8702351

Rutherford, Brian Milne

BOOTSTRAP AND RELATED METHODS FOR APPROXIMATE CONFIDENCE BOUNDS IN NONPARAMETRIC REGRESSION

The University of Arizona PH.D. 1986

BOOTSTRAP AND RELATED METHODS FOR APPROXIMATE CONFIDENCE BOUNDS IN NONPARAMETRIC REGRESSION

by

Brian Milne Rutherford

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF SYSTEMS AND INDUSTRIAL ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN SYSTEMS ENGINEERING

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 8 6

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Brian Milne Rutherford

entitled Bootstrap and Related Methods for Approximate Confidence Bounds in Nonparametric Regression

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Moshe Shaked — Date 5-19-86

Arthur Larry Wright — Date 5/19/86

Marcel F. Neuts — Date 5/19/86

Joseph Pignatiello Jr — Date 5/19/86

Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Sid Yakowitz — 5/19/86
Dissertation Director — Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Brian M Rutherford

ACKNOWLEDGMENTS

It would be terribly selfish to submit this dissertation without admitting that it required an effort by many.

My special thanks go to:

- My advisor, Dr. Sidney Yakowitz for his interesting direction and crucial mathematical insights;

- My wife, Kathy Rutherford for her encouragement and tolerance of long hours, and some bad moods;

- My typist, Kathy Rutherford for her excellent typing, some long hours and some bad moods;

- My committee for their many helpful suggestions;

- Dr. Schuster and Dr. Singh for their personal communications;

- Nick Nelson for going out of his way to make computer resources available as needed; and

- all my friends who provided endless encouragement.

TABLE OF CONTENTS

TABLE OF CONTENTS--Continued

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS--Continued

LIST OF TABLES

ABSTRACT

The problem considered relates to estimating an arbitrary regression function m(x) from sample pairs (X_i,Y_i) $1 \leq i \leq n$. A model is assumed of the form $Y = m(x) + \epsilon(x)$ where $\epsilon(x)$ is a random variable with expectation 0. One well known method for estimating m(x) is by using one of a class of kernel regression estimators say $m_n(x)$. Schuster (1972) has shown conditions under which the limiting distribution of the kernel estimator $m_n(x)$ is the normal distribution.

It might also be of interest to use the data to estimate the distribution of $m_n(x)$. One could, given this estimate, construct approximate confidence bounds for the function m(x).

Three estimators are proposed for the density of $m_n(x)$. They share a basis in non-parametric kernel regression and utilize bootstrap techniques to obtain the density estimate. The order of convergence of one of the estimators is examined and conditions are given under which the order is higher then when estimation is by the normal approximation. Finally the performance of each estimator for constructing confidence bounds is compared for moderate sample sizes using computer studies.

1. INTRODUCTION

Problems in nonparametric regression have recently received a great deal of attention largely as a result of the availability and affordability of computing resources. Some of the methods which have emerged are quite simple to compute and have a great deal of intuitive appeal.

The basic problem considered here is a nonparametric problem in regression where we wish to make approximate pointwise confidence statements concerning the regression function values. A general model for this problem can be expressed in two equivalent ways.

We may write Y explicitly as a random function of x: $y = m(x) + \epsilon(x)$ where $m(x)$ is the regression function and $\epsilon(x)$ represents a random term with mean zero. For this model and arbitrarily chosen x, we wish to make confidence statements for the value of $m(x)$. An alternative formulation for this problem is to assume we are working with a random pair (X,Y) with the goal of placing confidence bands on the expectation of Y for specific values of X. In either case we assume that both X and Y represent random variables with unknown distributions.

The estimators used to construct approximate confidence bounds use the methodologies of two fairly new but quite well-known nonparametric techniques. A review of relevant literature in the areas of bootstrapping and kernel regression are the subjects of the next two chapters of this dissertation.

Bootstrapping provides a method of obtaining estimates of the probability density of an arbitrary estimator, say R. Briefly, bootstrapping consists of constructing an empirical distribution (or some reasonable alternative). This is to be used as a 'resampling distribution'. Each time this distribution is randomly resampled a new value, say R*, for R is computed. The values R_1^*, $R_2^*,\ldots,R_B^*$ are calculated when the resampling procedure is repeated B times. The histogram or some other density approximation of these R_i^* provides an approximation to the density of R from which inferences about R can be made. This density approximation for R* is called the bootstrap density.

Chapter 2 describes the bootstrapping procedure in detail and gives some of the relevant results needed for this dissertation. An introductory section describes the mechanics of the bootstrap. The second section mentions some of the alternative methods which have been proposed for constructing the resampling distribution. The third section describes one way in which the bootstrap can be used to construct confidence limits. This is followed by a section which gives some of the primary theoretical results in the bootstrap literature. The chapter concludes with a section devoted to reviewing the bootstrap method in a regression setting.

Kernel regression estimates from a sample of n pairs (X_i,Y_i), i=1,...,n provide a nonparametric method for estimating the regression function m(x). Here, the estimate $m_n(x)$ is computed as a weighted combination of the realizations of Y_i, i=1,...,n where the weights depend upon the distance $|x-X_i|$. More specifically, it is assumed that

$$m_n(x) = \Sigma_{i=1}^n Y_i k_n(x-X_i)/\Sigma_{j=1}^n k_n(x-X_j)$$

where details of the restrictions on the kernel k_n will be given in Chapter 3. The kernel estimators, used for estimating conditional distributions which are needed in turn to obtain approximations of the density of $m_n(x)$, are ratios of two kernel estimators. The numerator being a two dimensional kernel and the denominator one dimensional. The first section of chapter 3 is an introduction giving more details of what follows in that chapter. Next, a section reviews the essential asymptotic properties of kernel estimators for univariate and multivariate probability densities needed in this dissertation. A third section reports an asymptotic result specifically for conditional kernel density estimation. The chapter concludes with a section reviewing the asymptotics for kernel regression estimators.

In the remainder of this introductory chapter, we describe briefly how this research utilizes these methodologies to construct confidence bounds in what is believed to be a novel way. This description will be given through an outline of the remaining chapters. Chapter 4 gives three different methods for estimating the confidence bounds for the regression function as discussed earlier. In each case, the density of $m_n(x)$, the kernel regression estimator, is approximated by a different technique and the confidence bounds are then taken as percentiles from these densities. The primary difference between the three methods is the way the conditional densities of Y given X are constructed. Each section of Chapter 4 gives the details of one of the estimators and includes a detailed algorithm.

Chapter 5 contains the major theoretical contribution of this dissertation. In that chapter we establish the order of convergence for one of the methods proposed in Chapter 4. Following an introduction, a section examines the order of convergence for an estimator which is implied by the results on asymptotic normality of $m_n(x)$. Next, preliminary theoretical results are given and a section is devoted to describing the notation needed for the final section. The chapter concludes with a section giving the order of convergence for the kernel conditional estimator of Chapter 4.

In Chapter 6 we present a series of examples illustrating several features of these estimators. Following the introduction a brief description of the simulations to be run is given. This is followed by a section which explains how various parameters of the kernels are computed. The remainder of the chapter describes particular computer simulations. The third section compares the two alternative methods of constructing density approximations of $m_n(x)$ from the approximate conditional densities. Computing time is considered in this analysis as well as accuracy. A fourth section compares the estimators under a variety of model assumptions and conditional distributions. The final section shows how the estimates of methods 2 and 3 described in Chapter 4 may fail when our assumptions about tne data are not satisfied.

2. THE BOOTSTRAP

Introduction

In this chapter we describe the bootstrap as a general technique and present a review of related research. The bootstrap method was introduced by Efron (1979) as a tool for estimating the variability and sampling distribution for an arbitrary estimator R. Consider, for example, the statistic

$$\bar{X} = \Sigma_{i=1}^{n} X_i/n$$

as an estimator of the mean μ from a single sample $X_1,\ldots,X_n$ with distribution F. For this particular estimator, we may use the data to obtain an estimate of the standard error of $\bar{X}$ namely

$$\sigma_{\bar{X}} = [(1/(n-1))\Sigma_{i=1}^{n}(X_i-\bar{X})^2]^{1/2}.$$

As Efron (1982) points out the problem with this is that such an estimate of standard error has no obvious generalization to an arbitrary estimator R.

Assume we are given a random sample of size n from an unknown distribution F. We assume $X_1, X_2, \ldots, X_n$ are independent. The general problem is, given a specific random variable $R(\mathbf{X},F)$ (where $X = (X_1,\ldots,X_n)$), to estimate the distribution of R from information provided by the data.

The basic single sample bootstrap method follows:

(i) Construct the empirical distribution

$$F_n(x) = \Sigma_{i=1}^{n} 1/n \; I_{\{X_i<x\}}$$

where $I_{\{A\}}$ is the indicator of the set A.

(ii) With F_n fixed, select a new sample $X_1^* = x_1^*,\ldots,X_n^* = x_n^*$ at random from F_n (This will be referred to as a bootstrap sample). From this sample, compute $R(\mathbf{X}^*,F_n)$.

(iii) Approximate the distribution of $R(\mathbf{X},F)$ by the bootstrap distribution of $R(\mathbf{X}^*,F_n)$.

Even in situations where R is a complicated function of the data, (i) and (ii) are easily accomplished. (iii), however, may and usually does require approximation. Three methods for calculating the bootstrap distribution are suggested by Efron (1979). Examples of each are given in that paper.

Method 1 Direct Analytic Calculation.

Method 2 Monte Carlo Approximation: Repeat (ii) above many (say B) times. Then approximate the bootstrap distribution by the empirical distribution of $R(X_1^*,F_n),\ldots,R(X_B^*,F_n)$.

Method 3 Taylor series expansion methods: These are used to obtain the approximate mean and variance. This latter method has used primarily for theoretical comparison with other methods in computing standard error such as the jackknife.

Research on the bootstrap has taken several directions and the remainder of this chapter is accordingly organized. The successive sections treat: a) the use of alternatives to the empirical distribution as a resampling distribution; b) a method for using the bootstrap distribution to obtain confidence bounds; c) demonstration that the bootstrap is consistent for specific situations; and d) showing that in specific cases it provides an advantage compared to asymptotic approximations. These last two are of greatest interest here. A final section in this chapter discusses the bootstrap literature dealing specifically with regression problems.

Alternative Resampling Distributions

Efron (1979), (1981a), (1981b), (1982) considers alternatives to using the empirical distribution as a resampling distribution for generating bootstrap samples or otherwise computing the bootstrap distribution. These are summarized below.

1. The symmetric bootstrap:

Here, we replace F_n by F_{sym} which is formed by reflecting F_n about its median $x_{(m)}$ That is, the empirical density say f_{sym} becomes

$$f_{(x)sym} = 1/(2n-1) \text{ at } x_{(1)},\ldots x_{(n)},(2x_{(m)}-x_{(1)}),\ldots,(2x_{(m)}-x_{(n)})$$

where $x_{(i)}$ is the ith ordered data value from the original sample.

One might use this modified resampling distribution if there were reason to assume that F has a symmetric density f.

2. The window estimates:

Here the empirical distribution is smoothed by convolving it with some continuous distribution. In essence, F_n is replaced with a smoothed distribution with the same mean and variance. In examples, Efron has tried smoothing using the normal distribution and also the uniform distribution.

3. The parametric bootstrap:

The data are used to compute parameter estimates for some prescribed family of distributions. The resulting fully specified distribution now replaces the empirical distribution in the basic bootstrap procedure.

Many other proposals could be made for smoothing the empirical distribution including computing the kernel density estimate and resampling from that. Efron has published simulation studies which include some of these alternatives. The results showed minor differences when compared to the basic bootstrap.

Constructing Confidence Limits

Next, Efron (1981b), (1982) has proposed several methods for obtaining confidence intervals once the bootstrap distribution has been obtained. For notational purposes, assume from here on that we are dealing with statistic $R(X,F) = \emptyset(F)$ which is invariant under permutations of the arguments. Hence we can replace the notation $\emptyset^* = \emptyset^*(F_n)$ for $R(X^*,F_n)$. Of those methods proposed in these papers, we mention only the percentile method which is used in this dissertation for the experimental results of Chapter 6. Many of the remaining methods were specifically designed to obtain bounds for the sample mean in a single sample situation.

Let CDF(t) represent the cumulative of the bootstrap distribution of $\emptyset^*$. That is, if we assume that we repeated the resampling portion of the bootstrap algorithm B times, we are interested in

$$CDF(t) = Prob^* \{\emptyset^* < t\} = B^{-1} \Sigma_{i=1}^{B} I_{\{\emptyset^* < t\}}$$

where $Prob^*$ indicates probability under the bootstrap resampling procedure and $I_{\{A\}}$ is the indicator of the set A. The percentile method assigns the $(1-2\alpha)$ confidence interval to be:

$$[\emptyset^*(\alpha), \emptyset^*(1-\alpha)] \text{ where } \emptyset^*(\alpha) = CDF^{-1}(\alpha)$$

hence, we are simply finding the appropriate percentiles of the bootstrap distribution.

Asymptotic Results for the Bootstrap

We have reviewed the mechanics of the bootstrap and one suggested method for using it to construct confidence limits. It is the purpose of this section to report results in the bootstrap literature which show conditions under which the use of the bootstrap is valid in the sense that the asymptotic distribution of the statistics $R(\mathbf{X}^*,F_n)$ and $R(\mathbf{X},F)$ are the same. The section will conclude with a theorem which shows one case where the bootstrap can provide an advantage when compared to the asymptotic distribution of an estimator to approximate its distribution for finite sample sizes.

The first results in demonstrating the validity of the bootstrap was again in Efron (1979). Efron considers the asymptotic justification for the bootstrap when the sample space is finite. Assume it is 1,...,L without loss of generality. When this is the case, we can write the statistics $R(\mathbf{X}^*,F_n)$ and $R(\mathbf{X},F)$ as:

$$R(\mathbf{X},F) = Q(f',f); \qquad \text{and}$$
$$R(\mathbf{X}^*,F_n) = Q(f^*,f')$$

where:

Q is some function;

$$\mathbf{f} = (f_1,\ldots,f_L) \quad , \; f_i = \text{prob}\{x = i\};$$
$$\mathbf{f}' = (f_1',\ldots,f_L') \quad , \; f_i' = \#(x=i)/n;$$
$$\mathbf{f}^* = (f_1^*,\ldots,f_L^*) \quad , \; f_i^* = \#(x^*=i)/n;$$

and #(x=i) is the number of times x=i.

