b is not only biased but inconsistent as well since the true covariance of the least square estimator (LSE), $\hat{\beta}$, is $Cov(\hat{\beta}) = (X^TX)^{-1}\sum\sigma_i^2x_ix_i^T(X^TX)^{-1}$. This deficiency of v_b is intrinsic. Drawing iid samples from $\{r_i\}$ only makes sense for exchangeable residuals $\{r_i\}$; iid sampling wipes out heterogeneity among the $\{r_i\}$ and this heterogeneity will not be reflected in v_b. To deal with this difficulty, Wu (1986) proposes another bootstrap (see also Efron 1986) described below.

Suppose we wish to assess the variability of a statistic computed from the random vector of observables $\mathbf{Y}$ which comes from a specific, but unknown structural/stochastic mechanism. Here we suppose $\mathbf{Y} = X\beta + \mathbf{e}$, with unknown parameters $\beta, \sigma_1^2, ..., \sigma_n^2$ and $\hat{\beta} = (X^TX)^{-1}X^T\mathbf{Y}$. The general bootstrap method would consist of: **(i)** estimating the entire probability mechanism, P, by say $\hat{P}$ [in this case the $\hat{P}$ associated with $(\hat{\beta}, \hat{\sigma}_1^2, ..., \hat{\sigma}_n^2)$]; **(ii)** after resampling the data vector y^* from $\hat{P}$, recalculating the statistic of interest, $\hat{\beta}^* = (X^TX)^{-1}X^Ty^*$; **(iii)** and finally using the observed variability of the $\hat{\beta}^*$'s to estimate the variability of $\hat{\beta}$. For (8.1) the bootstrap calculations can be carried out explicitly and they yield $v_b = (X^TX)^{-1}\sum \hat{\sigma}_i^2 x_i x_i^T (X^TX)^{-1}$.

Wu's (1986) particular implementation of the general bootstrap described above can be obtained by choosing $\hat{\sigma}_i^2 = (r_i(1-h_i)^{-1/2})^2$ for all i. To explicate these variance estimators, define $y_i^* = x_i^T\hat{\beta} + \hat{\sigma}_i t_i^*$, $i = 1, ..., n$; the $t^* = \{t_i^*\}_1^n$ obtain from resampling (denoted by *) with

$$E_* t^* = 0, \text{ and } Cov_*(t^*) = I. \tag{8.2}$$

Wu (1986) shows that (8.2) entails $E_*\hat{\beta}^* = \hat{\beta}$ and
$v_b = (X^TX)^{-1}\sum r_i^2(1-h_i)^{-1}x_i x_i^T(X^TX)^{-1}$. The latter is consistent in general.

Remarks

1 When $\sigma_i = \sigma$ for all i, v_b is unbiased and consistent. Indeed Efron's method (Efron 1979) may well be best if $e_1, ..., e_n$ are iid. If not and interest focuses on the distribution of $\hat{\beta}$ (and not just its covariance), the bootstrap encounters serious difficulties. A very simple example demonstrates this.

Example 8.1 *Consider the one dimensional regression model:*

$$Y_i = x_i\beta + e_i. \quad i = 1, ..., n$$

with $Ee_i = 0$, $Var(e_i) = \sigma^2$, *and* $Ee_i^3 = \mu_{i3}$. *Interest focuses on* $\hat{\beta}$*'s third moment as when the Edgeworth expansion and related results for the bootstrap method are required. The LSE is* $\hat{\beta} = (\sum x_i^2)^{-1} \sum x_i Y_i$ *and*

$$E(\hat{\beta} - \beta)^3 = (\sum x_i^2)^{-3} \sum \mu_{i3} x_i^3.$$

But the bootstrap estimate of the third moment is

$$E_*(\hat{\beta}^* - \hat{\beta})^3 = (\sum_1^n x_i^2)^{-3} \sum_1^n x_i^3 n^{-1} \sum_1^n r_i^3. \tag{8.3}$$

In general, (8.3) is biased and inconsistent. This is not surprising since the $\{r_i\}$ *are not exchangeable (the* $\{\mu_{i3}\}$ *being unequal).*

2 The properties of Wu's bootstrap depend on the distribution of t^*. Specifying that distribution becomes a new problem. (Wu's bootstrap uses information not in the original samples. This may make this method nonrobust in more general cases).

3 For heteroscedastic errors, Wu's method encounters the problem identified in (Remark 2.1). We can get a consistent third moment estimate by choosing the distribution of t^* to satisfy $E(t^{*3}) = 1$. But we may well be interested in other quantities as well. Manipulating the distribution of t^* may not simultaneously provide satisfactory estimators for all quantities of interest.

4 As noted by Wu (1986), bootstrap methods are hard to generalize to nonlinear situations; the heteroscedastic bootstrap is based on resampling the residuals which is not feasible in the nonlinear case.

8.2.2 Vector resampling

Freedman (1981) suggests a way of dealing with the correlation model used for a nonhomogeneous errors model. His method draws a simple random sample with replacement (* sampling) $\{(y_i^*,\ x_i^{*T})\}_1^n$ from $\{(y_i,\ x_i^T)\}_1^n$. The method computes the bootstrap LSE from $\{(y_i^*,\ x_i^{*T})\}_1^n$,

$$\hat{\beta}^* \ = \ (\sum_1^n x_i^* x_i^{*T}) \sum_1^n x_i^* y_i^*, \tag{8.4}$$

and the bootstrap covariance estimator

$$v_* \ = \ E_*(\hat{\beta}^* - \hat{\beta})(\hat{\beta}^* - \hat{\beta})^T. \tag{8.5}$$

This approach suffers from several drawbacks. Firstly injudicious use of this method can lead to inconsistent covariance estimators (Wu 1986 gives an example). Secondly, the method fails to incorporate the knowledge that $Var(e_i)$ changes smoothly with x_i when such knowledge obtains. Thirdly, on the general grounds of requiring inference to be conditional on the design $D \ = \ (x_1,\ ,...,\ x_n)$, one should not risk having simulated data sets whose designs $D^* \ = \ (x_1^*,\ ,...,\ x_n^*)$ are very different from D. Fourthly, if n or k is large, computational costs may be quite high. Fifthly, small sample sizes can entail singular design matrices, D^*.

8.3 A new bootstrap

Let $Y_1, ..., Y_n$ be a sequence of independent random variables with observed values, $y_1, ..., y_n$. Suppose for specified functions $\{g_i\}$, θ satisfies $E_\theta\{g_i(Y_i, \theta)\} \ = \ 0$ for all

$i = 1, ..., n$ and θ. Here and in the sequel E_θ denotes the conditional expectation of the $\{Y_i\}$ given θ; var_θ represents the corresponding conditional variance. For simplicity of presentation, we assume θ is a scalar but our results easily extend to vector-valued parameters. We let $\hat{\theta}$ be a solution of the estimating equation

$$\sum_{i=1}^{n} g_i(y_i, \theta) = 0,$$

and consider the distribution of $\hat{\theta} - \theta$.

A standard two step method for approximating that distribution when the $\{Y_i\}$ are independent and identically distributed can be described as follows:

1. use the Taylor expansion to get the approximation

$$\{-\sum \frac{\partial g_i(y_i, \theta)}{\partial \theta}\}(\hat{\theta} - \theta) \approx \sum g_i(y_i, \theta);$$

2. invoke the Central Limit Theorem to get

$$\sqrt{n}(\hat{\theta} - \theta) \approx N[0, \frac{nvar_\theta\{\sum g_i(y_i, \theta)\}}{E_\theta\{\sum \frac{\partial g_i(y_i, \theta)}{\partial \theta}\}^2}].$$

The standard bootstrap method exploits this approximation when the $\{Y_i\}$ are independent and identically distributed. However, that condition often fails. Then drawing a bootstrap sample directly from $\{y_1, ..., y_n\}$ proves unproductive, leading us to our alternative bootstrap method. First replace θ by $\hat{\theta}$ in $\sum g_i(y_i, \theta)$. Then define $z_i = g_i(y_i, \hat{\theta})$. Finally:

1. draw the bootstrap sample $\{z_1^*, ..., z_n^*\}$ from $\{z_1, ..., z_n\}$ as a simple random sample with replacement;

2. compute the bootstrap estimator as

$$\hat{\theta}^* = \hat{\theta} + \{-\sum \frac{\partial g_i(y_i, \hat{\theta})}{\partial \theta}\}^{-1} \sum z_i^*;$$

3. use the bootstrap, *i.e* empirical distribution of the $(\hat{\theta}^* - \hat{\theta})$'s obtained after many repetitions of 1-2 to approximate the distribution of $\hat{\theta} - \theta$.

Since our method approximates the distribution of $\hat{\theta} - \theta$, we get approximations to the distributions needed for inferences based on that distribution. The quality of those approximations depend of course, on the quality of our underlying approximation to the distribution. That notwithstanding, our method offers great flexibility. And in a manner of Liu (1988), we can prove that our bootstrap yields a better approximation than its competitors under certain conditions. However, our method cannot be used to estimate the bias.