Efron gives conditions under which

$\sqrt{n}(f'-f)/f \to$ Normal $(0,\Sigma)$ and

$\sqrt{n}(f^*-f')/f' \to$ Normal $(0,\Sigma)$

in distribution as $n \to \infty$. Here, Σ is the L by L covariance matrix which is the same in either case. Hence, he has demonstrated for this finite sample space case that the bootstrap is valid.

Bickel and Freedman (1981) (and also Singh (1981)) extend this concept to other statistics. The following is Bickel and Freedman's theorem for means where the sample space is assumed continuous. Note that here, the original sample size n and the bootstrap sample size m may be different. Note also that for this problem,

$$R(X,F) = \Sigma_{i=1}^{n} X_i/n = \mu_n$$

$$R(X^*,F_n) = \Sigma_{i=1}^{m} X_i^*/m = \mu_m^*.$$

Theorem 2.1 (Bickel and Freedman) Suppose $X_1, X_2 \ldots$ are independent identically distributed and have finite positive variance σ^2. Along almost all sample sequences $X_1^*, X_2^*, \ldots, X_m^*$ given $(X_1, \ldots, X_n)$ as n and m tend to ∞

(a) The conditional distribution of $\sqrt{m}(\mu_m^* - \mu_n)$ converges weakly to normal $(0,\sigma^2)$

(b) $s_m^* \dashrightarrow \sigma$ in conditional probability

i.e. $P\{|s_m^* - \sigma| > \epsilon \backslash X_1, \ldots X_n\} \dashrightarrow 0$ a.s.

as $n \dashrightarrow \infty$ and $m \dashrightarrow \infty$ where s_m^* is the sample standard deviation from the bootstrap samples. ■

Since it is known that the distribution of $\sqrt{n}(x-\mu)/s_n$ is asmptotically normal (0,1) Bickel and Freedman as in Efron's proof are just showing that the distributions are asymptotically the same. So again, validity is established by demonstrating how the limiting distribution of the bootstrap as m and n approach infinity is the same as the statistic R(X,F) whose distribution is our interest.

Bickel and Freedman (1981) went on to demonstrate the validity of the bootstrap for large classes of statistics where tne asymptotics are already known. Many pivotal statistics are functions of quantiles or U-statistics or of the values of the empirical distribution itself. Bickel and Freedman investigate these classes of statistics.

As was mentioned earlier in the introduction to this section, demonstrating that the bootstrap is valid for a class of statistics does not in itself suggest that the bootstrap is useful even for that specific class. Since we know for example that the difference between the sample mean and the population mean has a limiting normal distribution $N(0,\sigma^2/n)$, we can use the data in a large sample to estimate σ^2 and then invert the normal distribution in the usual way to obtain confidence bounds. We can do this with other statistics when the asymptotics are known.

We turn now to examine how the bootstrap would compare to this asymptotic approximation for finite sample size. One major effort in this direction is Singh (1981). Like Bickel and Freedman, Singh proves that the bootstrap is valid for means (and also for quantiles) but he takes an approach which allows for a comparison of the order of convergence.

Theorem 2.2 (Singh) If $E\{|X|^3\} < \infty$ and F is non-lattice, then

$$\text{Prob}^*\{n^{1/2}(\mu_n^*-\mu_n)/s_n \le x\} = \Phi(x)+\{\mu_{n3}(1-x^2)/(6s_n^3 n^{1/2})\}\phi(x)+o(n^{-1/2})$$

uniformly in x a.s. where $\Phi(x)$ and $\phi(x)$ are the standard normal distribution and density respectively, s_n^2 and μ_{n3} are the sample second and third moments and Prob^* indicates probability under F_n. Therefore, in this case

$$n^{1/2}||\text{Prob}\{n^{1/2}(\mu_n-\mu)/\sigma \le x\}-\text{Prob}^*\{n^{1/2}(\mu_n^*-\mu_n)/s_n \le x\}||_\infty \to 0 \text{ a.s.} \blacksquare$$

Note that this is important because the normal approximation utilizes only the first two moments. In Singh's words, "In the non-lattice case, it is shown that the bootstrap approximation of the distribution of the standardized sample mean is asymptotically more accurate than approximation by the limiting normal distribution". "The difference in accuracies of the two approximations decreases with decreasing skewness of the underlying distribution and is non-existent for symmetric distributions".

In Chapter 5 we will use a similar approach when examining the order of convergence of one of our proposed regression estimators.

IV. Bootstrap for Regression

The ideas presented for demonstrating validity for the bootstrap carry over to regression problems. Freedman (1981) establishes the validity of the bootstrap for linear regression models. First consider what Freedman calls the regression model

$$\mathbf{Y} = X\boldsymbol{\beta} + \epsilon$$

where ϵ represents a random vector of independent errors which are not effected by the levels of X. The unknown values $\boldsymbol{\beta}$ are the parameters of interest. If X is of full rank, we know the statistic $\bar{\beta} = (X'X)^{-1}X'\mathbf{Y}$ has expectation $\boldsymbol{\beta}$ and variance covariance matrix $\sigma^2(X'X)$ where σ^2 is the variance of Y_i, i=1,..,n.

The asymptotic result for this model is that if $X'X/n \to \Sigma$, positive definite as $n \to \infty$ then $n^{1/2}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically normal with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2(X'X)^{-1}$.

The mechanics of the bootstrap for this case are to compute the least squares linear model and to form an empirical distribution from the residual values e_i, i=1,...,n. Next compute a bootstrap sample where each row gives $Y_i^* = X_{(i)}\bar{\boldsymbol{\beta}} + e_i^*$ so that $X_{(i)}$ is as in the initial sample, $\bar{\beta}$ is the initial parameter estimates and e_i^* is sampled at random from the empirical distribution. The n values Y_i^*,i=1,...,n make up a single bootstrap sample yielding one estimate of the parameters β^*.

Freedman's proofs utilize many of the same tools which were used in Bickel and Freedman (1981). Freedman's theorem 2.2 follows:

Theorem (Freedman) Assume the regression model (described above) with the components $\epsilon_1,\ldots,\epsilon_n$ having distribution F, mean 0 and variance σ^2 unknown. Assume that $1/n(X'X) \rightarrow \Sigma$ where Σ is positive definite. Then along almost all sample sequences, given $Y_1,\ldots Y_n$, as m (the bootstrap sample size) and n approach ∞

a) the conditional distribution of $\sqrt{m}\{\beta_m^* - \bar{\beta}\}$ converges weakly to the normal distribution with mean 0 and variance-covariance matrix $\sigma^2(X'X)^{-1}$.

b) The conditional distribution of σ_m^* converges to a point mass at σ.

c) The conditional distribution of the statistic $\{X'X\}^{1/2}\{\beta_m^* - \bar{\beta}\}/\sigma_m^*$ converges to the standard normal in R^p. ■

Note that as in the single sample description of the bootstrap, m is allowed to differ from the original sample size n and hence the subscripts for β_m^* and σ_m^*. These quantities are just the usual estimates made from the bootstrap sample.

A second model considered in Freedman (1981) is one which he refers to as the correlation model. This model is the same as the regression model described above except that we allow the residual variance (and distribution) to depend upon the independent variables X.

For this model, the bootstrap procedure differs in that when resampling, we select vectors $(X_{1i},\ldots,X_{pi},Y_i)$ at random m times to obtain a single bootstrap sample of size m. In effect, we are reconstructing our X matrix which we did not do in the regression model. Note that here, m is definitely different from n and we would expect that it is much larger. From each bootstrap sample, we make the usual estimate of the unknown parameter $\beta_m = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$. Freedman (1981) gives two theorems showing that this bootstrap procedure is valid in the usual sense for this model.

In Chapter 4 three methods of bootstrapping a more general regression model are proposed. One of these, the kernel conditional method is valid in this correlation setting. The need to resample from only the original sample pairs is circumvented by resampling from conditional densities which were constructed from local data values. The other two proposed estimators are valid only for the regression model presented first in this section.

3. KERNEL REGRESSION

Introduction

The purpose of this chapter is to describe nonparametric kernel regression estimators and to present a partial literature review of research related to kernel regression. As a first step we introduce the kernel method of univariate and multivariate probability density estimation. We then describe conditional density estimation and conclude the chapter with kernel regression function estimation.

Probability Density Estimation

The nonparametric kernel regression estimators described later in this chapter have their roots in probability density estimation. The main goal here will be to give a summary of results in this area relevant to our needs. While some results, included for completeness, will not be used directly in the remainder of this dissertation, the subject is too vast to permit a comprehensive review.

The format for this section will be as follows. First there is an introduction of the general form of the estimator. Next, several pointwise results are reported. Theorems given for asymptotic pointwise properties concern the following:

a) weak pointwise convergence;

b) asymptotic distribution;

c) strong pointwise convergence;

d) order of strong convergence.

We next report several results giving uniform asymptotic properties. Here we concentrate on the following:

e) weak uniform convergence;

f) strong uniform convergence;

g) order of strong uniform convergence.

One of the first papers to consider a kernel estimator of the type reported here is Parzen (1962). Let X_i, i=1,...,n, be a sequence of independent random variables with density f(x) and distribution function F(x). Parzen takes a probability density estimator, suggested earlier by Rosenblatt (1956), of the form

$$f_n(x) = \{F_n(x+a_n)-F_n(x-a_n)\}/2a_n$$

where F_n is the empirical distribution function and a_n is a sequence of positive numbers converging to 0 as $n \to \infty$. He shows that this is a special case of the class of estimators formed by choosing

$$k(y) = \begin{cases} 1/2 & \text{if } |y| \leq 1 \\ 0 & \text{if } |y| > 1 \end{cases}$$

in the more general equation

$$f_n(x) = (1/a_n)\int k((x-y)/a_n)dF_n(y)$$

$$= (na_n)^{-1} \Sigma_{j=1}^{n} k((x-X_j)/a_n). \tag{3.1}$$

Parzen proceeds to analyze the above estimators specifically addressing a), b), and e).

The following theorems from Parzen (1962) are in answer to a) and b) above. Each gives conditions on the kernel k and the "bandwidth" sequence a_n to yield the desired properties.

<u>Theorem 3.1</u> (Parzen) Suppose k(y) is a Borel function satisfying:

$$\sup_y|k(y)| < \infty \tag{3.2}$$

$$\int|k(y)|dy < \infty \tag{3.3}$$

$$\lim_{y\to\infty}|yk(y)| = 0 \tag{3.4}$$

and let g(y) satisfy $\int|g(y)|dy < \infty$. Let a_n be a sequence of positive constants satisfying

$$\lim_{n\to\infty}a_n = 0 \tag{3.5}$$

and define

$$g_n(x) = a_n^{-1} \int k(y/a_n) g(x-y) dy.$$

Then at every point x of continuity of g

$$\lim_{n\to\infty} g_n(x) = g(x) \int k(y) dy.$$ ■

The next corollary which follows immediately shows the estimators are asymptotically unbiased.

Corollary 3.1 (Parzen) Under the conditions of Theorem 3.1 and under the additional condition that

$$\int k(u) du = 1 \tag{3.6}$$

we have that the estimates (3.1) are asymptotically unbiased. ■

Def 3.1 (Parzen) an even kernel function k satisfying all assumptions (3.2), (3.3), (3.4), (3.6) is called a **weighting function.** ■

This next theorem gives the limiting variance of $f_n(x)$.

Theorem 3.2 (Parzen) Under the conditions of Theorem 3.1 the estimator (3.1) has variance satisfying

$$\lim_{n\to\infty} (na_n) \mathrm{Var}[f_n(x)] = f(x) \int k^2(u) du$$

at all points x of continuity of f(x). ■

The remainder of question a) is answered by the following corollary to Theorem 3.2.

Corollary 3.2 (Parzen) Under the conditions of Theorem 3.1 and the additional condition that

$$\lim_{n\to\infty} na_n = \infty \qquad (3.7)$$

$f_n(x)$ is a consistent estimate of $f(x)$. ■

This shows consistency in mean square error which in turn implies weak pointwise convergence. Parzen (1962) shows that b) the question about the asymptotic distribution can be at least partially answered in the following way.

Theorem 3.3 (Parzen) Let k be a weighting function and let the sequence a_n satisfy (3.5) and (3.7). Then the sequence of estimates $f_n(x)$ are asymptotically normal. That is

$$\lim_{n\to\infty} P[(f_n(x)-E(f_n(x)))/Var[f_n(x)]^{1/2} < c] = \Phi(c).$$

where Φ is the standard normal cumulative distribution function. ■

One may wish to enhance this result by replacing $E[f_n(x)]$ by $f(x)$. This will be done later when we report extensions to multivariate densities.

In later chapters of this dissertation it will be of interest to estimate probability densities of two variables with the estimator

$f_n(x,y)$. Parzen's results have been extended to the multivariate case by Cacoullos (1965). In that paper the X_j, $j=1,\dots,n$ are independent random vectors with p components and the joint density $\mathbf{f}(\mathbf{x})$ is estimated by

$$f_n(\mathbf{x}) = (na_n^p)^{-1}\Sigma_{j=1}^n K((\mathbf{x}-\mathbf{X}_j)/a_n)$$

where the positive sequence $a_n \to 0$ as $n \to \infty$ and $K(\mathbf{x})$ is a kernel function chosen to satisfy suitable conditions. The following theorem is the multivariate analogue of Theorem 3.1.

<u>Theorem 3.4</u> (Cacoullos) Suppose $K(\mathbf{y})$ is a scalar Borel function on $\mathbf{R}^p$ such that

$$\sup_{\mathbf{y}\in\mathbf{R}^p} |K(\mathbf{y})| < \infty \; ; \tag{3.8}$$

$$\int_{\mathbf{R}^p} |K(\mathbf{y})|d\mathbf{y} < \infty; \qquad \text{and} \tag{3.9}$$

$$\lim_{||\mathbf{y}|| \to \infty} ||\mathbf{y}||^p K(\mathbf{y}) = 0 \tag{3.10}$$

where $||\mathbf{y}|| = \Sigma_{i=1}^p y_i^2$. Let g be a scalar function on $\mathbf{R}^p$ such that

$\int_{\mathbf{R}^p} |g(\mathbf{y})|d\mathbf{y} < \infty$ and define

$$g_n(\mathbf{x}) = \int_{\mathbf{R}^p} a_n^{-p} K(\mathbf{y}/a_n)g(\mathbf{x}-\mathbf{y})d\mathbf{y} \tag{3.11}$$

where $a_n \to 0$ as $n \to \infty$. Then at every point $\mathbf{x}$ of continuity of g we have

$$\lim_{n\to\infty} g_n(\mathbf{x}) = g(\mathbf{x})\int_{R^p} K(\mathbf{y})d\mathbf{y}. \qquad \blacksquare$$

The next corollary corresponds to Corollary 3.1.

Corollary 3.4 (Cacoullos) Under the conditions of theorem (3.4) and with the additional assumption that

$$\int_{R^p} K(\mathbf{y})d\mathbf{y} = 1 \tag{3.12}$$

then for every continuity point $\mathbf{x}$ of f,

$$\lim_{n\to\infty} E[f_n(\mathbf{x})] = f(\mathbf{x}). \qquad \blacksquare$$

For the remaining theorems on multivariate density function estimators we will assume (3.12) holds and make the additional assumptions that there exists a bound m where $K(\mathbf{y}) \le m < \infty$ and that $K(\mathbf{y}) \ge 0$ for all $\mathbf{y}$ and also that $K(\mathbf{y}) = K(-y)$ for all y. We are now working with a kernel function K which we shall call a **multivariate weighting function** similar to definition 3.1. We next give the results corresponding to Corollary 3.2.

Theorem 3.5 (Cacoullos) If the sequence of constants a_n satisfies $\lim_{n\to\infty} a_n = 0$ and $\lim_{n\to\infty} na_n = \infty$ and also K is a multivariate weighting function as defined above, then

$$\lim_{n\to\infty} E[f_n(\mathbf{x})-f(\mathbf{x})]^2 = \mathbf{0}$$

at every continuity point $\mathbf{x}$ of f where $\mathbf{0}$ is the zero vector. ∎

Again, we have established weak pointwise convergence. We turn now to the question of asymptotic distribution.

The next theorem is the multivariate version of Theorem 3.3 showing asymptotic normality. It is followed by a corollary which gives conditions under which one might replace $E[f_n(\mathbf{x})]$ by its limit $f(\mathbf{x})$. This corollary will also apply to the earlier theorem (Theorem 3.3) when $p = 1$.

<u>Theorem 3.6</u>. (Cacoullos) Let $\mathbf{t}_1,\mathbf{t}_2,\ldots,\mathbf{t}_k$ be any finite set of continuity points of the density f. If K is a multivariate weighting function and the sequence a_n satisfies $\lim_{n\to\infty} a_n = 0$ and $\lim_{n\to\infty} na_n^p = 0$ then the joint distribution of random variables $f_n(\mathbf{t}_1),\ldots,f_n(\mathbf{t}_k)$ is asymptotically a k-variate normal with independent components in the following sense:
For any real numbers $c_1,c_2,\ldots c_k$

$$\lim_{n\to\infty} P[(na_n^p)^{1/2}\,(f_n(\mathbf{t}_i)-E[f_n(\mathbf{t}_i)])<c_i, i=1,\ldots k] \qquad (3.13)$$

$$= \prod_{i=1}^{k} \Phi(c_i/\sigma_i)$$

where Φ again denotes the standard normal distribution function, and

$$\sigma_i^2 = f(\mathbf{t}_i) \int_{R^p} K^2(\mathbf{y})d\mathbf{y} \ , \ i=1,\ldots,k$$ ■

<u>Corollary 3.7</u> (Cacoullos) Assume the conditions of the above theorem are satisfied and that in addition for each i, $f(\mathbf{t}_i)$ has continuous partial derivatives of 3rd order in a neigborhood of $\mathbf{t}_i$ and $a_n = O(n^{-\alpha})$ where $(p+4)^{-1} < \alpha < p^{-1}$. Then (3.13) holds with $E[f_n(t_i)]$ replaced by its limit $f(\mathbf{t}_i)$. ■

Many other results specifying the asymptotic behavior of these kernel density estimates for various modes of convergence have been given. Another theorem regarding pointwise behavior is this following theorem giving conditions for pointwise convergence with probability one. This result of Van Ryzin (1969) is given for multivariate probability densities. The conditions of the theorem require the following definition.

<u>Def 3.2</u> (Van Ryzin) A real valued function on the real line g(x) is said to be **locally lipschitz** of order α, $\alpha > 0$, at x if there exists an $\epsilon > 0$ and $0 < m < \infty$ such that $|g(c)-g(x)| \le m|c-x|^{\alpha}$ for all c in the interval $(x-\epsilon, x+\epsilon)$. ■

Theorem 3.7. (Van Ryzin) Let K be a multivariate weighting function on $\mathbf{R}^p$ and let a_n satisfy $\lim_{n\to\infty} a_n \to 0$ and $\lim_{n\to\infty} n a_n^p \to \infty$. Let K satisfy in addition:

$g(c) = \sup_{||\mathbf{u}||\geq a} ||\mathbf{u}||^p \{K(c\mathbf{u})-K(\mathbf{u})\}$ is locally lipschitz of order α at $c = 1$ for some $a > 0$; and

$g(c) = \int \{K(c\mathbf{u})-K(\mathbf{u})\}du$ is locally lipschitz of order α at $c = 1$.

Let the sequence a_n be of the form $kn^{-\beta}$ where k is a constant and $0<\beta<\min\{\alpha/2,1\}/p$. Then if $\mathbf{x}$ is a continuity point of the density $f(\mathbf{x})$ we have with probability one $f_n(\mathbf{x}) \to f(\mathbf{x})$ as $n \to \infty$. ∎

This theorem as specified in Van Ryzin (1969) was given for a more general sequence a_n. We conclude our review of pointwise asymptotic properties with this theorem from Hall (1981). This theorem gives the order of pointwise strong convergence of $f_n(x)$ to $E[f_n(x)]$.

Theorem 3.8 (Hall) Let the kernel k be a function of bounded variation on the real line which satisfies $xk(x) \to 0$ as $|x| \to \infty$ and

$\int k^2(z)dz < \infty$. Let X have distribution F(x) and density f(x) where at a fixed point x we know $f(x) \neq 0$. Assume in addition that F satisfies a local lipshitz condition of order one at the point x. Let $a_n = cn^{-b}$, $c > 0$ and $0 < b < 1$.

Then we have

$$\lim_{n\to\infty}\sup \pm [f_n(x)-E(f_n(x))]na_n/(2\log\log(n))^{1/2})$$
$$= [f(x)\int k^2(z)dz]^{1/2} \quad \text{a.s.}$$ ■

We turn now to review of global results for density functions. Many asymptotic results have been proved in this area. We restrict attention to those stated in the introduction to this chapter.

The first two theorems in this area are by Parzen (1962) and Cacoullos (1966). These theorems concern uniform weak convergence for respectively univariate and multivariate densities.

Theorem 3.9 (Parzen) Let k be a weighting function and let the sequence a_n satisfy both $\lim_{n\to\infty} a_n = 0$ and $\lim_{n\to\infty} na_n^2 = \infty$. Assume further that the density function f(x) is uniformily continuous. Then for every $\epsilon > 0$ we have

$$\lim_{n\to\infty} P[\sup_x|f_n(x)-f(x)|>\epsilon] \to 0.$$ ■

Theorem 3.10 (Cacoullos) Assume the conditions of Theorem 3.5 are satisfied and that in addition we have

$$\lim_{n\to\infty} na_n^{2p} = \infty$$

and the Fourier transform

$$\delta(\mathbf{u}) = \int_{R^p} e^{i\mathbf{u}'\mathbf{y}}K(\mathbf{y})d\mathbf{y}$$

of $K(\mathbf{y})$ is absolutely integrable. Then for every $\epsilon > 0$

$$\lim_{n\to\infty} P[\sup_{R^p} |f_n(\mathbf{x})-f(\mathbf{x})|>\epsilon] = 0. \qquad \blacksquare$$

The next theorem reported here from Stute (1982) gives the order of uniform convergence for a univariate density function under certain conditions.

Def 3.3 (Stute) Let J be any interval on the real line. The function g(x) is said to satisfy a **lipschitz condition** of order α on J if

$$\sup_{x,y\in J}|g(x)-g(y)| \le m|x-y|^{\alpha} < \infty. \qquad \blacksquare$$

Theorem 3.11 (Stute) Let k be of bounded variation and suppose $k(x) = 0$ outside some finite interval $[r,s)$. If $F(x)$ satisfies a lipschitz condition of order 1 on $J = (a,b)$ and if a_n satisfies:

$a_n \to 0$ and $na_n \to \infty$ as $n \to \infty$;

$\log(a_n^{-1})/\log\log(n) \to \infty$ as $n \to \infty$; and

$\log(a_n^{-1})/na_n \to 0$ as $n \to \infty$.

Then for each $\epsilon > 0$

$$\lim\sup_{n\to\infty}\left[\frac{na_n}{2\log(a_n^{-1})}\right]^{\frac{1}{2}}\sup^{t\in J_*}|f_n(t)-E[f_n(t)]| = c$$

where c is a constant less then or equal to $(m(s-r))V_s^r$, the m from the lipschitz condition and V_s^r is the total variation of k over $[r,s]$. J_* represents the interval $(a+\epsilon,\ b-\epsilon)$. ■

Conditional Density Estimation

Two of the estimators described in the next chapter use kernel methods to construct the conditional density of Y given X=x. The approach here is as follows. We wish to approximate the density

$$f_x(y) = f(x,y)/h(x) \tag{3.14}$$

where h(x) is the marginal density of X. We utilize the kernel estimation procedures described earlier in this chapter to estimate independently the numerator and the denominator of (3.14).
We get

$$f_{nx}(y) = f_n(x,y)/f_n(x)$$

$$= \frac{(na_n^2)^{-1}\Sigma_{i=1}^n K(((x-X_i)/a_n)((y-Y_i)/a_n))}{(na_n)^{-1}\ \Sigma_{j=1}^n k((x-X_j)/a_n)} . \tag{3.15}$$

Most of the pointwise asymptotic properties proved for the univariate and multivariate kernel density estimators carry over to conditional densities. The ratio of $f_n(x,y)$ converging in probability to $f(x,y)$ and $f_n(x)$ converging in probability to $h(x)$ converges in probability to $f(x,y)/h(x)$ if $h(x) \neq 0$. The same applies for strong convergence.

The following theorem from Yakowitz (1985) gives the asymptotic distribution of the conditional distribution at X=x. The theorem given here for x of a single dimension completes this section.

Theorem 3.12 (Yakowitz) Let k and K be symmetric uniformly continuous weighting functions of dimensions one and two respectively with finite fourth moments. Assume that $na_n^2 \to \infty$ and $na_n^6 \to 0$ as $n \to \infty$.

Assume that f(x,y) and h(x) are bounded and twice continuously differentiable. Then for $f_{nx}(y)$ as specified in (3.15) we have that

$$(na_n^2)^{1/2}(f_{nx}(y) - f_x(y))$$

is asymptotically normally distributed with zero mean and variance equal to $f(x,y)\int K^2(u,v)dudv/(h(x))^2$ when $h(x) \neq 0$. ■

Regression Function Estimation

The extension of kernel estimation to regression was proposed in Watson (1964). The model and estimator to be described here are central to the remainder of this chapter and this dissertation. This description follows Watson (1964).

We assume n independent random pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ each with the joint density f(x,y). We write h(x) for the marginal density of X and f(y\X=x) for the density of Y conditioned on X.

The regression function of Y on X is given by

$$m(x) = E(Y\backslash X{=}x)$$

$$= \int yf(x,y)dy/h(x).$$

Watson does not initially assume a specific form for the kernel but writes

$$f_n(x,y) = n^{-1} \Sigma_{i=1}^{n} K_n(x-X_i,y-Y_i). \tag{3.16}$$

We will latter use such a two dimensional kernel estimator to estimate the density $f(y\backslash X{=}x)$. For that estimator we will use a kernel similar to those described earlier in this chapter, namely

$$f_n(x,y) = (na_n^2)^{-1} \Sigma_{i=1}^{n} K[((x-X_i)/a_n),((y-Y_i)/a_n)].$$

Now, using (3.16) to estimate $f(x,y)$ one may estimate $m(x)$ by

$$m_n(x) = \int yf_n(x,y)dy/\int f_n(x,y)dy$$

$$= \int y\Sigma_{i=1}^{n} K_n(x-X_i,y-Y_i)dy/\int \Sigma_{i=1}^{n} K_n(x-X_i,y-Y_i)dy$$

After adding the further assumption that

$$\int yK_n(x,y)dy = 0,$$

Watson computed

$$\int yf_n(x,y)dy = n^{-1} \int y\ \Sigma_{i=1}^{n} K_n(x-X_i,y-Y_i)dy$$

$$= n^{-1}[\Sigma_{i=1}^{n} \int (y-Y_i)K_n(x-X_i,y-Y_i)dy$$

$$+ \Sigma_{i=1}^{n} Y_i \int K_n(x-X_i,y-Y_i)dy]$$

$$= n^{-1} \Sigma_{i=1}^{n} Y_i \int K_n(x-X_i,y-Y_i)dy \ .$$

Finally, letting

$$k_n(x-X_i) = \int K_n(x-X_i,y-Y_i)dy$$

the resulting estimator becomes

$$m_n(x) = \Sigma_{i=1}^{n} Y_i k_n(x-X_i)/\Sigma_{i=1}^{n} k_n(x-X_i) \ . \qquad (3.17)$$

In the theoretical developments of Chapter 5, this estimator will take the form

$$m_n(x) = \Sigma_{i=1}^{n} Y_i k((x-X_i)/a_n)/\Sigma_{i=1}^{n} k((x-X_i)/a_n). \qquad (3.18)$$

Following an outline similar to that used earlier for the density function estimators, we now proceed to review issues relating to the asymtotics of the regression estimator $m_n(x)$. We will be interested in the following pointwise results.

a) weak pointwise convergence;

b) asymptotic distribution;

c) strong pointwise convergence;

d) order of strong convergence.

We next report several results giving global asymptotic properties.

Here we concentrate on the following:

e) weak uniform convergence;

f) strong uniform convergence;

g) order of strong uniform convergence.