The following example shows how our method differs from its conventional relative. Observations are made of n independent and identically distributed random variables, each with an unspecified probability distribution function, F. Inferential interest focuses on the mean, μ, of F. The usual bootstrap would: (i) draw the "bootstrap sample" $\{y_1^*, ..., y_n^*\}$ from $\{y_1, ..., y_n\}$ as a simple random sample with replacement; (ii) calculate the bootstrap sample mean $\bar{y}^* = n^{-1} \sum y_i^*$; and (iii) repeat (i)-(ii) sequentially to obtain a sequence of $\bar{y}^*$ values. The empirical distribution of the $(\bar{y}^* - \bar{y})$'s is the bootstrap approximation to the sampling distribution of $\bar{y} - \mu$.

Our approach begins with the sampling distribution model

$$Y_i = \mu + e_i \qquad i = 1, ..., n.$$

Given a sample $\{y_i\}$ generated by this distribution, we readily find that the least squares estimate of μ satisfies the "estimating equation ", $\sum(y_i - \mu) = 0$; for simplicity here and in the sequel we (usually) suppress the upper and lower summation limits, $i = 1$ and $i = n$. This estimating equation relies on the component functions $y_i - \mu$, $i = 1, ..., n$,

which may be estimated by $z_i = y_i - \bar{y},\ i = 1, ..., n$. Our method: (i) draws a bootstrap sample, $\{z_1^*, ..., z_n^*\}$ from $\{z_1, ..., z_n\}$ as a simple random sample with replacement; (ii) calculates the bootstrap estimator, $\hat{\mu}^* = \hat{\mu} + n^{-1} \sum z_i^*$; and (iii) repeats (i)-(ii) above sequentially to obtain a sequence of $(\hat{\mu}^* - \hat{\mu})$'s and the bootstrap approximation to the sampling distribution of $\hat{\mu} - \mu$.

From this general discussion of our method we turn in the following sections to linear regression and describe properties of our bootstrap estimator of the regression coefficient vector.

8.4 Asymptotics.

8.4.1 Preamble.

For the linear model (8.1), Efron (1979), Freedman (1981) and Wu (1986) have suggested bootstrap methods described in Section 8.2, based on either the residuals of the least squares fit or the complete observation vectors. Our method differs from theirs.

We start with the normal equations for the ordinary least squares estimate,

$$\sum_{i=1}^{n} x_i(y_i - x_i^T \beta) = 0 \tag{8.6}$$

and their solution $\hat{\beta}$. Let

$$z_i = x_i(y_i - x_i^T \hat{\beta}),\ i = 1, ..., n.$$

The bootstrap estimator defined in Section 8.3 becomes

$$\hat{\beta}^* = \hat{\beta} + (X^T X)^{-1} \sum_{i=1}^{n} z_i^*, \tag{8.7}$$

$\{z_1^*, ..., z_n^*\}$ being a bootstrap sample from $\{z_1, ..., z_n\}$.

8.4.2 Consistency of $\hat{\beta}^*$.

For brevity, we now state our main results. All proofs except that of Theorem 8.1, and underlying assumptions appear in Section 8.8.

Theorem 8.1 *Suppose Assumptions 8.1 and 8.2 hold for model (8.1). Then*

$$E(v_*) = cov(\hat{\beta})\{1 + O(n^{-1})\},$$

where $v_* = E_*(\hat{\beta}^* - \hat{\beta})(\hat{\beta}^* - \hat{\beta})^T$ *and* E_* *represents expectation with respect to the distribution induced by bootstrap sampling.*

Proof. This is an immediate consequence of Lemma 8.1 in Section 8.8. □

Because we use an estimate of β in resampling, we lose degrees of freedom. This suggests renormalising the terms on the left side of equation (8.6). Two alternatives suggest themselves, $z_i^{(1)} = (1 - n^{-1}k)^{-1/2}x_i(y_i - x_i^T\hat{\beta})$ and $z_i^{(2)} = (1 - h_i)^{-1/2}x_i(y_i - x_i^T\hat{\beta})$, $i = 1, ..., n$, where the $\{h_i\}$ represent the diagonal elements of the hat matrix. The bootstrap estimators corresponding to these two renormalisations are $\hat{\beta}^{*(1)}$ and $\hat{\beta}^{*(2)}$ and the covariance estimators are $v_{*(1)}$ and $v_{*(2)}$. The asymptotic properties of $\hat{\beta}^*$, $\hat{\beta}^{*(1)}$ and $\hat{\beta}^{*(2)}$ are equivalent.

The proof of the next theorem, the counterpart of Theorem 8.1 for the renormalized "bootstrappands", follows immediately from Lemma 1 and so it is omitted.

Theorem 8.2 *For model (8.1),* $Ev_{*(2)} = Cov(\hat{\beta})$ *if the assumptions in (8.24) hold. If Assumptions 8.1 and 8.3 hold as well,* $E(v_{*(i)}) = Cov(\hat{\beta})(1 + O(n^{-1}))$, *for* $i = 1, 2$.

An asymptotic analysis must explore the third moments needed for Edgeworth expansions. The use of such expansions in bootstrap regression models seems to be have been proposed by Navidi (1989) who considers only Efron's method for independent and identically distributed errors. Liu (1988) investigates the third moment bootstrap estimator of $l^T\hat{\beta}$ based on Wu's bootstrap for the case of heteroscedastic errors, l being a fixed $k \times 1$ vector. For model (8.1), we also consider the sampling distribution of the least squares estimator of any such specified linear functional of β, $l^T\beta$.

For model (8.1), let

$$E(e_i^3) = \mu_{3,i}, \quad E(e_i^4) = \mu_{4,i}, \quad i = 1, ..., n. \tag{8.8}$$

From the bootstrap estimate (8.7), the third and fourth moment estimators for $l^T\beta$ are defined as $\mu_{3,*} = E_*\{l^T(\hat{\beta}^* - \hat{\beta})\}^3$ and $\mu_{4,*} = E_*\{l^T(\hat{\beta}^* - \hat{\beta})\}^4$.

It is easily shown that

$$\mu_{3,*} = \sum_{i=1}^{n} w_i^3 r_i^3, \tag{8.9}$$

and

$$\mu_{4,*} = \sum_{i=1}^{n} w_i^4 r_i^4 + \sum_{i \neq j} w_i^2 w_j^2 r_i^2 r_j^2, \tag{8.10}$$

where $w_i = l^T(X^TX)^{-1}x_i$ and $r_i = y_i - x_i^T\hat{\beta}$.

With the notation of (8.8), the third and fourth moments of $l^T\hat{\beta}$ are

$$\mu_3 = E\{l^T(\hat{\beta} - \beta)\}^3 = \sum_{i=1}^{n} w_i^3 \mu_{3,i}, \tag{8.11}$$

and

$$\mu_4 = E\{l^T(\hat{\beta} - \beta)\}^4 = \sum_{i=1}^{n} w_i^4 \mu_{4,i} + \sum_{i \neq j} w_i^2 w_j^2 \sigma_i^2 \sigma_j^2. \tag{8.12}$$

Theorem 8.3 *Assume the elements of l are bounded. Then under model (8.1) and (8.8),*

(i) $E(\mu_{3,*}) = \mu_3 + O(n^{-3})$ *when Assumptions 8.1 and 8.5 hold and*

(ii) $E(\mu_{4,*}) = \mu_4 + O(n^{-3}) = \mu_4\{1 + O(n^{-1})\}$ *when Assumptions 8.1 and 8.6 hold.*

Remarks:

1 The covariance estimators v_*, $v_{*(1)}$ and $v_{*(2)}$ corresponding to the bootstrap estimator triplet $\hat{\beta}^*$, $\hat{\beta}^{*(1)}$ and $\hat{\beta}^{*(2)}$ are asymptotically equivalent. As well, in estimating $cov(\hat{\beta})$, $v_{*(2)}$ is unbiased for homoscedastic models and consistent for heteroscedastic models. So it has the desiderata set out by Wu (1986, Section 5).

2 Wu's method (Wu 1986) fails in general to give a consistent estimator of the third moment. This difficulty can be overcome with Wu's approach simply by modifying the distribution of his random variable t to achieve a consistent third moment estimator (Liu 1988). But then an inconsistent fourth moment estimator may result. By contrast, our method yields consistent 3rd and 4th moment estimators.

8.4.3 Asymptotic normality of $\hat{\beta}^*$.

In this section, we describe the asymptotic distribution of $\hat{\beta}^*$ under general conditions including the case of independent and identically distributed errors and the correlation model as special cases. Freedman (1981) investigates the asymptotic theory of Efron's residual resampling method for the case of independent and identically distributed errors and the vector resampling method for the correlation model.

To simplify the statement of our results, denote by $\mathcal{L}(U|V)$ the distribution of U conditional on V for any random objects U and V and P_n denotes the probability distribution

condition on $\{(y_i,\ x_i^T),\ i = 1, ..., n\}$. Let $N_k(\mu, \Gamma)$ be the multivariate normal distribution with mean μ and covariance Γ.

The next result involves two distinct sampling processes, the first leading to the original data set of size n, and the second created by bootstrap resampling, say m times. With the original sample fixed classical limit theory applies when $m \to \infty$. But complications arising from the need to allow $n \to \infty$ entail more delicate analysis, as emphasized in the work of Bickel and Freedman (1981). Such analysis leads to the following results.