A partial answer to the first of these questions was given huristically by Watson (1964) and more rigorously by Johnston (1979). The theorem taken here from Johnston (1979) is given only for the specific kernel type estimator (3.18) even though Johnston has proved them for the more general form (3.17) satisfying appropriate conditions.

Theorem 3.13 (Johnston) Assume that the estimator $m_n(x)$ is of the form (3.18) where $\{a_n\}$ satisfies (3.5) and (3.7) and k is a weighting function satisfying $\int k^2(u)du < \infty$. Assume that for the random pair (X,Y), $E(Y^2) < \infty$ and that the fixed point x is a continuity point of h(x) the marginal in X and also a continuity point of the functions

$$\int yf(x,y)dy/h(x) \qquad \text{and} \qquad \int y^2f(x,y)dy/h(x)$$

where we assume $h(x) > 0$. Then $m_n(x) \to m(x)$ in probability. ■

Schuster (1972) has given the asymptotic variance of $(na_n)^{1/2}(m_n(x)-m(x))$ to be

$$\text{Var}[(na_n)^{1/2}(m_n(x)-m(x))] = \text{Var}(Y\backslash X=x)\int k^2(u)du/h(x)$$

for $h(x) > 0$. In this same paper, Schuster answered the second question above concerning the asymptotic distribution of $m_n(x)$.

Theorem 3.14 (Schuster) Assume that the kernel function $k(u)$ is bounded, $|uk(u)|$ is bounded and that the conditions $\int uk(u)du = 0$ and $\int u^2k(u)du < \infty$ are satisfied. Assume the sequence a_n satisfies $\lim na_n^3 = \infty$ and $\lim na_n^5 = 0$. Suppose that $x_1, x_2, \ldots, x_k$ are distinct points with marginal density values $h(x_j) > 0$, $j=1,\ldots,k$. Assume that $E|Y^3|$ is finite and that $h'(x)$, $w'(x)$, $v'(x)$, $h''(x)$ and $w''(x)$ exist and are bounded where

$$h(x) = \int f(x,y)dy;$$
$$w(x) = \int yf(x,y)dy;$$
$$v(x) = \int y^2f(x,y)dy.$$

Then

$$(na_n)^{1/2}(m_n(x_1)-m(x_1),\ldots,m_n(x_k)-m(x_k))^t$$

converges in distribution to a multivariate normal with mean vector **0** and diagonal covariance matrix with ith diagonal element equal to

$$\mathrm{Var}(Y\backslash X=x_i)\int k^2(u)du/h(x_i)$$

for $i=1,\ldots,k$. ■

The following theorem establishes conditions for strong convergence of the kernel regression function. It is taken from Noda (1976).

<u>Theorem 3.15</u> (Noda) Assume that at the point x we have that m(x), h(x) and f(x,y), for any fixed y, are continuous. Assume that the kernel k(x) is a weighting function. Assume further that the bandwidth sequence a_n satisfies $a_n = n^{-\alpha}$ for $0 < \alpha < 1$. If $h(x) \neq 0$ then with probability one,

$$\lim_{n \to \infty} m_n(x) = m(x). \qquad \blacksquare$$

This last theorem on pointwise convergence gives the rate of strong convergence of $m_n(x)$ to m(x). The theorem is taken from Hardel (1984).

<u>Theorem 3.16</u> (Hardel) Let $v(x)= \int y^2 f(x,y)dy$ and f(x,y), h(x) and m(x) are also defined in Theorem 3.14. Assume that h(x) and m(x) are twice differentiable and that v(x) is continuous. Assume further that the kernel k is continuous and has compact support (-1,1) and that

$\int_{-1}^{1} uk(u)du = 0$. Finally assume that the sequence a_n satisfies:

$$a_n = n^{-\alpha}, \quad 0 < \alpha < 1;$$

$$na_n^5/\log\log(n) \to 0 \text{ as } n \to \infty;$$

$\Sigma_{n=1}^{\infty} n^{-2} a_n^{-1} < \infty$ and

$\Sigma_{n=3}^{\infty} (a_n/\text{loglog}(n)) E[Y^2 I(|Y|>b_n)] < \infty$;

where I(A) is the indicator of the set A and b_n satisfies

$b_n = o((na_n^{-1}\text{loglog}(n))^{1/2}/\log(n)^2)$ Then

$$\lim\sup_{n\to\infty} \pm[m_n(x)-m(x)](na_n/(2\text{loglog}(n))^{1/2}$$

$$= [\text{Var}(Y\backslash X=x)\int k^2(u)du/h(x)]^{1/2} \quad \text{a.s}$$ ∎

Turning now to uniform convergence of the kernel regression estimator, we consider first, weak uniform convergence. This first theorem taken from Schuster and Yakowitz (1979) gives conditions for weak uniform convergence of the kernel regression estimator and its derivatives.

Theorem 3.17 (Schuster and Yakowitz) Let $w(x) = \int yf(x,y)dy$. Assume that w has continuous derivatives $w^{(j)}$, j=1,...,N+1 where for each j, $w^{(j)}$ is a function of bounded variation. Let h(x) be the marginal density of X. Assume that for some interval [a,b], h(x) is bounded away from zero and that h(x) and its first N+1 derivatives are continuous and bounded. Let m(x) = w(x)/h(x) be the regression function and assume that its Nth derivative is continuous.

Let k the kernel function be a univariate probability density function satisfying $\int |u|k(u)du < \infty$ and $\lim_{|u|\to\infty}|uk(u)| = 0$. Assume that k has n+1 continuous derivatives $k^{(j)}$ each of bounded variation and that δ the characteristic function of k satisfies

$\int |u|^N |\delta(u)| du < \infty$. Finally, assume that $E(Y^2) < \infty$ and $\lim_{n \to \infty} a_n = 0$. Then there is a constant C such that for any postive ϵ and n sufficiently large

$$P[\sup_{a \le x \le b} |m_n^{(N)}(x) - m^{(N)}(x)|] \le C/(n a_n^{2N+2} \epsilon^2). \qquad \blacksquare$$

Strong uniform bounds for the regression function over a finite interval were established by Mack and Silverman (1982).

<u>Theorem 3.18</u>. (Mack and Silverman) Let the kernel function k be uniformily continuous and of bounded variation. Assume that

$\int |k(u)| du < \infty$, $\quad k(x) \to 0$ as $x \to \infty$ $\quad$ and

$\int |x \log|x||^{1/2} |dk(x)| < \infty$.

Assume the $E(|Y|^s) < \infty$ $\quad$ and

$\sup_x \int |y|^s f(x,y) dy < \infty$

for some $s \ge 2$ and that f, w and h as defined in the Theorem 3.15 are continuous in an open interval containing the bounded interval [a,b] over which h(x) is bounded away from zero. Assume finally that the sequence $a_n = n^{-\alpha}$, $0 < \alpha < 1 - s^{-1}$ Then

$$\sup_{[a,b]} |m_n(x) - m(x)| = o(1)$$

with probability 1. $\blacksquare$

A second theorem in Silverman and Mack (1982) gives rates for both weak and strong uniform convergence under slightly stricter conditions.

4. THE CONFIDENCE BOUND PROCEDURES

Introduction

In this chapter, we introduce three methods for constructing confidence bounds for a regression function. We also use an example from chapter 6 to demonstrate how the methods differ.

The two basic models considered here are of the general forms:

a) $Y = m(x) + e$

where e is an arbitrary random variable centered at 0;

b) $(Y \backslash X=x)$ has distribution $F_x(y)$.

Model a) assumes that the difference in the conditional distribution of Y at two distinct X values is only a difference in mean. Model b) is more general. Here we allow not only for differences in variability of Y for different values of X but also for changes in the shape of the conditional distribution itself.

In the description which follows, Method 1, 2, and 3 are applicable to model a). Only Method 1 is applicable to model b) although ways to use Methods 1 and 2 for model b) are suggested at the conclusion of this chapter. We now proceed to an outline of the steps for constructing the estimators.

It was mentioned briefly in Chapter 2 that each of the procedures which are described in detail in this chapter consist of constructing an estimate of the density of $Y\backslash X=x$. We then use these estimated conditional densities to form an estimate of the density of $m_n(x)$. The confidence bounds are taken as the percentiles of this last density estimate.

The procedures discussed here for approximating the density of $m_n(x)$ can be viewed as a bootstrap method but the resampling procedure or Monte-Carlo method described in Chapter 2 provides only one alternative method for their computation. The bootstrap density as defined in Chapter 2 can also be computed by taking the n-fold convolution of the weighted approximate conditional densities at the n realized values x_j, $j=1,\ldots,n$ from the original sample. More will be said on this after the first method has been described.

After each method has been described an example is given showing the actual construction of the conditional densities and the resulting bootstrap density of $m_n(x)$ from which the percentiles are taken.

<u>Method 1</u> (The Conditional Kernel Method)

The conditional kernel method utilizes both a 2 dimensional kernel and a 1 dimensional kernel. The forms of these kernels and the methods for obtaining parameters for them are discussed in later chapters.

The conditional density is computed at each realized sample value x_j, j=1,...,n as

$$g_{nx_j}(y) = \Sigma_{i=1}^{n} k_n'[(y-Y_i),(x_j-X_i)]/\Sigma_{i=1}^{n} k_n(x_j-X_i) \quad (4.1)$$

where k_n' and k_n are the 2 and 1 dimensional kernels respectively and the subscript n is used to indicate that these kernels are different for different values of n. The kernels used in this dissertation for development of the results in the following chapter are of the form

$$k_n(u) = k(u/a_n)$$

where k is a probability density symetric about 0 and a_n a sequence of positive real numbers converging to zero but satisfying $na_n \rightarrow \infty$ as $n \rightarrow \infty$.

We now have n estimated densities g_{nx_j} conditioned on the values of the random variable X_j, j=1,...,n. We compute for any value x the density approximation of $m_n(x)$ by taking the convolution of these estimated densities weighted appropriately

$$g_{nx}^{m}(y) = *_{j=1}^{n} \beta_j(x) g_{nx_j}(y) \quad (4.2)$$

where * represents the convolution operator,

$$\beta_j(x) = \underline{K}_n(x-X_j)/\Sigma_{i=1}^{n} \underline{K}_n(x-X_i) \quad (4.3)$$

and $\underline{K}_n$ is another kernel function used to weight the conditional densities. Details of these kernel functions and their associated bandwidth parameters is delayed until Chapter 5.

This final approximation gives us what we need. It is the percentiles of the density $g_{nx}^{m}(y)$ from which the approximate confidence bounds are constructed. A more algorithmic presentation of the conditional kernel method follows.

Algorithm:

1. Take a sample of random pairs yielding the data (x_j,y_j), $j=1,\ldots,n$.

2. Using this data, estimate the parameters for the kernels $k',\underline{K}$ and k in (4.1) and (4.3).

3. At each point x_j, $j=1,\ldots,n$ compute the kernel approximation to the conditional density of Y by (4.1).

4. For any x compute the kernel approximation to the conditional density of $m_n(x)$ using (4.2).

5. Use as $(1-\alpha)$ percent confidence bounds for $m(x)$ the $(\alpha/2)$ and $(1-\alpha/2)$ percentiles of the density in 4.

Before proceeding to Method 2, we return briefly to our earlier discussion about the bootstrap. Step 4 above is the computation of the bootstrap density. This is one of the instances in the bootstrap literature where this density can be computed without using the Monte-Carlo methods described in Chapter 2. To see that the two are identical, note that if $m_n^*(x)$ is a bootstrap value obtained by resampling from the estimated conditional densities then

$$\text{Prob}(m_n^*(x)<y) = \text{Prob}([\Sigma_{j=1}^n Y_i^* k_n(x-x_j)/\Sigma_{i=1}^n k_n(x-x_i)]<v) \qquad (4.4)$$

where Y_i^* is sampled at random from the kernel conditional density approximation of Y at $X=x_i$. But Y_i^* are the only random variables in (4.4) so this is just the n-fold convolution of weighted kernel density approximations $g_{nx_i}(y), i=1,\ldots,n$. This is exactly what we had in the algorithm above. The weights are the same provided $k_n(.) = \underline{K}_n(.)$

$$\beta_i(x) = \underline{K}_n(x-X_i)/\Sigma_{j=1}^n \underline{K}_n(x-X_j).$$

Hence we could replace 4 and 5 in the above algorithm to be:

4a. Resample the n conditional densities of 3 (in the original algorithm specification) and obtain a new estimate of $m_n(x)$ call it $m_n^*(x)$.

4b. Repeat 4a a large number, say B, times recording the values $m_n^*(x)$.

5. Use as (1-α) percent confidence bounds at any point x the upper and lower value of the middle (1-α)B values of $m_n^*(x)$ obtained in 4b.

Given now that resampling is only one alternative method for computing the bootstrap density, we will compare the two methods as one of the computer studies in Chapter 6. These alternative approaches for obtaining the bootstrap distribution are also available for the two methods which follow. Note that the differences between the three estimators proposed in this chapter is the construction of the approximation to the density of Y at the points x_j, j=1,...,n. From that point on the procedures are identical.

Figure 4.1 shows one approximation using the conditional kernel method. Figure 4.1a shows the data, the model (dotted line) and the kernel regression estimate (dashed line) as well as 80% bounds (solid lines) on the true model as computed by the conditional kernel method. Figures 4.1b and 4.1c show the conditional kernel density approximations for Y at the points x = 4 and x = 7 respectively. These (and the 38 other) conditional densities are resampled as described above 150 times to get the bootstrap density approximations shown in Figures 4.1d and 4.1e. These latter densities are the ones used to get the percentiles to form the bounds in 4.1a. Here, 4.1d is taken at x = 4, 4.1e at x = 7.

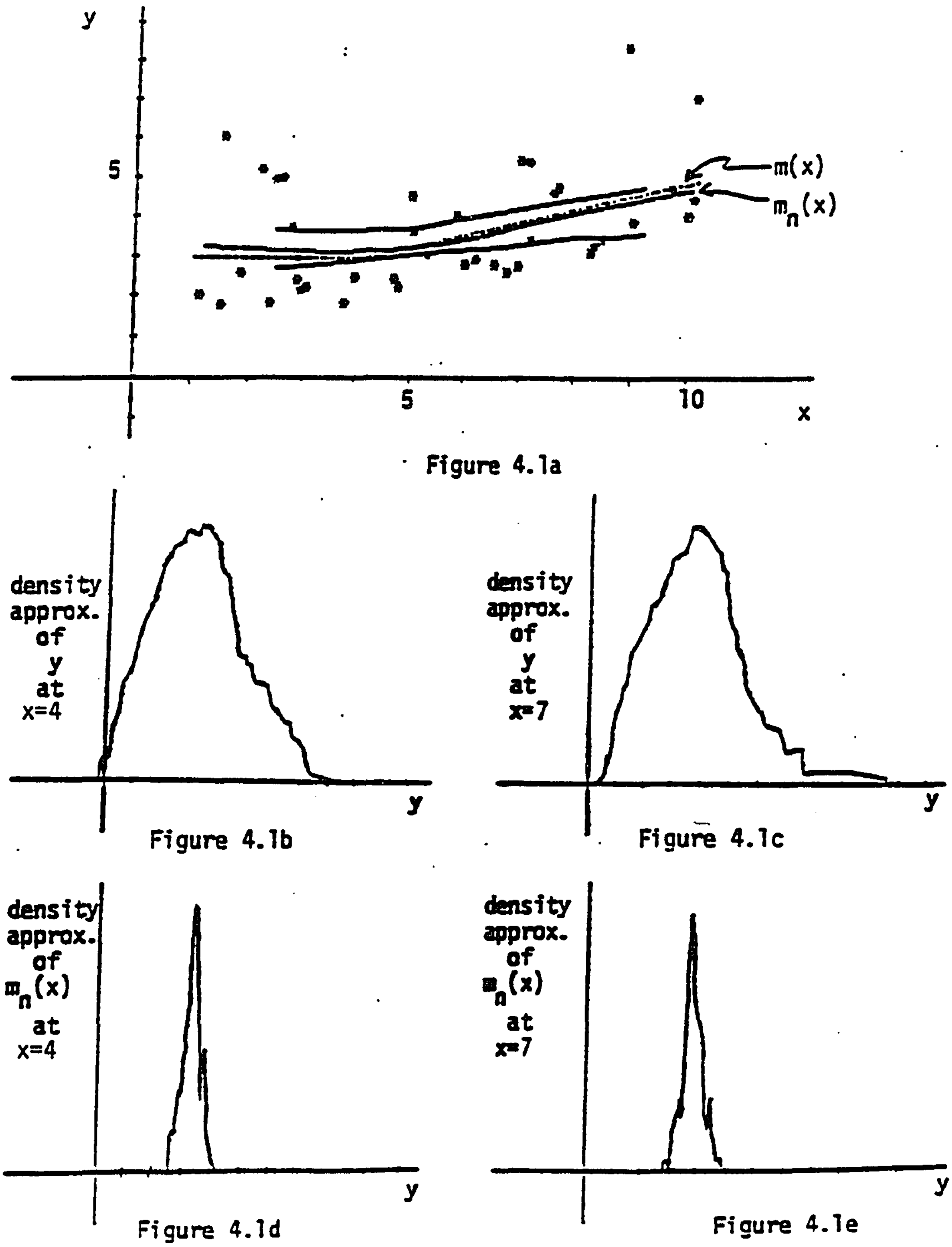


Figure 4.1 Example: Conditional Kernel Method

Method 2 (The Averaged Kernel Method)

For the averaged kernel method and also Method 3, we assume that the conditional density of $Y \backslash X=x$ is fixed about $m(x)$. That is for any x_1, x_2 the densities $f(y-m(x_1)\backslash X=x_1) = f(y-m(x_2)\backslash X=x_2)$. These methods apply only to model a) described in the introduction to this chapter.

Given that this is the case, intuitively, we would get a better approximation to the conditional density of Y if we pool or average the kernel density approximations after adjusting them for differences in mean. This is exactly the motivation for the Averaged Conditional Method. An algorithmic presentation follows.

Algorithm:

1. Take a sample of random pairs yielding the data (x_j, y_j), $j=1,\ldots,n$.

2. Using this data, estimate the parameters for the kernels k', $\underline{K}$ and k in (4.1)and (4.3).

3. At each point x_j, $j=1,\ldots,n$, compute the kernel approximation to the conditional density of Y by (4.1).

4. After subtracting $m_n(x)$ from each of the conditional density estimates in 3, average them so that

$$g_{n,avg}(y) = (1/n)\Sigma_{j=1}^{n}[g_{nx_j}(y-m_n(x_j))].$$

5. At each x_j, j=1,...,n use as an approximated conditional estimate of the density of Y, the density $m_n(x_j) + g_{n,avg}(y)$. Call this $g_{n,avg}(y\backslash X=x_j)$.

6. For any x compute the average conditional approximation of the density of $m_n(x)$ as $g_{nx}^{m}(y) = *_{j=1}^{n}\beta_j(x)g_{n,avg}\ (y\backslash X=x_j)$ where the $\beta_j(x)$ are as in (4.3).

7. Use as $(1-\alpha)$ percent confidence bounds for $m_n(x)$ the $(\alpha/2)$ and $(1-\alpha/2)$ percentiles of the density in 6.

ss with Method 1, we could replace step 6 with a Monte Carlo routine to get the bootstrap distribution.

Figure 4.2 shows the approximation for the average kernel method Figure 4.2a is the same as 4.1a except the confidence bounds were constructed by the average kernel method. Figures 4.2b and 4.2c give the conditional density estimates using this method at the same points as Figures 4.1b and 4.1c. Note that unlike the former approximations the densities here differ from one another only in their location parameter. They are centered at the kernel regression

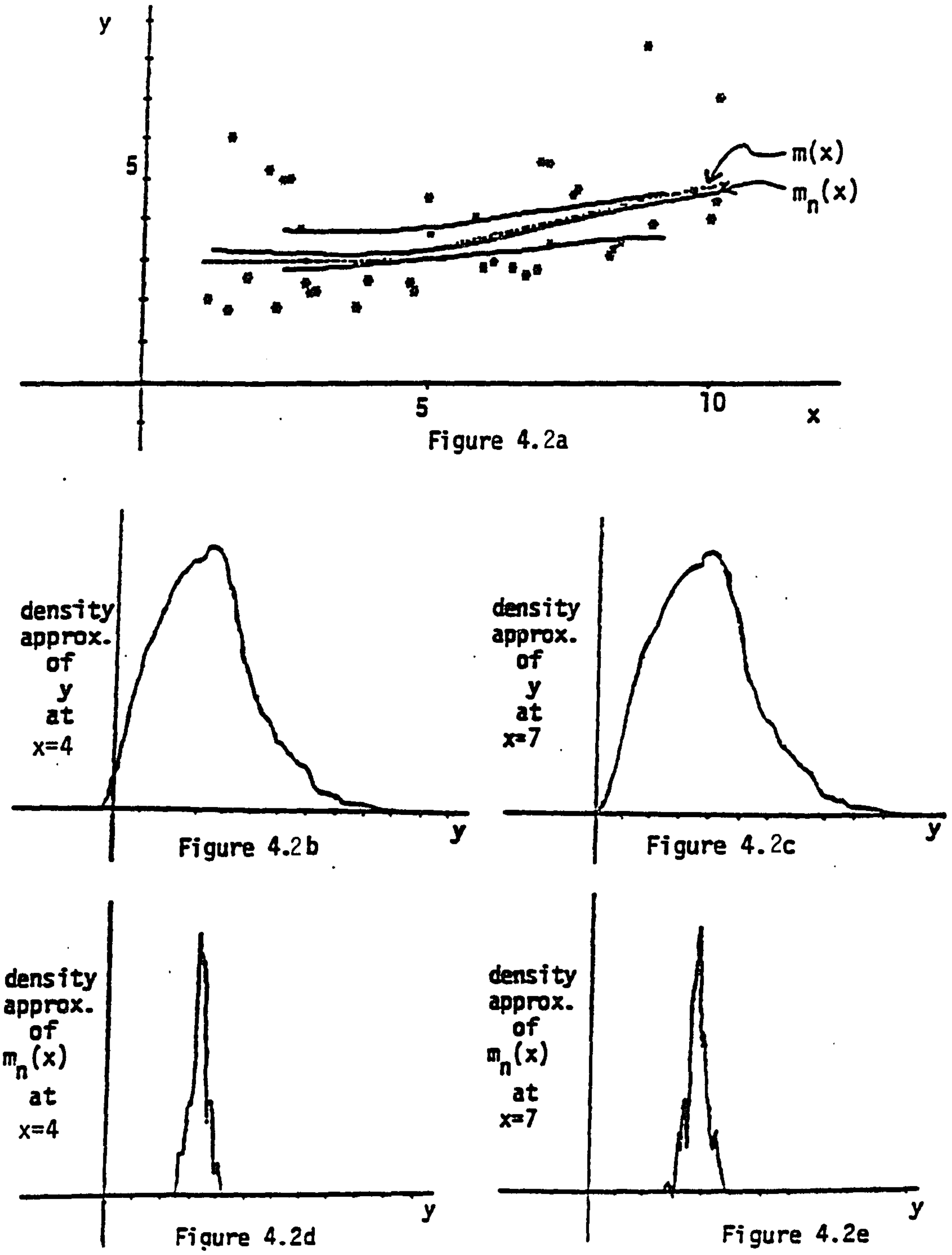


Figure 4.2a

Figure 4.2b

Figure 4.2c

Figure 4.2d

Figure 4.2e

Figure 4.2 Example: Average Kernel Method approximation for each x. Resampling, using these conditional density approximations yielded the bootstrap densities in Figures 4.2d and 4.2e. The x coordinates of these approximations are the same as those used in Figures 4.1d and 4.1e.

Method 3 (The Bootstrap Method)

The bootstrap method was so named because, unlike the previous two methods which were constructed from the reasoning presented in the first algorithm, this method was derived from Freedman's approach (Freedman (1981)) to a bootstrap for linear regression models. It turns out however that there are striking similarities between this and the preceeding algorithms. This method also is designed for homoscedastic data.

Consider the reasoning behind using kernel estimators as they were presented in Chapter 3. Since we only have one Y data value at each X_j, j=1,...,n we utilize surrounding data as information to aid in the construction of an approximate density function. Now with the averaging idea presented in the average kernel method description, we are pooling the information from various points concerning the density function. The idea behind this final method, as compared to the previous methods, is that perhaps since we are pooling we don't need the additional information provided initially by including the surrounding data values. The algorithm for this method follows:

Algorithm:

1. Take a sample of random pairs yielding the data (x_j, y_j), $j=1,\ldots,n$.

2. Using this data, estimate the parameters for the kernel k in (4.1) and $\underline{K}$ in (4.3).

3. Compute the kernel estimate $m_n(x)$ from the data using the kernel regression estimator of Chapter 3.

4. Construct the "resampling" density from the residuals $r_j = Y_j - m_n(x_j)$, $j=1,\ldots,n$. Resampling is in quotes because in this method as in the previous two, we can compute the bootstrap distribution without resorting to the Monte-Carlo approach. Call this density $g_{n,r}(Y)$.

5. For each j approximate the conditional density of Y at $X=x_j$ by $m_n(x_j) + g_{n,r}(Y)$. Call this density $g_{n,r}(Y \backslash X=x_j)$.

6. For any x compute the bootstrap approximation of the density of $m_n(x)$ as $g^m_{nx}(y) = *^n_{j=1}\beta_j(x)\ g_{n,r}(y \backslash X=x_j)$ where the $\beta_j(x)$ are as in (4.3).

7. Use as $(1-\alpha)$ percent confidence bounds for $m_n(x)$ the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the density in 6.

Figure 4.3 gives the approximation for the bootstrap method. Figure 4.3a shows the confidence bounds constructed using this method and Figures 4.3b - 4.3e correspond to the same letters as Figures 4.1 and 4.2 but for the bootstrap method. Note that as in the averaged kernel method, the conditional distribution approximations for different values of x differ only in location parameter.

As a final note in this chapter, recall that only the first of the three methods was to be used with data assumed to be of the form in model b). One possible way to adjust the other two would be to do the averaging procedure only locally and hence use a different averaged distribution at each X_j, $j=1,\ldots,n$.

In the next two chapters we will further investigate these estimators. In Chapter 5 we examine the order of convergence of the conditional kernel estimator to the real distribution of $m_n(x)$ and in Chapter 6 we do a number of computer studies using all three estimators.