Theorem 8.4 *For the regression model (8.1) with fixed regressor variables, suppose Assumptions 8.7-8.10 obtain. Then*

$$\mathcal{L}[\sqrt{n}(\hat{\beta}^* - \hat{\beta})|\{(y_i, x_i^T) : i = 1, ..., n\}] \to N_k(0, V^{-1}WV^{-1})$$

for almost all sequences $\{(y_i, x_i^T) : i = 1, ..., n\}$ *as* $n \to \infty$.

Theorem 8.5 *For the regression model (8.1), with random regressor variables, suppose Assumptions 8.11-8.13 obtain. Then* $\mathcal{L}[\sqrt{n}(\hat{\beta}^* - \hat{\beta})|\{(y_i, x_i^T) : i = 1, ..., n\}] \to N_k(0, \Omega_1^{-1} M \Omega_1^{-1})$ *for almost all sequences* $\{(y_i, x_i^T) : i = 1, ..., n\}$ *as* $n \to \infty$, *where* $M = E x_i x_i^T e_i^2$.

Remarks:

3 Freedman (1981) suggests different bootstrap methods for different models. If the model is changed, the normality of the bootstrap method can be lost. In contrast, Theorems 8.4 and 8.5 assert that the one bootstrap method discussed in Section 8.4.1 yields asymptotic normality for both models.

8.5 Comparison and Simulation Study.

In this section we compare our method with those of Efron (1979), Wu (1986) and Freedman (1981). We do this by highlighting the simple structural differences between these methods in the way they simulate resampling of the regression coefficient estimation vectors. Then we present the results of simulation studies in support our method.

Suppose we wish to estimate the distribution of

$$\hat{\beta} - \beta = (X^T X)^{-1} \sum x_i e_i.$$

Then we need only estimate the distribution of $\sum x_i e_i$, $(X^T X)^{-1}$ being fixed.

Efron's bootstrap resamples from the residuals $r_1, ..., r_n$, $r_i = y_i - x_i\hat{\beta}$ and determines the corresponding bootstrap estimators by

$$\hat{\beta}^* - \hat{\beta} = (X^T X)^{-1} \sum x_i r_i^*.$$

Efron's $\{r_i^*\}$ are independent and identically distributed samples from $\{r_i\}$. While this procedure seems natural for exchangeable residuals $\{r_i\}$, doubt about the validity of his procedure arises when they are not. And these residuals are definitely not exchangeable in heteroscedastic regression models, where Efron's estimators can be inconsistent.

In Wu's approach, bootstrap estimators are found from

$$\hat{\beta}^* - \hat{\beta} = (X^T X)^{-1} \sum x_i r_i t_i^*,$$

the independent and identically distributed sampling model, denoted by *, for $t^* = (t_1^*, ..., t_n^*)$ satisfying

$$E_*(t^*) = 0, \ and \ cov_*(t^*) = I.$$

So we see that Wu's method uses $\sum x_i r_i t_i^*$ to simulate $\sum x_i e_i$. The distribution of $\sum x_i r_i t_i^*$ will depend on the choice of that of the $\{t_i^*\}$. Specifying that distribution becomes a new problem and in any event, means that Wu's bootstrap requires information not in the original sample. In general the method risks possible nonrobustness. And if n is insufficiently large, the distribution of $\sum x_i r_i t_i^*$ can vary greatly for varying distributions of t^*.

Freedman (1981) suggests a way of dealing with the correlation model used for a non-homogeneous errors model. His "pairs" method draws a simple random sample with replacement, $\{(y_i^*, x_i^*)\ i = 1, ..., n\}$ from $\{(y_i, x_i)\ i = 1, ..., n\}$ and computes the bootstrap least squares estimate from $\{(y_i^*, x_i^*)\ i = 1, ..., n\}$,

$$\hat{\beta}^* = (\sum_{i=1}^{n} x_i^* x_i^{*T})^{-1} \sum_{i=1}^{n} x_i^* y_i^*.$$

Then the distribution of $\hat{\beta}^* - \hat{\beta}$ approximates that of $\hat{\beta} - \beta$. As indicated in Wu (1986), Hinkley (1988) and Section 8.2, this approach suffers from several drawbacks.

As shown in equation (8.7), our approach uses

$$\hat{\beta}^* - \hat{\beta} = (X^T X)^{-1} \sum z_i^*.$$

The permutability of the $z_1, ..., z_n$ in (8.6) leads us to make the $\{z_i\}$ independent and identically distributed. We see that of the four methods, ours is the most direct in simulating the distribution of interest.

We now compare the four methods again, this time through a simulation study designed to compare their performance. From the simulation study of Wu (1986), we know that our bootstrap method will perform well in variance-covariance estimation. We know this because our method has the desiderata set out by Wu (1986, Section 5).

We study the following regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1(1)10,$$

$$x_i = 0,\ 1,\ 1.5,\ 2,\ 3,\ 3.5,\ 4,\ 4.5,\ 4.75,\ 5.$$

Two error distributions are used in our study: equal variances $e_i \sim N(0,1)$ and unequal variances $e_i \sim x_i N(0,1)$ $i = 1, ..., n$. In all cases, the $\{e_i\}$ are independent.

We consider four bootstrap methods: (i) Efron's method; (ii) Wu's method, the distribution of t_i^* being given by the "wild bootstrap" (Hardle and Marron 1991); this choice for the distribution of the $\{t_i^*\}$ ensures fulfilment of 1st, 2nd and 3rd moments conditions (Liu 1988); (iii) Freedman's "pairs" method; (iv) our new bootstrap method.

We have run many simulations and chosen two Figures to illustrate the results obtained for the homoscedastic and heteroscedastic model simulation experiments. Each figure displays the cumulative frequency plots for simulated error distributions. They depict the distributions associated with each of the four methods under investigation, labelled b through e. That associated with the true estimation error distribution is labelled by a. The five variables underlying the displays a through e are represented generically by $w = \hat{\beta}_o - \beta_o$. A single simulation run begins with a $\hat{\beta}$ from a single sample generated in accord with the sampling distribution. Then 1000 bootstrap values w^* of w are generated by each of the four methods under consideration.

The results in Table 1 and Table 2 unlike those in the Figures offer an aggregate view of performance and are based on 500 simulations of B = 1000 bootstrap samples.

We use the Kolmogorov-Smirnov statistic to index the accuracy various distribution estimators. In Table 1, the Kolmogorov-Smirnov statistic is defined as $\sup |F^*(w) - F(w)|$, where $F^*(\cdot)$ denotes the empirical distribution of w^* and $F(\cdot)$ that of the true

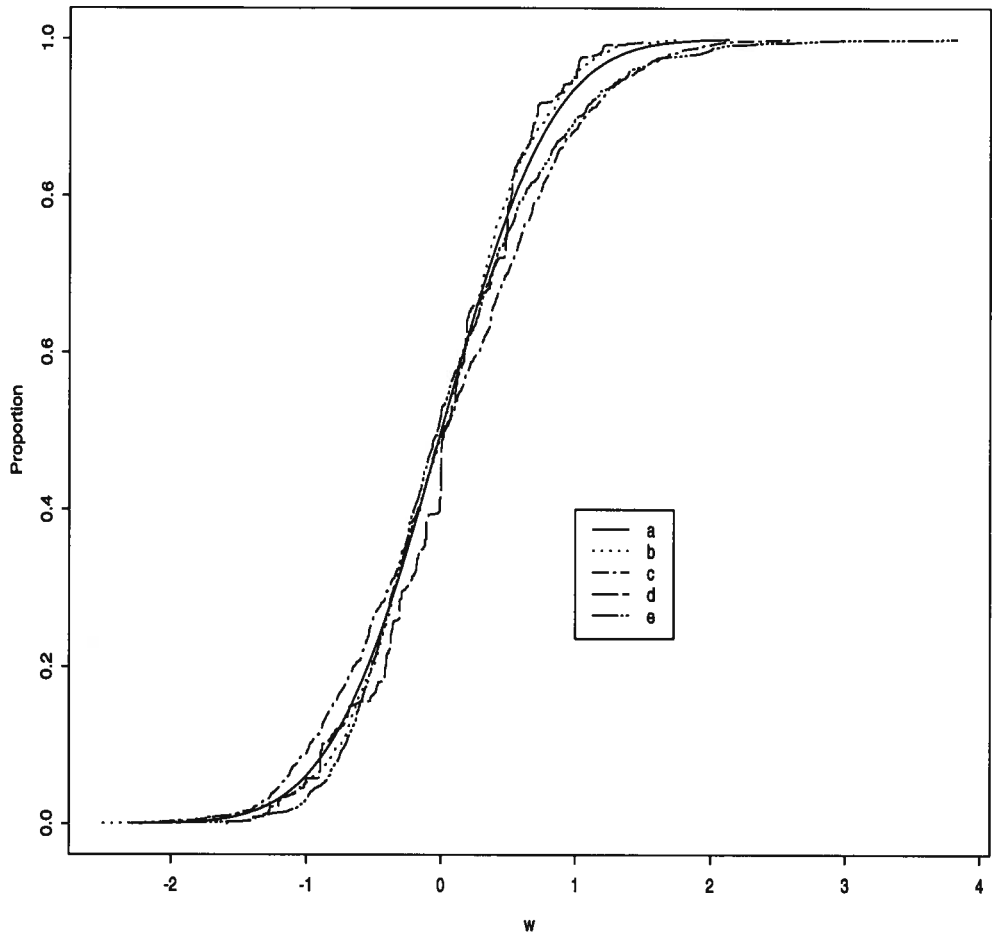

Figure 8.1: *A comparison of bootstrap distribution estimators for regression with homoscedastic errors. We depict the distributions of* $w = \hat{\beta}_0 - \beta_0$ *induced by the true distribution (labelled a), using our bootstrap estimator (b), Efron's estimator (c), Wu's estimator (d), and Freedman's estimator (e).*