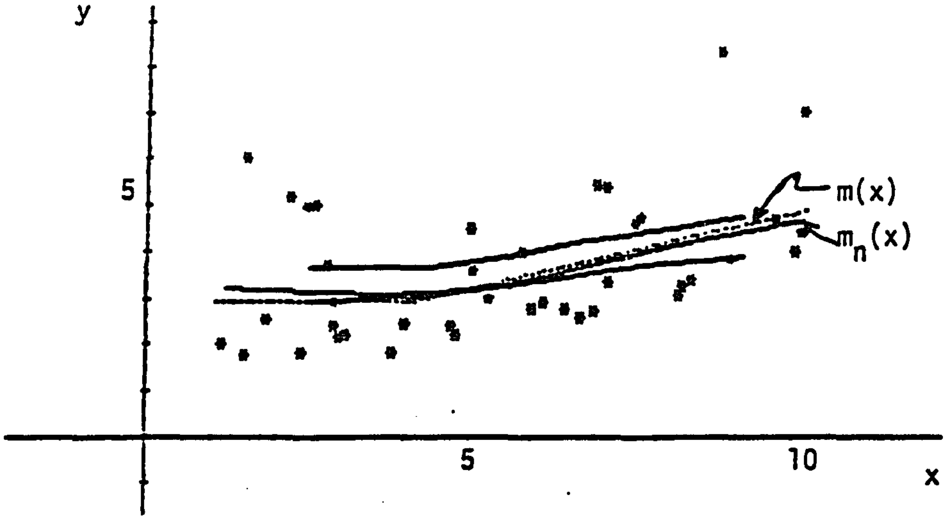


Figure 4.3a

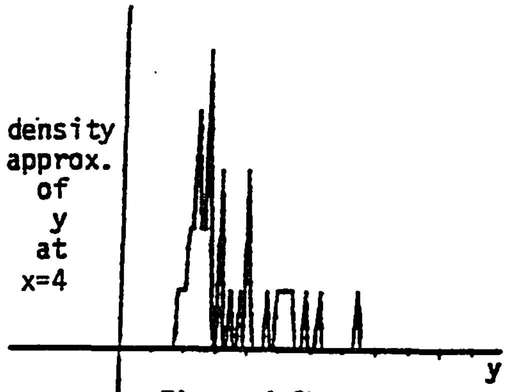


Figure 4.3b

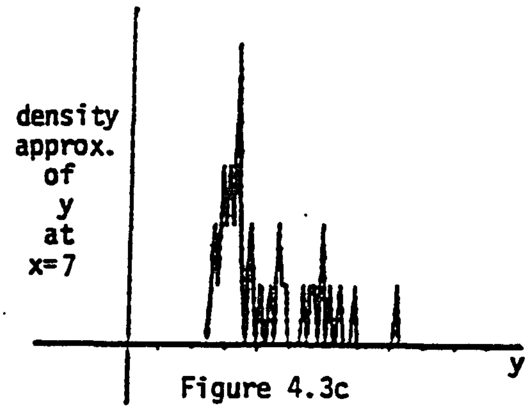


Figure 4.3c

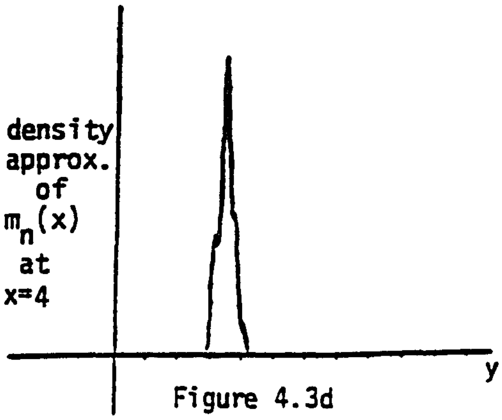


Figure 4.3d

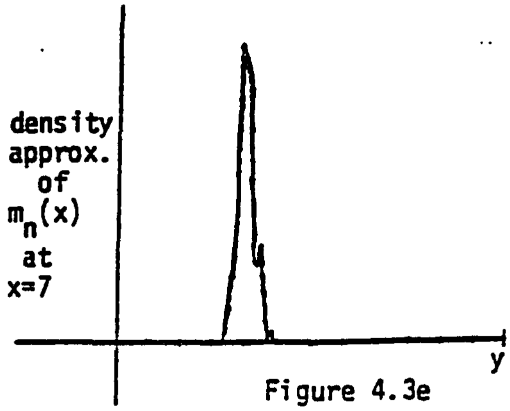


Figure 4.3e

Figure 4.3 Example: Bootstrap Method

5. CONVERGENCE RESULTS

Introduction

Percentiles used in Chapter 6 to give approximate pointwise confidence intervals are taken from approximations to the conditional densities of $m_n(x)$ described in Chapter 4. In the current chapter we investigate the conditions under which these approximations are in some sense better then using the limiting normal density to get percentiles for the approximation. In particular, we compare the order of convergence in distribution to the distribution of $(m_n(x)-m(x))$ of the conditional kernel method, with that of an estimator implied by the asymptotic normality results given by Schuster (1971). We begin the analysis by examining the normal estimator. We briefly review the convergence properties of this estimator as it will be used as a basis of comparison for later results.

The next section introduces all the notation which will be required for this chapter. A fourth section will review several background results which provide the theoretical basis for proving the theorems which follow.

The concluding section gives the main result of this chapter which will be proved through three theorems. Here we give sufficient conditions for a higher order of convergence to the distribution of

$(m_n(x)-m(x))$ of the estimate made by the kernel conditional method when compared to the asymptotic normal approximation.

Order Implied by Asymptotic Normality

The results (which were reviewed in chapter 3) of Schuster (1972) and Johnston (1979) show that under standard conditions the limiting distribution of the random variable $(na_n)^{\frac{1}{2}}(m_n(x)-m(x))$ at any fixed point x is normal with mean zero and standard deviation

$$\sigma^m_{nxf} = [\sigma^2_x \int k^2(u)du/h(x)]^{\frac{1}{2}}$$

where σ^2_x is the variance of $(Y\backslash X=x)$ and $h(x)$ is the marginal density of X.

Alternatively, one can use the normal distribution above to give an approximation to the distribution of $(m_n(x)-m(x))$ by using some available method to estimate $h(x)$ and σ^2_x. This normal approximation will be used here to provide a basis of comparison for the order of convergence.

Directly from the asymptotic normality results we have that

$$\text{Prob}[(na_n)^{\frac{1}{2}}(m_n(x)-m(x))/\sigma^m_{nxf} \le y] - \Phi(y) = o(1)$$

where $\Phi(y)$ is the distribution of the standard normal. One or both of $h(x)$ and σ^2_x may still require estimation. We know that we can approximate $h(x)$ by its kernel density approximation $h_n(x)$ and that $h_n(x) \dashrightarrow h(x)$ with probability 1. A nearest neighbor or kernel approach could also be used to estimate the value of σ^2_x. One such estimator has been proposed in Yakowitz and Szidarovszky (1985) which is shown there to converge in probability to σ_X^2.

It is the primary goal of the remainder of this chapter to give conditions under which the kernel conditional estimate will provide a higher order of convergences then that given above.

Notation

The purpose of this section is to establish the notation essential to proving the theorems which will follow. For the technical development of the proofs, we concern ourselves primarily with an estimated distribution of $(m_n(x)-m(x))$, say $(m_n^*(x)-m_n(x))$, which uses the kernel approximation of the conditional densities of $(Y\backslash X=x_j)$, $j=1,\ldots,n$ for it's construction. This distribution is also conditioned on the sample points determined by the random variables X_j, $j=1,\ldots,n$.

In order to fully describe these distributions and random variables we must assign symbols to represent their characteristic functions, means, variances etc. First we give symbols for the kernel functions which will be required.

Consider the density function estimate for the kernel conditional estimator. We have

$$g_{nx}(y) = a_n^{-1} \frac{\Sigma_{j=1}^{n} K((x-X_j)/a_n,(y-Y_j)/a_n)}{\Sigma_{i=1}^{n} k_x((x-X_i)/a_n)} \tag{5.1}$$

We denote the two dimensional kernel for this estimator by $K(.,.)$ and for the theorems which follow, it will be assumed that $K(.,.)$ can be factored as the product $K_1(.)K_2(.)$ where $K_2(.)$ has characteristic function $\delta_2(t)$. We will refer to $k_x(.)$ as the marginal kernel. The last kernel function we will need will be called the regression kernel. This is the kernel used in the regression model to determine the weights assigned to the conditional densities when approximating the density of $m_n(x)$. The regression kernel will be denoted $\underline{K}((x-X_i)/b_n)$. These kernel functions are listed in Table 5.1 which follows a brief description of the random variables we will be using.

$(Y\backslash X=x_i)$, $i=1,\ldots,n$ are assumed to be the conditional random variables after the X_i have been realized from a sample of independent pairs (X_i,Y_i), $i=1,\ldots n$. $(Y_n\backslash X=x_i)$ is a sequence of random variables again conditional on the sample values of X_i, $i=1,\ldots,n$ but here the Y_i component is assumed to have the density given by the kernel approximation (5.1) at $x = x_i$. These random variables are tabulated in rows one and two of Table 5.2.

The next two random variables in the symbol table are the normalized values of the regression function estimator at the point x taken from a sample of size n and conditioned on the values of X_j, $j=1,\ldots,n$. Row three gives the symbols associated with the random variable where this sample is of the $(Y\backslash X=x_i)$, $i=1,\ldots,n$ while the fourth row is for samples of $(Y_n\backslash X=x_i)$, $i=1,\ldots,n$.

The final two entries in Table 5.2 are not particularly interesting themselves but are essential to the proofs of Theorems 5.1 and 5.2. The random variable in row 5 is used in Theorem 5.1 and the random variable in row 6 is used in Theorem 5.2.

Table 5.1: Summary of Kernel Functions and Related Parameters

Kernel	Discription	Char Fcn.	Bandwidth Param.
$\underline{K}(.)$	regression kernel		b_n
$K(.,.)$	two dimensional kernel		a_n, a_n
$K_1(.)$	first factor of $K(.,.)$		a_n
$K_2(.)$	second factor of $K(.,.)$	$\delta_2(t)$	a_n
$k_x(.)$	marginal kernel		a_n

Table 5.2: Summary of Random Variables, Moments and Char. Functions

R.V.	Mean	Var.	Ch.fcn.	dist.	den
$(Y\backslash X=x)$	$m(x)$	σ_x^2	$\delta_x(t)$	$F_x(y)$	$f_x(y)$
$(Y_n\backslash X=x)$	$m_n(x)$	σ_{nx}^2	$\delta_{nx}(t)$	$G_{nx}(y)$	$g_{nx}(y)$
$(m_n(x)-m(x))$	0	$\Sigma_{i=1}^n \beta_i^2(x)\sigma_{x_i}^2$		$F_{nx}^m(y)$	$f_{nx}^m(y)$
$(m_n^*(x)-m_n(x))$	0			$G_{nx}^m(y)$	$g_{nx}^m(y)$
$(m_n(x)-\bar{m}(x))$	0	$\Sigma_{i=1}^n \beta_i^2(x)\sigma_{x_i}^2$	$\delta_{nxf}^m(t)$	$H_{nx}^m(y)$	$h_{nx}^m(y)$
$(m_n^*(x)-\bar{m}_n(x))$	0		$\delta_{nxg}^m(t)$	$H_{nx}^{m'}(y)$	$h_{nx}^{m'}(y)$

where: $\beta_i(x) = \underline{K}((X_i-x)/b_n)/\Sigma_{j=1}^n \underline{K}((X_j-x)/b_n)$,

$\bar{m}(x) = \Sigma_{i=1}^n \beta_i m(x_i)$ and

$\bar{m}_n(x) = \Sigma_{i=1}^n \beta_i m_n(x_i)$.

We now give the actual formulas for two of the characteristic functions listed above and used in theorems (5.1), (5.2), and (5.3). In those theorems, the following assumptions are in force. Let b_n be the sequence of positive constants associated with the regression kernel so that for any n we have $\underline{K}(x) = \underline{K}(x/b_n)$. Let a_n be the sequence of positive constants associated with the two-dimensional kernel K. Assume further that K satisfies

$$K((x)/a_n,(y)/a_n) = K_1((x)/a_n)K_2((y)/a_n)$$

and let $\delta_2(t)$ be the characteristic function corresponding to $K_2(.)$. Then

$$\text{i)}\quad \delta^m_{nxf}(t) = \pi^n_{j=1}\exp(-itm(x_j)\beta_j(x))\delta_{x_j}(\beta_j(x)t)$$

$$\text{ii)}\quad \delta^m_{nxg}(t) = \pi^n_{j=1}\exp(-itm_n(x_j)\beta_j(x))\Sigma^n_{l=1}\alpha_l(x)\delta_{x_l}(t)\delta_y(a_nt)$$

where:

$$\beta_k(x_j) = \underline{K}((X_k-x_j)/b_n)/\Sigma^n_{l=1}\underline{K}((X_l-x_j)/b_n);$$

$$\alpha_k(x_j) = K_1((X_k-x_j)/a_n)/\Sigma^n_{l=1}k_x((X_l-x_j)/a_n);\ \text{and}$$

$$i^2 = -1.$$

Proof.

Each of the formulas is computed directly from the definition of characteristic function and the relationship that if $\delta_x(t)$ is the characteristic function of X then aX+b has characteristic function $e^{itb}\delta_x(at)$.

Proof of i)

$$\delta^m_{nxf}(t) = \pi^n_{j=1}\delta'_{x_j}(t)$$

where $\delta'_{x_j}(t)$ is the characteristic function of the centered random variables $\beta_j(x)[(Y\backslash X=x_j)-m(x_j)]$. Using the results stated above, we have immediately

$$\delta^m_{nxf}(t) = \pi^n_{j=1}\exp(-itm(x_j)\beta_j(x))\delta_{x_j}(\beta_j(x)t).$$

Proof of ii)

From the proof above we have

$$\delta^m_{nxg}(t) = \pi^n_{j=1}\exp(-itm_n(x_j)\beta_j(x))\delta^*_{nx_j}(\beta_j(x)t)$$

where $\delta^*_{nx_j}(t)$ is the characteristic function associated with the random variable $(Y_n\backslash X=x_j)$. But note that using the relationship $E(U) = E[E(U\backslash Y)]$ where the outer expectation is over the values of Y, we get

$$\begin{aligned}\delta^*_{nx}(t) &= \int\exp(itu)g_{nx}(u)du\\ &= \Sigma^n_{j=1}\int\int\exp(itu)a_n^{-1}\alpha_j(x)K_2((u-y)/a_n)dudF(y\backslash X=x_j)\\ &= \Sigma^n_{j=1}\alpha_j(x)\int\int\exp(it(y+a_nv))K_2(v)dvdF(y\backslash X=x_j)\\ &= \Sigma^n_{j=1}\alpha_j(x)\delta_2(a_nt)\delta_{x_j}(t)\end{aligned}$$

from which ii) follows. Note that both i) and ii) are conditioned on the values of $X_1,X_2,\ldots,X_n$ ∎

Preliminary Results

Before proceeding to the convergence results, it is worthwhile to briefly mention the major theoretical results upon which much of this chapter is based. Edgeworth's expansion for distributions is most important. Let $X_1,\ldots,X_n$ be a sample of size n, distributed as

F. Assume that σ is the population standard deviation μ and μ_3 its mean and third moment respectively and that its characteristic function is given by $\delta^*(t)$. Let F_n^μ represent the distribution of the sum of weighted random variables $(1/n)(X_i-\mu)/(\sigma/\sqrt{n})$. The following theorem is in Feller (1966 Chapter 16).

<u>Theorem</u> (Edgeworth) If F is a non-lattice distribution and μ_3 the 3rd moment exists. Then if $\Phi(x)$ and $\varphi(x)$ represent respectively the standard normal distribution and density functions then

$$F_n^\mu(x) - \Phi(x) - (\mu_3/6\sigma^3\sqrt{n})(1-x^2)\varphi(x) = o(1/\sqrt{n})$$

holds uniformly in x. ∎

The proof of this theorem is also worth mention here. It follows from Esséen's lemma (also given in Feller (1966)).