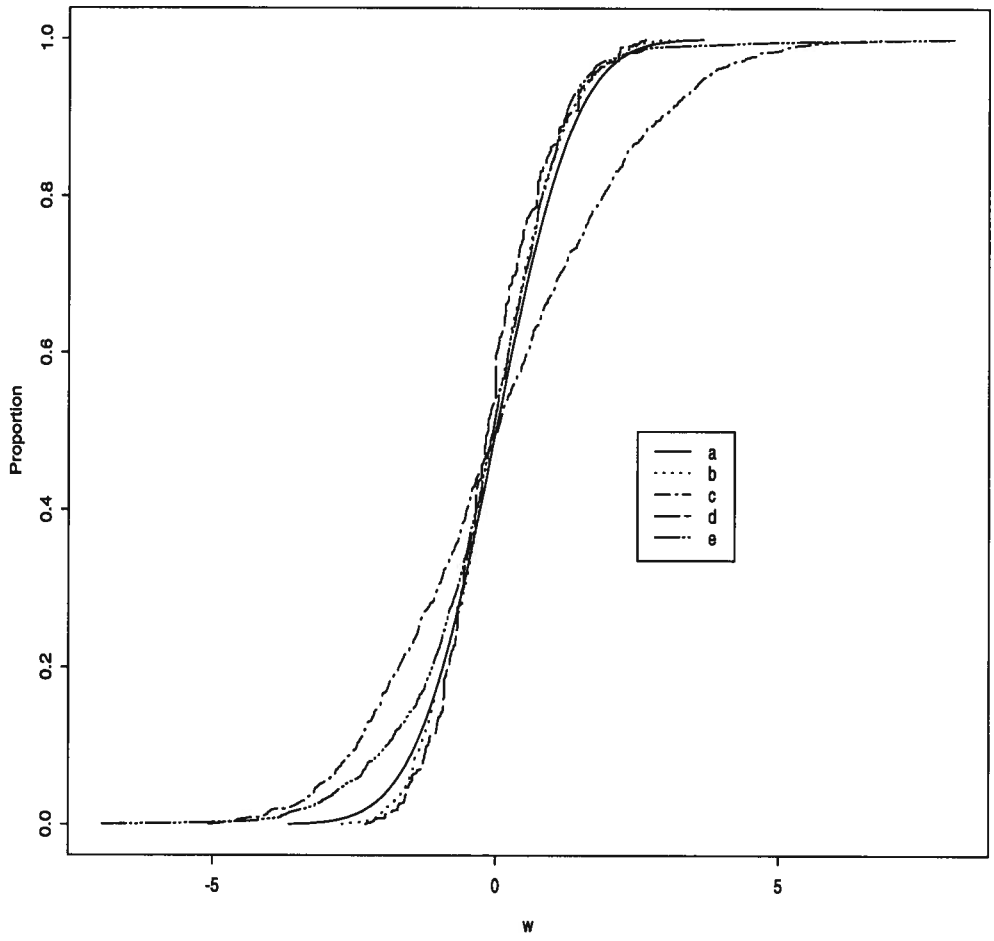

Figure 8.2: *A comparison of bootstrap distribution estimators for regression with heteroscedastic errors. We depict the sampling distributions of* $w = \hat{\beta}_0 - \beta_0$ *induced by the true distribution (labelled a), using our bootstrap estimator (b), Efron's estimator (c), Wu's estimator (d), and Freedman's estimator (e).*

Table 8.1: Averages of the Kolmogorov-Smirnov Statistics for Competing Bootstrap Distribution Estimators

	Equal Variance	Unequal Variance
New Method	0.115 (0.0029)	0.080(0.0021)
Efron	0.079(0.0019)	0.136(0.0028)
Wu	0.158(0.0024)	0.113(0.0019)
Pair	0.126(0.0025)	0.095(0.0017)

Table 8.2: Absolute Biases of the Competing Bootstrap Quantile Estimators

	Equal Variance		Unequal Variance	
	5%	95%	5%	95%
New Method	0.328(0.009)	0.320(0.009)	0.411(0.014)	0.403(0.014)
Efron	0.224(0.007)	0.220(0.007)	1.298(0.041)	1.276(0.039)
Wu	0.344(0.009)	0.338(0.009)	0.436(0.014)	0.431(0.014)
Pair	0.345(0.012)	0.343(0.011)	0.785(0.033)	0.680(0.030)

distribution. The mean values of the Kolmogorov-Smirnov statistics from these 500 simulations are summarized in Table 1. The value besided the mean is the correspond standard error.

Since the confidence interval is one commonly used inferential procedure, studying the quantile estimators obtained from the four methods seemed worthwhile. The absolute bias of 95%-quantile estimator is defined as $|q_{.95}(F^*) - q_{.95}(F)|$, where $q_{.95}(F)$ denotes the 95%-quantile for the distribution, F. The mean of the absolute biases of 5%- and 95%-quantile estimators are summarized in Table 2. The value besided the mean is the correspond standard error.

The results of our simulation studies can be summarized as follows.

1. As expected, for estimating a distribution Efron's method performs best when the error distributions are independent and identically distributed (see Table 1). The new bootstrap method also works well in that case, though not quite as well as Efron's. Freedman's pairs and Wu's method seem to yield rough estimators and to be qualitatively poorer than the other two methods. This may be because Wu's method stresses consistency of the estimator's moments at the expense of estimating the overall distribution (see Figure 1 also).

   From Table 1, we see that in regression with heteroscedastic errors, Efron's approach does badly, again in accord with our expectations; the model errors are not exchangeable and Efron's estimator is not consistent. The new bootstrap method works best, Freedman's reasonably well and Wu's not so well (see also Figure 2).

2. For quantile estimation, Efron's method performs best for both 5%- and 95%-quantiles when the error distributions are independent and identically distributed (see Table 2). The new method ranks second and Wu's third.

   For regression with heteroscedastic errors, the new method is best, Wu's second and Freedman's third. However, Efron's method gives an inconsistent estimator and seems totally unsatisfactory.

3. Overall, we can conclude that Efron's method is not robust against heteroscedastic errors. Wu's method seems unsatisfactory for estimating a distribution. We found Freedman's method effective but computationally expensive in the examples considered; presumably the method would be unrealistic for use in nonlinear situations. Even with just two parameters, the method required much more time than the others.

We next consider the use of pivotal quantities in bootstrapping. The bootstrap-t statistic

Table 8.3: The Pivot Quantile Estimators

	0.05	0.10	0.25	0.50	0.75	0.90	0.95
	Equal Variance						
New Method	-1.72	-1.34	-0.70	0.052	0.72	1.30	1.67
Efron	-1.86	-1.36	-0.71	-0.004	0.71	1.49	1.97
Pair	-1.51	-1.08	-0.60	0.122	0.58	1.16	1.61
t(8)	-1.86	-1.40	-0.71	0	0.71	1.40	1.86
N(0,1)	-1.65	-1.28	-0.67	0	0.67	1.28	1.65
	Unequal Variance						
New Method	-1.77	-1.34	-0.69	-0.029	0.68	1.31	1.84
Efron	-1.99	-1.55	-0.76	-0.033	0.64	1.36	1.86
Pair	-1.63	-1.06	-0.55	-0.011	0.45	0.90	1.17

is constructed as

$$t^*(b) = (\hat{\beta}^*(b) - \hat{\beta})/\hat{se}^*(b), \tag{8.13}$$

where b denotes bootstrap sample and se means "standard error". In view of our earlier findings, we consider only three bootstrap methods, Efron's, Freedman's, and ours.

For both the case of equal and that of unequal variances, we ran a single simulation. The pivot quantile estimators in Table 3 are based on 1000 bootstrap samples. In Table 3, $t(8)$ denotes the t-distribution with degree of freedom 8 and $N(0,1)$ denotes the standard normal distribution.

For the case of equal variances, $t(8)$ is a good reference distribution. From Table 3, we find that both Efron's method and the new bootstrap method perform well. The pairs estimator is not as good as that given by the other two methods.

For the case of unequal variances, we do not know a good reference distribution. If we use $t(8)$ as before, both Efron's and the new method give good estimators. The pairs method

fails to give reasonable results. But since Efron's method yields an inconsistent variance estimator, we cannot use Efron's method to construct a useful confidence interval.

8.6 Bootstrapping in Nonlinear Situations

In this section we merely sketch extensions of our bootstrap in the nonlinear situations considered by Wu (1986). The extensive detail needed for the required analysis of lower-order terms will be presented elsewhere.

8.6.1 Regression M-estimator

An M-estimate $\tilde{\beta}$ of the regression coefficient vector β is found by minimizing

$$\sum_{1}^{n} \rho(y_i - x_i^T \beta); \tag{8.14}$$

over β, where ρ is usually assumed to be symmetric. Choices of ρ can be found in Huber (1981). The M-estimate is found by solving:

$$\sum_{1}^{n} x_i \rho'(y_i - x_i^T \beta) = 0 \tag{8.15}$$

Call $\sum_1^n x_i \rho'(y_i - x_i^T \beta)$ the "score function". From the method of Section 8.3, we obtain a bootstrap sample from the estimated score function with β replaced by $\tilde{\beta}$. To be precise, let $z_i = x_i \rho'(y_i - x_i^T \tilde{\beta})$, $i = \cdots, n$. Next: **(1)** draw the bootstrap sample, $\{z_1^*, \ldots, z_n^*\}$, a simple random sample with replacement from $\{z_1, \ldots, z_n\}$; **(2)** construct the bootstrap M-estimator $\tilde{\beta}^*$ as $\tilde{\beta}^* = \tilde{\beta} + (\sum x_i x_i^T \rho''(\tilde{e}_i))^{-1} \sum z_i^*$, where $\tilde{e}_i = y_i - x_i^T \tilde{\beta} : i = 1, \ldots, n$.

For estimating $Var(\tilde{\psi})$, $\tilde{\psi} = \psi(\tilde{\beta})$, we use an analogue of the estimator in Section 8.4, $v_* = E_*(\tilde{\psi}^* - \tilde{\psi})((\tilde{\psi}^* - \tilde{\psi})^T$ where $\tilde{\psi}^* = \psi(\tilde{\beta}^*)$. Principal interest centers on $\tilde{\psi} = \tilde{\beta}$, and $v_* = E_*(\tilde{\beta}^* - \tilde{\beta})(\tilde{\beta}^* - \tilde{\beta})^T$ where

$$v_* = (\sum_1^n x_i x_i^T \rho''(\tilde{e}_i))^{-1}(\sum_1^n x_i x_i^T r_i^2)(\sum_1^n x_i x_i^T \rho''(\tilde{e}_i))^{-1} \tag{8.16}$$

and $r_i = \rho'(y_i - x_i^T\tilde{\beta})$.

Now let us calculate the asymptotic approximation covariance matrix of $\tilde{\beta}$. Under appropriate conditions (not stated here for brevity) $0 = \sum x_i\rho'(y_i - x_i^T\tilde{\beta}) = \sum x_i\rho'(y_i - x_i^T\beta) - (\tilde{\beta} - \beta)\sum x_i x_i^T\rho''(y_i - x_i^T\beta) + o(\tilde{\beta} - \beta)$. This implies $\tilde{\beta} - \beta \simeq [\sum x_i x_i^T\rho''(y_i - x_i^T\beta)]^{-1}\sum x_i\rho'(y_i - x_i^T\beta)$. Alternatively, $v = E(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T$ where

$$v \simeq [\sum_1^n x_i x_i^T\rho''(e_i)]^{-1}[\sum_1^n x_i x_i^T(\rho'(e_i))^2][\sum_1^n x_i x_i^T\rho''(e_i)]^{-1}; \tag{8.17}$$

here $e_i = y_i - x_i^T\beta : i = 1, ..., n$. Comparison of (8.16) and (8.17) suggest the bootstrap estimator is a reasonable estimate of the covariance matrix of $\tilde{\beta}$.

8.6.2 Nonlinear regression.

In nonlinear regression $y_i = f_i(\beta) + e_i \quad i = 1, ..., n$, where f_i is a nonlinear smooth function of β and e_i satisfies (8.1). The LSE $\hat{\beta}$ is obtained by minimizing $\sum(y_i - f_i(\beta))^2$. The score function is $\sum x_i(\beta)(y_i - f_i(\beta))$, where $x_i = \partial f_i/\partial\beta$. Put the LSE into the score function. The bootstrap method of Section 8.3 is with samples, $\{z_1^*, ..., z_n^*\}$ from $\{z_i = x_i(\hat{\beta})(y_i - f_i(\hat{\beta}))\ i = 1, ..., n\}$. The bootstrap coefficient estimators are $\hat{\beta}^* = \hat{\beta} + (X(\hat{\beta})^T X(\hat{\beta}))^{-1}\sum_1^n z_i^*$, where $X = (x_1, ..., x_n)$. It is easily shown that $E_*\hat{\beta}^* = \hat{\beta}$ and $v_* = E_*(\hat{\beta}^* - \hat{\beta})(\hat{\beta}^* - \hat{\beta})^T$, with

$$v_* = (X(\hat{\beta})^T X(\hat{\beta}))^{-1}\sum_1^n x_i(\hat{\beta})x_i(\hat{\beta})^T r_i^2(X(\hat{\beta})^T X(\hat{\beta}))^{-1}, \tag{8.18}$$

where $r_i = y_i - f_i(\hat{\beta})$.

If we calculate the asymptotic approximation covariance matrix of the LSE, $\hat{\beta}$, we get $\hat{\beta} - \beta \simeq (X(\beta)^T X(\beta))^{-1} X(\beta)e$ and $v = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T$ where

$$v \simeq (X(\beta)^T X(\beta))^{-1} \sum_1^n x_i(\beta) x_i(\beta)^T \sigma_i^2 (X(\beta)^T X(\beta))^{-1}. \tag{8.19}$$

From (8.18) and (8.19), we find that the bootstrap estimator is consistent and robust in general.

8.6.3 Generalized Linear models

A generalized linear model is characterized by three components (McCullagh and Nelder, 1989):

(i) an error distribution from the exponential family $f_\phi(y;\psi) = \exp([y\psi - a(\psi) + b(y,\psi)]/\phi)$,

(ii) a systematic component $\eta = x_i\beta$, and

(iii) a monotonic, differentiable link function $\mu = g(\eta)$, where $\mu = E(y)$.

Here we consider generalized linear models with independent observations. Let $y = (y_1, \ldots, y_n)$, $E(y) = \mu = (\mu_1, \ldots, \mu_n)^T$ and

$$Cov(y) = diag(\sigma_1^2, \ldots, \sigma_n^2)\ V(\mu) = diag(\sigma_i^2 \nu_i(\mu)). \tag{8.20}$$

The mean, μ_i, is related to the regressor, x_i, by the link function g, i.e. $\mu_i = g(x_i^T\beta)$. The full likelihood may not be available. Inference is instead based on the log quasilikelihood

(see Wedderburn 1974, McCullagh 1983), $L(\mu, y)$, defined by

$$\partial L(\mu, y)/\partial \mu \;=\; V(\mu)^{-1}(y - \mu)$$

A generalized least squares estimator (GLSE), $\hat{\beta}$, is defined as a solution of

$$D^T V^{-1}(y - \mu(\beta)) \;=\; 0, \tag{8.21}$$

where $\mu(\beta) = (\mu_i)_1^n = (g(x_i^T\beta))_1^n$ and $D = d\mu/d\beta = diag(g'(x_i^T\beta))X$. Here $D^T V^{-1}(y - \mu(\beta))$ is called the quasi-score function.

Moulton and Zeger (1991) consider two bootstrap methods for generalized linear models. One corresponds to standardized Pearson residual resampling, which extends Efron's residual bootstrapping method using standardized Pearson residuals. The method is similar to that for the ordinary regression model, with heteroscedastic errors (i.e. unequal σ_i^2's in (8.20)); the bootstrapped covariance estimator is in general inconsistent for the true covariance of GLSE, $\hat{\beta}$. The other method of Moulton and Zeger, involves observation vector resampling; this extends ordinary vector resampling using a one-step approach.

We use the bootstrap method proposed in Section 8.3. By substituting the estimator $\hat{\beta}$ in (8.21) and rewriting it, we find the quasi-score function

$$\sum x_i g'(x_i^T\hat{\beta})\nu_i(\mu)^{-1}(y_i - \hat{\mu}_i)$$

Let $z_i = x_i g'(x_i^T\hat{\beta})\nu_i(\mu)^{-1}(y_i - \hat{\mu}_i)$, $i = 1, \ldots, n$. The bootstrap is thus based on the uniform distribution, $\tilde{F}_b$ with support $z_1, \ldots, z_n$ and the calculation of the bootstrap

coefficients by

$$\hat{\beta}^* = \hat{\beta} + (D^T V^{-1} D)^{-1} \sum_1^n z_i^*; \tag{8.22}$$

here we have inserted the estimator $\hat{\beta}$ into the righthand side of the (8.22). For estimating $Var(\hat{\psi})$, $\hat{\psi} = \psi(\hat{\beta})$, the bootstrap covariance estimator is $v_* = E_*(\hat{\psi}^* - \hat{\psi})(\hat{\psi}^* - \hat{\psi})^T$. For $\psi = \beta$, $v_* = E_*(\hat{\beta}^* - \hat{\beta})(\hat{\beta}^* - \hat{\beta})^T$ where

$$v_* = (D^T V^{-1} D)^{-1} \sum_1^n (g'(x_i^T \hat{\beta}) \nu_i(\mu)^{-1})^2 r_i^2 x_i x_i^T (D^T V^{-1} D)^{-1},$$

where $\{r_i = y_i - \hat{\mu}_i : i = 1, ..., n\}$ are the prediction errors.