<u>Lemma</u> (Esséen) Let F(x) be a probability distribution with zero gxpectation and characteristic function Γ. Suppose that F-G vanishes at $+\infty$ and at $-\infty$ and that G(x) has a derivative g(x) such that $|g(x)| \leq M$ for all x. Finally, suppose that g has a continuously differentiable Fourier transform δ such that $\delta(0) = 1$ and $\delta'(0) = 0$. Then

$$|F(x)-G(x)| \leq (1/\pi)\int_{-T}^{T}|(\Gamma(t)-\delta(t))/t|dt + 24M/\pi T$$

holds for all x and for T > 0. ∎

This lemma is used for the previous theorem with:

$$G(x) = \Phi(x) - (\mu_3/6\sigma^3\sqrt{n})(x^2-1)\varphi(x);$$

$$F(x) = F_n^{\mu}(x); \quad \text{and}$$

$$T = a\sqrt{n}$$

where a is chosen to be large enough to satisfy $24|g(x)| < \epsilon a$ for arbitrary but fixed ϵ and all x. This leaves us with the following inequality

$$|F_n^{\mu}(x) - \Phi(x) - (\mu_3/6\sigma^3\sqrt{n})(x^2-1)\varphi(x)|$$

$$\leq \int_{-a\sqrt{n}}^{a\sqrt{n}} \left| \frac{[\delta^*(t/\sigma\sqrt{n})]^n - \exp(-1/2t^2)[1+(\mu_3/6\sigma^3\sqrt{n})(it)^3]}{t} \right| dt + \epsilon/\sqrt{n}.$$

The proof is completed by showing that the right hand side above has the stated order of convergence to zero.

The results reported above have also been extended to the case where the random variables are not identically distributed. This next theorem is also in Feller (1966).

<u>Theorem</u> (Cramér) Assume that for each i, the random variable X_i has the non lattice distribution $F_i(x)$ with mean 0. Assume further that there exist constants c, C, and $q_\tau < 1$ for any $\tau > 0$ such that for all i

$$c < E(|X_i|^k) < C, \quad k = 1, 2, 3, 4 \qquad \text{and}$$

$$\max_{t>\tau}|\delta_i(t)| < q_\tau$$

where $\delta_i(t)$ is the characteristic function of X_i. Then if we let $s_n^2 = \Sigma_{i=1}^n \sigma_i^2$, $F_n^*(x/s_n)$ be the distribution of $(X_1+X_2+\ldots+X_n)/s_n$ and $\mu_{n3} = \Sigma_{i=1}^n E(X_i^3)$ we have that

$$F_n^*(x/s_n) - \Phi(x) - (\mu_{n3}/6s_n^3)(1-x^2)\phi(x) = o(1/\sqrt{n})$$

holds uniformly in x. ∎

In the section which follows, we investigate the conditions under which we can show the order of convergence for the distribution of the kernel conditional approximation $(m_n^*(x)-m_n(x))$ to the true distribution of $(m_n(x)-m(x))$. In the process, we will use the Cramér's theorem for both $F_{nx}^m(y)$ and $G_{nx}^m(y)$.

Main Theorems

The main result of this chapter derives a lower bound for the order of convergence of our estimated distribution for $(m_n^*(x)-m_n(x))$ to the real distribution of $(m_n(x)-m(x))$, both variables being conditioned on our selection of sample values x_i, $i=1,\ldots,n$.

We saw in a previous section (Order Implied by Asymptotic Normality) that the normal approximation to the normalized distribution of $(m_n(x)-m(x))$ had order of convergence $o(1)$. In this section we will show that the conditional kernel estimator (method 1) gives a higher order of convergence under the conditions given in the theorems. We show that under the specified conditions

$$|G^m_{nx}(y/\sigma_m)-F^m_{nx}(y/\sigma_m)| = o((nb_n')^{-1/2})$$

uniformly in y where $G^m_{nx}(y)$ is the kernel approximation of the distribution of $(m_n(x)-m(x))$ which we are calling $(m_n^*(x)-m_n(x))$, $F^m_{nx}(y)$ is the real distribution of $(m_n(x)-m(x))$, σ_m is the appropriate sequence of standard deviations which will be defined in Theorem 5.1 and $b_n' = n^{-\epsilon}b_n$, $\epsilon > 0$. The following technical results will be used latter in the theorems of this section. These first two lemmas are proved for example in Chung (1968).

Lemma 5.1 If the distribution function F(x) has a finite absolute moment of positive integral order k, then its characteristic function has a bounded continuous derivative of order k given by

$$\delta^{(k)}(t) = \int (ix)^k e^{itx} dF(x)$$

where $i^2 = -1$. ■

Lemma 5.2 Let F be a distribution function and δ be the corresponding characteristic function. If F has a finite absolute moment of integral order $k \geq 1$ then $\delta(t)$ has the following expansions in a neighborhood of $t = 0$:

i) $\delta(t) = \Sigma_{j=0}^{k}(i^j/j!)\mu_j t^j + o(|t|^k)$;

ii) $\delta(t) = \Sigma_{j=0}^{k-1}(i^j/j!)\mu_j t^j + (\emptyset_k/k!)\mu_{|k|}|t|^k$

where μ_j is the jth moment of F, $\mu_{|k|}$ is the kth absolute moment of F and $\emptyset_k$ is a specific value satisfying $|\emptyset_k| \leq 1$. ■

This next result applies to the log of the characteristic function.

Lemma 5.3 Let F and δ be as in Lemma (5.2) then $\log[\delta(t)]$ has the following expansions in the neighborhood of t=0:

i) $\log[\delta(t)] = \Sigma_{j=1}^{k}(i^j/j!)\log[\delta(0)]^{(j)}t^j + o(|t|^k)$;

ii) $\log[\delta(t)] = \Sigma_{j=1}^{k-1}(i^j/j!)\log[\delta(0)]^{(j)}t^j$

$+ (i^k/k!)\log[\delta(\emptyset_k t)]^{(k)}t^k$

where the superscript (j) implies jth derivative , and again $\emptyset_k$ is a specific value satisfying $|\emptyset_k| \leq 1$. ∎

Proof of the above lemma follows directly from Taylor's expansion about t=0.

For our purposes F will be centered at zero, which significantly simplifies the derivatives above. Note that:

$$\log[\delta(0)]^{(0)} = \log[1] = 0;$$

$$\log[\delta(0)]^{(1)} = \delta'(0)/\delta(0) = 0;$$

$$\begin{aligned}\log[\delta(0)]^{(2)} &= \delta''(0)/\delta(0)-(\delta'(0))^2/(\delta(0))^2\\ &= \delta''(0)\\ &= \sigma^2;\end{aligned}$$

$$\log[\delta(0)]^{(3)} = \delta'''(0)/\delta(0) - 3\delta''(0)\delta'(0)/(\delta(0))^2 + 2(\delta'(0))^3/(\delta(0))^3$$
$$= \mu_3;$$

and also

$$\log[\delta(t)]^{(4)} = \delta^{(iv)}(t)/\delta(t) - 4\delta'''(t)\delta'(t)/(\delta(t))^2$$
$$- 3(\delta''(t))^2/(\delta(t)^2 + 6\delta''(t)(\delta'(t))^2/(\delta(t))^3$$
$$+ 6(\delta'(t))^2\delta''(t)/(\delta(t))^3 - 6(\delta'(t))^4/(\delta(t))^4. \quad (5.2)$$

We will use these formulas in bounding the errors for several Taylor expansions in Theorem 5.1.

This next technical result used in Feller (1966) is an inequality valid for all complex α and β.

Lemma 5.4 Let α and β be two complex numbers. Then

$$|e^{\alpha}-1-\beta| \le (|\alpha-\beta|+(1/2)\beta^2)e^{\emptyset}$$

where $\emptyset \ge \max(|\alpha|,|\beta|)$. ∎

This last lemma from Devroye (1982) will be used to show that our convergence results hold with probability one.

<u>Lemma 5.5</u> (Devroye) Let $p \in (0,1/2)$ and $n \geq 1$ be given: p may depend upon n. Let $b(i,n,p) = \binom{n}{i} p^i (1-p)^{n-i}$ be the ith binomial probability and let $B(k,n,p) = \Sigma_{i=0}^{k} b(i,n,p)$. If k and p vary in such a way that as $n \to \infty$

i) $k \to \infty$

ii) $k^2/n \to 0$

iii) $k/(np) \to 0$ and

iv) $(np)/\log n \to \infty$, then

$\Sigma_{n=1}^{\infty} B(k,n,p) < \infty$. ■

We now proceed to the main results of this chapter. The first two theorems use Cramer's theorem expansion to compare respectively $F_{nx}^{m}(y)$ and $G_{nx}^{m}(y)$ to $(\Phi(y)+(\mu_{m3}/6\sigma_m^3)(1-y^2)\phi(y))$. The final theorem then compares $F_{nx}^{m}(y)$ to $G_{nx}^{m}(y)$ directly.

To prove this first theorem we will need symbols for the moments and characteristic function of the random variable

$$(m_n(x)-\bar{m}(x)) = \Sigma_{i=1}^{n} \beta_i(x)[(Y_i \backslash X_i=x_i)-m(x_i)] \tag{5.3}$$

where $\bar{m}(x) = \Sigma_{i=1}^{n} \beta_i(x) m(x_i)$ and

$$\beta_i(x) = \underline{K}((x-X_i)/b_n)/\Sigma_{j=1}^{n} \underline{K}((x-X_j)/b_n).$$

Let

$$\sigma_{mf}^2 = \Sigma_{i=1}^{n} \beta_i^2(x) \sigma_{x_{if}}^2$$

$$\mu_{m3f} = \Sigma_{i=1}^{n} \beta_i^3(x) \mu_{x_i 3f} \quad \text{and}$$

$$\mu_{m4f} = \Sigma_{i=1}^{n}\beta_i^4(x)\mu_{x_i4f}$$

be the second, third, and fourth moment of the random variable (5.3) where μ_{x_i3f} and μ_{x_i4f} are the third and fourth moment respectively of $[(Y\backslash X=x_i)-m(x_i)]$. Similarly for absolute moments we use for example $\mu_{m|3|f}= \Sigma_{i=1}^{n}\beta_i^3(x)\mu_{x_i|3|f}$ where $\mu_{x_i|3|f}$ is the third absolute moment of $[(Y\backslash X=x_i)-m(x_i)]$. Also, we write the characteristic function of the random sum (5.3)

$$\delta_{nxf}^{m}(t) = \pi_{i=1}^{n}\delta_{x_i}^{*}(\beta_i(x)t)$$

where $\delta_{x_i}^{*}$ is the characteristic function of $(Y\backslash X=x_i)-m(x_i)$.

To simplify the notation for this next theorem only, we will drop the subscript f from σ_{mf}, μ_{m3f}, etc. We will need these subscripts later in this chapter however as it will be necessary to compare these to σ_{mg}, μ_{m3g} etc. which are defined at that time.

Theorem 5.1 Consider the random variables

$$(m_n(x)-m(x)) = \Sigma_{i=1}^{n}\beta_i(x)(Y\backslash X=x_i)-m(x) \tag{5.4}$$

where again

$$\beta_i(x) = \underline{K}((x-X_i)/b_n)/\Sigma_{j=1}^{n}\underline{K}((x-X_j)/b_n).$$

Assume that the sequence $b_n = n^{-\alpha}$ where $1/2 < \alpha < 1$, $b_n' = n^{-\alpha+\epsilon^*}$

and ϵ^* some positive constant. Assume further that for all x, any $\tau > 0$, and positive constants c, C, B, D, Ω_1, Ω_2 and $q_\tau < 1$ which don't depend upon n, the following assumptions are satisfied:

i) $c < E(|Y \backslash X=x|^k) < C$ for k=1,...,4;

ii) $\max_{t>\tau}|\delta_x^*(t)| < q_\tau$;

iii) The regression kernel $\underline{K}$ is a probability density centered at 0 which has finite support concentrated on an interval [-z,z] with z > 0 and satisfies $0 < \Omega_1 \le K(x) \le \Omega_2 < \infty$ for all $x \in [-z,z]$;

iv) There exists a bound B such that for any x and all y, the conditional density $f(y \backslash X=x) < B$;

v) The model m(x) has a bounded derivative m'(x) satisfying $|m'(x)| < D$ for all x.

vi) h(x) the marginal density of X has a bounded derivative h'(x).

Then if $F_{nx}^m(y/\sigma_m)$ represents the distribution function of the random variable (5.4) after normalizing and $\Phi(y)$ and $\phi(y)$ respectively represent the distribution function and density of the standard normal, we have with probability one

$$|F_{nx}^m(y/\sigma_m)-\Phi(y)-(\mu_{m3}/6\sigma_m^3)(1-y^2)\phi(y)| = o((nb_n')^{-\frac{1}{2}}) \qquad (5.5)$$

uniformly in y for any x satisfying h(x) > 0.

Proof.

First, we show that $|F^m_{nx}(y/\sigma_m) - H^m_{nx}(y/\sigma_m)| = o((nb_n)^{-1/2})$ uniformly in y where $F^m_{nx}(y/\sigma_m)$ is the normalized distribution function of (5.4) and $H^m_{nx}(y/\sigma_m)$ is the normalized distribution function of (5.3). To complete the proof, we then show that

$$|H^m_{nx}(y/\sigma_m) - \Phi(y) - (\mu_{m3}/6\sigma^3_m)(1-y^2)\phi(y)| = o((nb_n')^{-1/2}) \qquad (5.6)$$

uniformly in y.

Consider the difference $|F^m_{nx}(y/\sigma_m) - H^m_{nx}(y/\sigma_m)|$.

These distributions are identical except for their location parameter. This difference is bounded for any y by

$$k[\sup_x m'(x)(b_n z)][\sup_{y,x} f(y\backslash X=x)]/\sigma_m = k(b_n z)DB/\sigma_m$$
$$= k'b_n/\sigma_m$$

for some positive constants k and k' where we are using assumptions iii), iv) and v). The first term in brackets represents the maximum change possible in location parameter for fixed n while the second is proportional to the maximum of the density function at any x and y. Now, using the assumption that $nb_n^2 \to 0$ as $n \to \infty$, we conclude that this difference has order of convergence $o((nb_n)^{-1/2})$. For the remainder of the proof we will work with the distribution $H^m_{nx}(.)$.

The rest of the proof of this theorem follows many of the developments of Feller (1966 Chapter 16). The primary difference here is that weights are changing with n according to the kernel weights. We are in effect dealing with a triangular array of random variables rather then a single sequence. Aside from that, the remainder of this result and what has been called Cramer's theorem (Feller (1966)) appear the same.