For the GLSE, $\hat{\beta}$, we have the approximation (McCullagh, 1983)

$$\hat{\beta} - \beta \simeq (D^T V^{-1} D)^{-1} D^T V^{-1} (y - \mu),$$

so that

$$Cov(\hat{\beta}) \simeq (D^T V^{-1} D)^{-1} D^T V^{-1} Cov(y - \mu) V^{-1} D (D^T V^{-1} D)^{-1}.$$

Under model (8.20) $Cov(y - \mu) = diag(\sigma_i^2 \nu_i(\mu))$ and

$$Cov(\hat{\beta}) = (D^T V^{-1} D)^{-1} \sum (g'(x_i^T \beta))^2 \nu_i^{-1}(\mu) \sigma_i^2 x_i x_i^T (D^T V^{-1} D)^{-1}.$$

We show elsewhere v_* is a consistent estimator for $Cov(\hat{\beta})$.

The homoscedastic model ($\sigma_i^2 = \sigma^2$) is the most interesting special case; then the covariance of $\hat{\beta}$ is $Cov(\hat{\beta}) = \sigma^2 (D^T V^{-1} D)^{-1}$ and the bootstrap covariance estimator becomes

$$v_* = (D^T V^{-1} D)^{-1} D V^{-1} Diag\{r_i^2\} V^{-1} D (D^T V^{-1} D)^{-1}. \tag{8.23}$$

The bootstrap covariance estimator (8.23) is the same as that employed by Cox (1961), Huber (1967), White (1982), and Royall (1986) for handling quite general model misspecification.

8.7 Concluding Remarks.

The conventional approach to the bootstrap has been through quasi replicates of the original sample possibly subject to reweighting or some constraints. The bootstrap distribution is then obtained from the empirical distribution or some smoothed version of the empirical distribution of the succession of realized estimators obtained from these quasi replicate samples.

Even when the resampling population consists of (possibly renormalized) model fit residuals the approach has been to construct successive pseudo estimator values and then the bootstrap distribution for the estimator of interest, through the construction of bootstrap data sets.

What we have done is to by-pass the data sets. Instead we have taken the estimator of interest as the baseline, generated successive estimator fit residuals (rather than model fit residuals). The approximate sampling distribution for the estimator is then found by taking the resulting empirical distribution of the realizations of estimator plus residual. A critical feature of our approach is the use of the components of the estimating function itself to transform to residuals to the appropriate scale. Heuristically, we have perturbed the estimator by bootstrapped realizations of a normalized first order term from a Taylor expansion of the estimator about the true value of the parameter. This approach seems natural. Generating bootstrap replicates of the original sample seems unnecessary when the object of interest is the sampling distribution of an estimator and it can be realized directly from suitably expressed estimator fit residuals.

The idea of by-passing the data sets is in itself not new. Moulton and Zeger (1991) use this idea in what they call their "one step procedure". They intend their procedure

for use in estimating coefficients in link functions of generalized linear models. Their goal is computational simplicity. If they were to follow thw traditional practice of bootstrapping the data set, they would need to find the bootstrap coefficient estimator by an iterative process for each successive bootstrap replicate sample. The resulting intolerable computational burden leads them to use the estimating equations (for the generalized linear models) directly in much the same way as we suggest in this paper for the general case.

Apart from the difference in motivation and intended domain of application, our method differs from that of Moulton and Zeger (1991) in the way we resample residuals. They use a method like that of Wu described above. Their method then inherits the potential deficiencies discussed above of Wu's approach.

Not surprisingly, our bootstrap has the asymptotic properties up to second order which one might expect from the Taylor expansion heuristic. However, the unexpected robustness of our method up to at least nonhomogeneity in four order moments is unexpected and encouraging. Overall our method has promise. As well, its underlying Taylor expansion heuristic suggests other ways of perturbing the estimator to approximate its sampling distribution and these are currently under investigation.

We noted above a number of potential applications of our method. In current work we are adapting our method for use with longitudinal count data series. As well we are seeing if our method can be used to find standard errors for estimators computed from data obtained through complex survey designs. Binder (1991), building on the work of Godambe and Thompson (1986), has shown how estimating equations can be used in that context. That work provides the platform on which we are attempting to build our bootstrap variance estimation procedure.

8.8 Proofs

Our asymptotic theory requires certain conditions.

Assumptions 8.1 *For $X_n = X$ the design matrix in (8.1) for n observations, $\max_{1\le i\le n} x_i^T(X_n^TX_n)^{-1}x_i \le c/n$ for some scalar $c > 0$ independent of n.*

Assumptions 8.2 *For the error variances in model (8.1), $\max_{1\le i<\infty}\sigma_i^2 < \infty$.*

Assumptions 8.3 *The minimum and maximum eigenvalues of $n^{-1}X_n^TX_n$ are uniformly bounded away from 0 and ∞.*

Assumptions 8.4 *The elements of X_n are uniformly bounded.*

We also require the following lemma.

Lemma 8.1 (Wu 1986) *If*

$$h_{ij} = x_i^T(X_n^TX_n)^{-1}x_j = 0 \text{ for any } i,j \text{ with } \sigma_i \ne \sigma_j, \tag{8.24}$$

then $E(r_i^2) = (1-h_i)\sigma_i^2$. More generally, Assumptions 8.1 and 8.2 imply $E(r_i^2) = (1-h_i)\sigma_i^2 + O(n^{-1}) = \sigma_i^2 + O(n^{-1})$.

Comparing (8.9) with (8.11), and (8.10) with (8.12) suggests $\mu_{3,*}$ and $\mu_{4,*}$ are robust for estimating the third and four moments, respectively under substantial departures from the model assumptions. We prove this below under additional assumptions.

Assumptions 8.5 $\max_{1\le i<\infty}|\mu_{3,i}| < \infty$;

Assumptions 8.6 $\max_{1\leq i<\infty}|\mu_{4,i}|<\infty$.

To prove the main theorems, we also need the following lemma.

Lemma 8.2 *If* $h_{ij}=x_i^T(X^TX)^{-1}x_j=0$ *for any* $i\neq j$, *then*

$$E(r_i^3)=(1-h_i)^3\mu_{3,i},\quad E(r_i^4)=\ (1-h_i)^4\mu_{4,i}, \tag{8.25}$$

where $h_i=x_i^T((X^TX)^{-1}x_i$. *More generally, Assumptions 8.1 and 8.5 imply*

$$E(r_i^3)=(1-h_i)^3\mu_{3,i}+O(n^{-2})=\mu_{3,i}+O(n^{-1}) \tag{8.26}$$

and Assumptions 8.1, 8.2 and 8.6 imply

$$E(r_i^4)=(1-h_i)^4\mu_{4,i}+O(n^{-1})=\mu_{4,i}+O(n^{-1}). \tag{8.27}$$

Proof. From $r_i=y_i-x_i^T\hat{\beta}=e_i-x_i^T(X^TX)^{-1}X^Te$,

$$r_i=e_i-\sum_{j=1}^{n}h_{ij}e_j=(1-h_i)e_i-\sum_{i\neq j}h_{ij}e_j. \tag{8.28}$$

From (8.28)

$$E(r_i^3)=(1-h_i)^3E(e_i^3)-\sum_{i\neq j}h_{ij}^3E(e_j^3)=(1-h_i)^3\mu_{3,i}-\sum_{i\neq j}h_{ij}^3\mu_{3,j}. \tag{8.29}$$

The first equality in (8.29) follows from the independence of $e_1,...,e_n$. The second inequality in (8.25) obtains in a similar way.

Now more generally when $h_{ij}\neq 0, i\neq j$,

$$|\sum_{i\neq j}h_{ij}^3\mu_{3,j}|\leq n\max\{|\mu_{3,j}|\}\max_{1\leq j\leq n}\{|h_{ij}|^3\}.$$

Assumptions 8.1 and 8.5 imply the right hand side of these last inequalities is of order n^{-2} since ($|h_{ij}|\leq\sqrt{h_ih_j}$) by the Cauchy-Schwarz inequality . Therefore, the first equality

in (8.26) follows from (8.29) and the second from (8.29) and $h_i \leq c/n$. This completes the proof of (8.26).

Now from (8.28)

$$\begin{aligned} E(r_i^4) &= (1-h_i)^4\mu_{4,i} + \sum_{i\neq j} h_{ij}^4\mu_{4,j} + 6(1-h_i)^2\sigma_i^2\sum_{j\neq i} h_{ij}^2\sigma_j^2 \\ &+ 6\sum_{j\neq j(1)\neq i} h_{ij}^2 h_{ij(1)}^2\sigma_j^2\sigma_{j(1)}^2 \\ &= (I)+(II)+(III)+(IV), \end{aligned}$$

say. Assumptions 8.1, 8.2 and the Cauchy-Schwarz inequality imply

$$(IV) \leq 6n^{-2}c^4 \max_j \sigma_j^4 \tag{8.30}$$

$$(III) \leq 12(n^{-1}c + n^{-3}c^4)\max_j \sigma_j^4. \tag{8.31}$$

From Assumptions 8.1, 8.5 and the Cauchy-Schwarz inequality,

$$(II) \leq n^{-3}c^4 \max_j \mu_{4,j} \tag{8.32}$$

$$(I) = (1-4h_i+6h_i^2-4h_i^3+h_i^4)\mu_{4,i} = \mu_{4,i} + O(n^{-1}). \tag{8.33}$$

The first equality in (8.27) follows from (8.30), (8.31) and (8.32) and the second from (8.33). This completes the proof. □

Proof of Theorem 8.3 From Lemma 8.2,

$$\begin{aligned} E(\mu_{3,*}) &= \sum_{i=1}^n w_i^3 E r_i^3 \\ &= \sum_{i=1}^n w_i^3(\mu_{3,i} + O(n^{-1})) \\ &= \sum_{i=1}^n w_i^3\mu_{3,3+6*9i} + \sum_{i=1}^n w_i^3 O(n^{-1}) \\ &= \mu_3 + O(n^{-3}) \end{aligned}$$

The last equality follows from

$$|\sum_{i=1}^{n} w_i^3| \leq n \max_i |w_i^3| = O(n^{-2}).$$

This proves part (i). Result (ii) is reached by similar reasoning. □

Let X_n denote the X matrix in (8.1). Key assumptions for the case of nonrandom $\{x_i\}$ are:

Assumptions 8.7 *The residuals e_i are independent with distribution F_i having mean 0 and variance σ_i^2, both F_i and σ_i^2 being unknown, $i = 1, ..., n$.*

Assumptions 8.8 *There exists a matrix $V > 0$ such that $V_n \stackrel{\text{def}}{=} n^{-1} X_n^T X_n \to V$.*

Assumptions 8.9 *There exists a matrix $W > 0$ such that $W_n \stackrel{\text{def}}{=} n^{-1} \sum \sigma_i^2 x_i x_i^T \to W$.*

Assumptions 8.10 *x_i and $E(e_i^4)$ are uniformly bounded, $i = 1, ..., n$.*

For random $\{x_i\}$, the corresponding assumptions are:

Assumptions 8.11 *The vectors $\{(Y_i, x_i^T)\}$ are independent and identically distributed with common unknown distribution function F on R^{k+1}, and $E \parallel (Y_i, x_i^T) \parallel^4 \leq \infty$, where $\|\cdot\|$ is Euclidean length.*

Assumptions 8.12 *The $k \times k$ covariance matrix $\Omega_1 = E(x_1 x_1^T)$ is positive definite.*

Assumptions 8.13 *$Ex_i e_i = 0$, $i = 1, ..., n$ where $e_i = Y_i - x_i^T \beta$.*

Proof of Theorem 8.4 We will first show $\tilde{W}_n \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n x_i x_i^T r_i^2 \to W$ almost surely as $n \to \infty$ where $r_i = y_i - x_i^T\hat{\beta}$ for all i. Then in bootstrap sampling we may condition on the set of unit probability where the just asserted convergence takes place and let $m \to \infty$.

To obtain this almost sure result, observe that

$$\begin{aligned}\tilde{W}_n &= n^{-1}\sum_{i=1}^n x_i x_i^T (y_i - x_i^T\beta)^2 + n^{-1}\sum_{i=1}^n x_i x_i^T [x_i^T\{\beta - \hat{\beta})\}^2 \\ &+ 2n^{-1}\sum_{i=1}^n x_i x_i^T x_i^T(\beta - \hat{\beta})(y_i - x_i^T\beta) \\ &= (I) + (II) + (III)\end{aligned}$$

say. Assumptions 8.8, 8.9 and Kolmogorov's strong law (*c.f.* Chung, 1968, p 119 and the Corollary given there for the case, $p = 2$) entail $\hat{\beta} - \beta \to 0$ almost surely as $n \to \infty$. Using the trace norm for matrices and the usual norm for vectors, $||(II)|| \leq n^{-1}\sum ||x_i||^4 \; ||(\beta - \hat{\beta})||^2 \leq c^4 \; ||(\beta - \hat{\beta})||$ where $c > 0$ is a uniform upper bound for $||x_i||$ obtained from Assumption 8.10. So $(II) \to 0$ a.e.. At the same time, $||(III)|| = ||2n^{-1}\sum x_i x_i^T x_i^T e_i(\beta - \hat{\beta})|| \leq 2n^{-1}\sum |e_i| \; ||x_i||^3 \; ||\beta - \hat{\beta}|| \leq 2c^3 n^{-1}\sum |e_i| ||\beta - \hat{\beta}||$ for the same constant $c > 0$ used above. So Assumption 8.10 and Kolmogorov's strong law cited above imply $(III) \to 0$ almost surely. Finally, $(I) = n^{-1}\sum x_i x_i^T e_i^2 = n^{-1}\sum x_i x_i^T(e_i^2 - \sigma_i^2) + n^{-1}\sum x_i x_i^T \sigma_i^2$. But $n^{-1}\sum x_i x_i^T(e_i^2 - \sigma_i^2) \to 0$ almost surely by Assumption 8.10 and Kolmogorov's strong law of large numbers. Then $I \to W$ almost surely by Assumption 8.9.

Now

$$\sqrt{n}(\hat{\beta}^* - \hat{\beta}) = \sqrt{n}^{-1} V_n^{-1} \sum_{i=1}^n z_i^*.$$

Let $\xi_i = l^T z_i$, $i = 1, ..., n$ for any fixed k dimensional vector l with $||l|| = 1$. Observe that conditional on the actual sample, the $\{l^T z_i^*, \; i = 1, ..., n\}$ are independent and identically

distributed random variables with zero means and common standard deviation, $\sigma_n = (l^T \tilde{W}_n l)^{1/2} = (n^{-1} \sum \xi_i^2)^{1/2}$. The Berry-Esseen Theorem implies

$$\sup_x |P_n(\sigma_n^{-1} \sqrt{n}^{-1} \sum_{i=1}^{n} l^T z_i^* \leq x) - \Phi(x)| \leq A \rho_n \sqrt{n}^{-1}, \tag{8.34}$$

where $\rho_n = n^{-1} \sum |\xi_i|^3 / (n^{-1} \sum |\xi_i|^2)^{3/2}$. But $n^{-1} \sum |\xi_i|^2 \to l^T W l > 0$ a.e.. Moreover, $n^{-3/2} \sum |\xi_i|^3 \to 0$ as $n \to \infty$ a.e.. The last result obtains since by our assumptions, the $\{E|\xi_i|^4\}$ are uniformly bounded; we may then invoke a general version of the strong law of large numbers (Chung, 1968, p117, Theorem 5.4.1 with $\phi(\cdot) = (\cdot)^{4/3}$ and $a_n = n^{3/2}$) to obtain the result.

Since inequality (8.34) holds for all l, the conclusion of the theorem now follows. □

Corollary 8.1 *Under the assumptions of the last theorem,*

$$\mathcal{L}[\{V_n^{-1} W_n V_n^{-1}\}^{1/2} \sqrt{n}^{-1} (\hat{\beta}^* - \hat{\beta}) | \{(y_i, x_i^T) : i = 1, ..., n\}]$$

converges to the standard normal on R^k for almost all sequences $\{(y_i, x_i^T) : i = 1, ..., n\}$ as $n \to \infty$.

Proof: This result is an immediate consequence of the result of the last theorem. □

Proof of Theorem 8.5 This proof is similar to that of Theorem 8.4 and the details will be omitted. That $\hat{\beta} - \beta \to 0$ almost surely follows from 8.11 and 8.13 and the strong law of large numbers. Assumptions 8.11 and 8.12 imply $(I) \to 0$, $(II) \to 0$, and $(III) \to M$ almost surely. These results and Assumption 8.12 imply the conclusion of the Theorem. □

Bibliography

[1] Akaiki, H. (1973). Information theory and an extension of entropy maximization principle. *2nd International Symposium on Information Theory,* eds. B. N. Petrov and F. Csak, Kiado: Akademia, p. 276-281.

[2] Akaiki, H. (1977). On entropy maximization principle. In P. R. Krishnaiah, (ed.), *Applications of Statistics.* Amsterdam: North-Holland, 27-41.

[3] Akaiki, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann.Inst.Statist.Math.*, 30A, 9-14.

[4] Akaiki, H. (1982). On the fallacy of the likelihood principle. *Statistics and Probability Letters*, 1, 75-78.

[5] Akaike, H. (1983). Information measures and model selection. *Proceedings of the 44th Session of ISI*, 1, 277-291.

[6] Akaike, H. (1985). Prediction and entropy. *A Celebration of statistics, The ISI Centenary Volume*, Berlin, Spring-Verlag.

[7] Bahadur, R.R.(1966). A note on quantiles in large sample, *Ann.Math.Statist.* 37, 577-580.

[8] Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

[9] Basu, D. (1975). Statistical information and likelihood (with discussions). *Sankhya*, Ser. A. 37, 1-71.

[10] Bickel, P.J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* 9, 1301-1309.

[11] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9, 1196-1217.

[12] Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys (with discussion), *Proc. Sect. Res. Methods,* Amer. Statist. Assoc., 34-44.

[13] Birnbaum, A. (1962). On the foundation of statistical inference (with discussions). *JASA*, 57, 269-306.

[14] Booth, J. Hall, P. and Wood, A. (1992). Bootstrap estimation of conditional distributions. *Ann. Statist.*, 20, 1594-1610.

[15] Buehler, R.J. (1971). Measuring information and uncertainty. *Foundations of Statistical Inference.* Eds: Godambe and Sprott; Holt, Rinehart and Winston.

[16] Casella, G. and Strawderman, W.E. (1981). Estimating a bounded normal mean. *Ann. Statist.*, 9, 870-878.

[17] Chow, Y.S. and Lai, T.L. (1973). Limiting behavior of weighted sums of independent random variables. *Ann. of Prob.*, 1, 810-824.

[18] Chu, C.K. and Marron, J.S. (1992). Choosing a kernel regression estimator. *Statistical Science*, 6, 404-436.

[19] Chung, K.L. (1968). *A Course in Probability Theory*, New York: Harcourt, Brace & World, Inc..

[20] Cox, D.R. (1961). Tests for separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, (University of California Press, Berkeley), 105-123.

[21] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* Chapman and Hall, Landon.

[22] Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

[23] Edwards, A.W.F. (1972). *Likelihood.* C.U.P., Cambridge.

[24] Efron,B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7,1-26.

[25] Efron,B. (1982). *The jackknife, the bootstrap and other resampling plans*, SIAM, Philadelphia.

[26] Efron,B. (1987). Better bootstrap confidence intervals and bootstrap approximations. *JASA,* 82, 171-185.

[27] Efron,B. (1990). More efficient bootstrap computations. *JASA*, 85, 79-89.

[28] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to Bootstrap.* Chapman & Hall, New York.

[29] Eubank, P.L. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

[30] Fan, J. (1992). Design-adaptative nonparametric regression. *JASA*, 87, 998-1004.

[31] Fan, J. (1993). Local linear regression smoothers and their Minimax efficiencies. *Ann. Statist.*, 21, 196-216.

[32] Fan, J., Heckman, N.E. and Wand M.P. (1993). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *JASA*, in press.

[33] Fan, J. and Marron, J.S. (1992). Best possible constant for bandwidth selection. *Ann. Statist.*, in press.

[34] Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Cambridge Phil. Soc.*, 22, 700-725.

[35] Fisher, R.A. (1956) *Statistical Methods and Scientific Inference.* Oliver and Boyd. Edinbugh.

[36] Fraser, D.A.S. (1965). On information in statistics. *Ann. Math. Statist.*, 36 890-896.

[37] Fraser, D.A.S. (1968). *The Structure of Inference.* New York: John Wiley.

[38] Freedman, D.A. (1981) Bootstrapping regression models. *Ann. Statist.*, 9, 1218-1228.

[39] Freedman, D.A. and Peters,S.C. (1984). Bootstrapping a regression equation: some empirical results. *JASA,* 79, 97-106.

[40] Gasser, T. and Muller, H.G. (1979). Kernel estimation of regression function. in *Smoothing Techniques of Curve Estimation.* Lecture Notes in Mathematics 757, eds. T. Gasser and M. Resenblatt. Heidelberg: Springer-Verlag, 23-68.

[41] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, 31, 1208-1212.

[42] Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63, 277-284.

[43] Godambe, V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *Int. Statist. Rev.*, 55, 231-244.

[44] Godambe, V.P. and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation, *International Statist. Rev.*, 54, 127-138.

[45] Good, I.J. (1971). The probability explication of information, evidence, surpise, causality, explanation, and utility. *Foundations of Statistical Inference*, Eds: Godambe and Sprott; Holt, Rinehart and Winston, 108-141.

[46] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York, Berlin.

[47] Hardle, W. (1991). *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

[48] Hardle, W. and Marron, J.S. (1990). Bootstrap simulaneous error bars for nonparametric regression. *Ann. Statist.*

[49] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London, New York.

[50] Hinkley, D.V. (1988). Bootstrap methods. *J. Roy. Statist. Soc. Ser.B*, 50,321-337.

[51] Hoeffding, W.(1963). Probability Inequalities for sum of bounded random variables, *JASA*, 58, 13-30.

[52] Hu, F. (1994). The consistency of the maximum relevance weighted likelihood estimator. *Unpublished Manuscript.*

[53] Hu, F. and Zidek, J.V. (1993a). An approach to bootstrapping through estimating equations. *Tech. Report of University of British Columbia,* No.126.

[54] Hu, F. and Zidek, J.V. (1993b). A relevance weighted nonparametric quantile estimator. *Tech. Report of University of British Columbia,* No.134.

[55] Hu, F. and Zidek, J.V. (1993c). Relevance weighted likelihood estimations. *Unpublished Manuscript.*

[56] Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, Berkeley) 221-233.

[57] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[58] Jennen-Steinmetz, C. and Gasser, T. (1988). A unifying approach to nonparametric regression estimation. *JASA*, 83, 1084-1089.

[59] Kullback, S. (1954). Certain inequalities in information theory and Cramer-Rao inequality. *Ann. Math. Statist.*, 25, 745-751.

[60] Kullback, S. (1959). *Information Theory and Statistics*, New York: Wiley.

[61] Kullback, S. and Leibler, R.A. (1951). On information and sufficient. *Ann. Math. Statist.*, 22, 79-86.

[62] Lahiri, S.N. (1992). Bootstrapping M - estimators of a multiple linear regression parameter. *Ann. Statist.*, 20, 1548-1570.

[63] Lee, K.W. (1990) Bootstrapping logistic regression model with regressors. *Commun. Statist. Theory Meth.*, 19, 2527-2539.

[64] Lehmann, E.L. (1983). *Theory of Point Estimation.* New York: Wiley.

[65] Lehmann, E.L. (1986). *Testing Statistical Hypothesis.* New York: Wiley.

[66] Lindley, D.V. (1956). On a measure of the information provided by an experiment . *Ann. Math. Statist.*, 27, 980-1005.

[67] Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2. Inference.* Cambridge: Cambridge University Press.

[68] Liu, R.Y. (1988). Bootstrap procedures under some non-iid models. *Ann. Statist.*, 16, 1696-1708.

[69] Liu, R.Y. and Singh, K. (1992). Efficiency and robustness in resampling. *Ann. Statist.*, 20, 370-384.

[70] Mack, Y.P. and Muller, H.G. (1989). Convolution-type estimators for nonparametric regression estimation. *Statistics and Probability Letters*, 7, 229-239.

[71] Marcus, M.B. and Zinn, J.(1984). The bounded law of the iterated logarithm for the weighted empirical distribution process in the non-iid case, *Ann.Prob.*, 12, 335-360.

[72] Marshall, A.W. and Olkin, I.(1979). *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

[73] McCullagh, P.J. (1983). Quasi-likelihood functions. *Ann. Statist.*, 11, 59-67.

[74] McCullagh, P.J. and Nelder, J.A. (1989). *Generalized linear models*, 2nd edition, Chapman and Hall, London.

[75] Moulton, L.H. and Zeger, S.L. (1991). Bootstrapping generalized linear models. *Computational Statistics and Data Analysis*, 11, 53-63.

[76] Muller, H.G. (1988). *Nonparametric Analysis of Longitudinal Data*, Berlin: Springer-Verlag.

[77] Muller, H.G. and Stadtmuller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, 15, 610-635.

[78] Navidi, W. (1989). Edgeworth expansions for bootstrapping regression models. *Ann. Statist.*, 17, 1472-1478.

[79] Owen, A.B. (1991). Empirical likelihood for linear models. *Ann. Statist.*, 19, 1725-1747.

[80] Park, B.U. and Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *JASA*, 85, 66-72.

[81] Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *Int. Statist. Rev.*, 54, 221-226.

[82] Rubin, D.B. (1981). The Bayesian bootstrap. *Ann. Statist.*, 9, 130-134.

[83] Sasieni, P. (1992). Maximum weighted partial likelihood estimators of Cox model. *JASA,* 88, 144-152.

[84] Sarndal, C.-E., Swenson, B. and Wretman, J.(1992). *Model Assessed Survey Sampling*, New York: Springer-Valag.

[85] Savage (1954). *The Foundations of Statistics.* New York: Wiley.

[86] Serfling, R.J.(1980). *Approximation Theorems of Mathematical Statistics* , Wiley, New York.

[87] Serfling, R.J.(1984). Generalized L-, M- and R-statistics, *Ann.Statist.*, 12, 76-86.

[88] Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27, 379-423; 623-656.

[89] Shorack, G.R.(1978). Convergence of reduced empirical and quantile processes with applications of order statistics in the non-iid case. *Ann.Statist.*, 1, 146-152.

[90] Shorack, G.R. and Wellner J.A.(1986). *Empirical Processes with Applications to Statistics*, Wiley, New York.

[91] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.*, 9, 1187-1195.

[92] Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models, *JASA*, 84, 276-283.

[93] Stout, W.F. (1968). Some results on complete and almost sure convergence of linear combinations of independent random variables and martingale differences. *Ann. Math. Statist.,* 39, 1549-1562.

[94] Stout, W.F. (1974). *Almost Sure Convergence.* Academic Press, New York.

[95] Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *JASA,* 82, 559-567.

[96] Wahba, G. and Wold, S. (1975). A completely automatic French curve: fitting spline functions by cross-validation. *Communications in Statistics, Part A–Theory and Methods*, 4, 1-7.

[97] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, 20,595-601.

[98] Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61, 439-447.

[99] Weerahandi, S. and Zidek, J.V. (1988). Bayesian nonparametric smoothers for regular processes. *The Canadian Journal of Statistics*, 16, 61-74.

[100] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrika*, 50, 1-25.

[101] Wiener, N. (1948). *Cybernetics*. New York: Wiley.

[102] Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.*, 14, 1261-1295.