Choose x where $h(x) > 0$. We will show that the left hand side of (5.6) is of order $o(\sigma_m^3 \tilde{n})$ where $\tilde{n}$ is the number of X_i satisfying the condition that $|x-X_i|/b_n$ falls within the support of $\underline{K}$. The desired result then follows from the relationship

$$\lim_{n\to\infty} (nb_n')^{1/2}\sigma_m^3\tilde{n} \le B^* \quad \text{a.s.} \tag{5.7}$$

for some bound $B^* > 0$.

To show this we will show that $\lim_{n\to\infty} nb_n'/\tilde{n} \le k$ a.s. (since $\tilde{n}\sigma_m^2$ is bounded) for k a constant. We show that

$$\Sigma_{i=1}^{\infty} p(\tilde{n} < knb_n') < \infty \ . \tag{5.8}$$

Then by the Borel Cantelli lemma we have $P((\tilde{n} < knb_n')$ infinately often) = 0 and the desired result follows.

To get (5.8) we use Lemma 5.5 and note that for fixed n and any point x where $h(x) > 0$, $P(\tilde{n} > j)$ is bounded by a binomial random variable with parameters n and $p = 2b_n zh(r)$ where $h(r) = \sup_x h(x)$. It follows using Lemma 5.5 with $k = nb_n'$ and $p = 2zb_nh(r)$ that as $n \to \infty$: $(nb_n')^2/n \to 0$; $nb_n'/2nzb_nh(r) \to 0$; and $2znh(r)b_n'/\log(n) \to \infty$. (5.8) now follows and we may conclude that $nb_n'/\tilde{n}$ is bounded almost surely. We may conclude that (5.7) is true.

We will now use Esseen's Lemma replacing:

$F(y)$ by $H^m_{nx}(y/\sigma_m)$;

$G(y)$ by $\Phi(y)-(\mu_{m3}/6\sigma^3_m)(y^2-1)\phi(y)$; and using

$T = a/\sigma^3_m\tilde{n}$

where a is chosen small enough so that the derivative of $\Phi(y)-(\mu_{m3}/6\sigma^3_m)(y^2-1)\phi(y)$ is bounded by $\epsilon a\pi/24$ and ϵ is arbitrary but fixed. We get from Esseen's Lemma that

$$|H^m_{nx}(y/\sigma_m)-\Phi(y)-(\mu_{m3}/6\sigma^3_m)(1-y^2)\phi(y)|$$

$$\leq \int_{-a/\sigma^3_m\tilde{n}}^{a/\sigma^3_m\tilde{n}} \left|\frac{\delta^m_{nxf}(t/\sigma_m)-\exp(-t^2/2)[1+(\mu_{m3}/6\sigma^3_m)(it)^3]}{t}\right| dt$$

$$+ \epsilon\sigma^3_m\tilde{n} \qquad (5.9)$$

Since ϵ was chosen arbitrarily, we see that the second term on the right hand side of (5.9) above is of the stated order of convergence $o(\sigma^3_m\tilde{n})$. We must show that the same is true for the integral in (5.9). Assume τ is fixed and satisfies $0 < \tau < a$. We split the range of integration into two regions

$r_1 = |t| < \tau/\sigma^3_m\tilde{n}$

$r_2 = \tau/\sigma^3_m\tilde{n} < |t| < a/\sigma^3_m\tilde{n}$.

The reason for this split is we want to choose the value τ.

The proof will be complete when we show that the integrals over r_1 and r_2 have order of convergence $o(\sigma_m^3\tilde{n})$.

First, consider integration over the region r_2.

$$\int_{r_2}\left|\frac{\delta_{nxf}^m(t/\sigma_m)-\exp(-t^2/2)[1+(\mu_{m3}/6\sigma_m^3)(it)^3]}{t}\right|dt$$

$$=\int_{\tau/\sigma_m^2\tilde{n}<|u|<a/\sigma_m^2\tilde{n}}\left|\frac{\delta_{nxf}^m(u/\sigma_m^2)-\exp(-u^2/2\sigma_m^2)[1+(\mu_{m3}/6\sigma_m^6)(iu)^3]}{u}\right|du$$

where we have used the substitution $u = t\sigma_m$. Note that using assumptions i) and iii)

$$\begin{aligned}\sigma_m^2\tilde{n} &= \Sigma_{i=1}^{\tilde{n}}\beta_i^2\sigma_{x_i}^2\tilde{n}\\ &\le \Sigma_{i=1}^{\tilde{n}}(\Omega_2/(\tilde{n}\Omega_1))^2C\tilde{n}\\ &= (\Omega_2/\Omega_1)^2C\end{aligned}$$

and also

$$\begin{aligned}\sigma_m^2\tilde{n} &\ge \Sigma_{i=1}^{\tilde{n}}(\Omega_1/\tilde{n}\Omega_2)^2c\tilde{n}\\ &= (\Omega_1/\Omega_2)^2c.\end{aligned}$$

So if we let $a_1 = \tau(\Omega_1/\Omega_2)^2c$ and $a_2 = a(\Omega_2/\Omega_1)^2C$ we have that

$$\int_{r_2}\left|\frac{\delta^m_{nxf}(t/\sigma_m)-\exp(-t^2/2)[1+(\mu_{m3}/6\sigma^3_m)(it)^3]}{t}\right|dt$$

$$\leq \int_{a_1<|u|<a_2}\left|\frac{\delta^m_{nxf}(u/\sigma^2_m)-\exp(-u^2/2\sigma^2_m)[1+(\mu_{m3}/6\sigma^6_m)(iu)^3]}{u}\right|du$$

$$\leq \int_{a_1<|u|<a_2}\left|\frac{\delta^m_{nxf}(u/\sigma^2_m)}{u}\right|du$$

$$+\int_{a_1<|u|<a_2}\left|\frac{\exp(-u^2/2\sigma^2_m)[1+(\mu_{m3}/6\sigma^6_m)(iu)^3]}{u}\right|du \qquad (5.10)$$

we show next that the integrals in 5.10 go to zero faster then any power of $1/\tilde{n}$.

Proof for first term in (5.10).

For the first term in (5.10) we have

$$\int_{a_1<|u|<a_2}\left|\frac{\delta^m_{nxf}(u/\sigma^2_m)}{u}\right|du = \int_{a_1<|u|<a_2}\left|\frac{\pi^{\tilde{n}}_{i=1}\delta^*_{x_i}(\beta_i(x)(u/\sigma^2_m))}{u}\right|du \qquad (5.11)$$

where again $\tilde{n}$ is the number of X_i falling within the support of $\underline{K}$ in a sample of size n and the product in (5.11) extends over this set of the X_i renumbered appropriately. We now show that $\beta_i(x)(u/\sigma^2_m)$ is bounded below regardless of n. Using assumptions i) and iii),

$$\beta_i(x)(u/\sigma_m^2) = \beta_i(x)u/\Sigma_{j=1}^n \beta_j^2(x)\sigma_{x_j}^2$$

$$\geq (\Omega_1/\tilde{n}\Omega_2)u/(\tilde{n}(\Omega_2/\tilde{n}\Omega_1)^2 C)$$

$$= (\Omega_1/\Omega_2)^3 u/C.$$

If we let the quantity $(\Omega_1/\Omega_2)^3/C$ be called k_1 then it follows using assumption ii) that

$$\int_{r_2} \left|\frac{\delta^m_{nxf}(t/\sigma_m)}{t}\right| dt \leq \int_{a_1<|u|<a_2} \left|\frac{\pi_{j=1}^{\tilde{n}} q_{a_1}k_1}{u}\right| du$$

$$\leq q^{\tilde{n}}_{a_1 k_1} 2(a_2-a_1)/a_1.$$

This goes to zero faster than any power of $1/\tilde{n}$.

Proof for the second term in (5.10).

Now, for the second term in (5.10) note that $\mu_{m3}/6\sigma_m^6$ goes to infinity with $\tilde{n}$, that is $(\mu_{m3}/6\sigma_m^6)/\tilde{n} = \mathbf{O}(1)$ and $\exp(-u^2/2\sigma_m^2)$ goes to zero faster then any power of $1/\tilde{n}$.

We have now established that the contribution to the value of the integral from the range r_2 goes to zero faster than $1/\tilde{n}$. To conclude the proof, we need only show that the integral in (5.9) over the region r_1 has the order of convergence $\mathbf{o}(\sigma_m^3\tilde{n})$.

Now consider the integral over the range r_1. First, note that

$$\tau/\sigma_m^3\tilde{n} = \tau/[\tilde{n}(\Sigma_{j=1}^{\tilde{n}}\beta_j^2(x)\sigma_{x_j}^2)^{3/2}]$$

$$\le \tau/[\tilde{n}(\Sigma_{j=1}^{\tilde{n}}(\Omega_1/(\Omega_2\tilde{n}))^2 c)^{3/2}]$$

$$= \tau\tilde{n}^{1/2}(\Omega_2/\Omega_1)^3/c^{3/2}.$$

We will instead consider integration over the region r_3 which includes t satisifying

$$|t| < \tau\tilde{n}^{1/2}(\Omega_2/\Omega_1)^3/c^{3/2}.$$

We now show that the integral over this region has the stated order of convergence.

$$\int_{r_3}\left|\frac{\delta_{nxf}^m(t/\sigma_m)-\exp(-t^2/2)[1+(\mu_{m3}/6\sigma_m^3)(it)^3]}{t}\right|dt$$

$$= \int_{r_3}\left|\frac{\exp(-t^2/2)[\exp\{\log[\delta_{nxf}^m(t/\sigma_m)]+t^2/2\}-1-(\mu_{m3}/6\sigma_m^3)(it)^3]}{t}\right|dt$$

$$\le \int_{r_3}\left[\frac{\exp(-t^2/2)\exp(\emptyset)|\log[\delta_{nxf}^m(t/\sigma_m)]+t^2/2-(\mu_{m3}/6\sigma_m^3)(it)^3|}{t}\right.$$

$$\left.+\frac{\exp(-t^2/2)\exp(\emptyset)(1/2)(\mu_{m3}/6\sigma_m^3)^2(it)^6}{t}\right]dt \qquad (5.12)$$

where we have used the inequality given in Lemma 5.4

$$|\exp(\alpha)-1-\beta| \le \exp(\emptyset)(|\alpha-\beta|+(1/2)\beta^2).$$

For the above equation,

$$\emptyset \geq \max\{|\log[\delta^m_{nxf}(t/\sigma_m)]+t^2/2|,|(\mu_{m3}/6\sigma^3_m)(it)^3|\}.$$

We now proceed to determine bounds for each of the terms

$$|\log[\delta^m_{nxf}(t/\sigma_m)]+t^2/2-(\mu_{m3}/6\sigma^3_m)(it)^3|; \quad (5.12a)$$

$$|\log[\delta^m_{nxf}(t/\sigma_m)]+t^2/2|; \quad \text{and} \quad (5.12b)$$

$$|(\mu_{m3}/6\sigma^3_m)(it)^3| \quad (5.12c)$$

in (5.12).

Bound for (5.12a).

To get a bound for (5.12a) we compare $\log[\delta^*_{x_i}(t)]$ to its four term Taylor expansion using the bound provided by it's fourth derivative term and then show that (5.12a) is a weighed sum of these differences. To do this we compute the four term Taylor expansion of $\log[\delta^*_{x_i}(t)]$ about t=0 as

$$(1/2)[\delta^{*''}_{x_i}(0)/\delta^*_{x_i}(0)](it)^2 + (1/6)[\delta^{*'''}_{x_i}(0)/\delta^*_{x_i}(0)](it)^3$$

$$= -\sigma^2_{x_i}t^2/2 + \mu_{x_i3}(it)^3/6.$$

Here we have used the fact that $\delta^*_{x_i}(0) = 1$ and $\delta^{*'}_{x_i}(0) = 0$ as the corresponding random variable $[(Y\backslash X=x_i)-m(x_i)]$ has mean 0. Also using the 4th derivative term's remainder for the Taylor series and

(5.2) we have a bound on the error of this expansion given by $t^4/24 \log[\delta^*_{X_i}(\emptyset t)]^{(iv)}$ where $\emptyset \in [0,1]$ and

$$\log[\delta^*_{X_i}(\emptyset t)]^{(iv)}$$

$$= \delta^{*\,(iv)}_{X_i}(\emptyset t)/\delta^*_{X_i}(\emptyset t) - 4\delta^{*\,'''}_{X_i}(\emptyset t)\delta^{*\,'}_{X_i}(\emptyset t)/(\delta^*_{X_i}(\emptyset t))^2$$

$$-3(\delta^{*\,''}_{X_i}(\emptyset t))^2/(\delta^*_{X_i}(\emptyset t))^2 + 6\delta^{*\,''}_{X_i}(\emptyset t)(\delta^{*\,'}_{X_i}(\emptyset t))^2/(\delta^*_{X_i}(\emptyset t))^3$$

$$+6(\delta^{*\,'}_{X_i}(\emptyset t))^2\delta^{*\,''}_{X_i}(\emptyset t)/(\delta^*_{X_i}(\emptyset t))^3 - 6(\delta^{*\,'}_{X_i}(\emptyset t))^4/(\delta^*_{X_i}(\emptyset t))^4. \quad (5.13)$$

We need to show that this expression is bounded above regardless of i. To do this note that the denominators in (5.13) are constructed solely from factors of $\delta_{X_i}(t)$ which is bounded arbitrarily close to 1 by selecting t close to zero. This follows from examination of its three term Taylor expansion about zero and the corresponding remainder from the second derivative term.

$$|\delta^*_{X_i}(t)-1| = (t^2/2)\delta''_{X_i}(\emptyset' t)$$

$$\leq (t^2/2)C$$

for any i where $\emptyset' \in [0,1]$, C is the bound in assumption i) and we have used Lemma 5.1. Next each of the terms of the numerator in (5.13) is bounded above. We have

$$\delta^{*(iv)}(t) \le \mu_{x_i|4|} \le C;$$

$$\delta^{*'''}(t) \le \mu_{x_i|3|} \le C;$$

$$\delta^{*''}(t) \le \sigma^2_{x_i} \le C; \qquad \text{and}$$

$$\delta^{*'}(t) \le \mu_{x_i|1|} \le C$$

where as before $\mu_{x|k|}$ is the kth absolute moment of Y at X=x. These inequalities follow directly from Lemma 5.1 and from our assumption i). Let $0 < \tau_1 < \min[((2/C)^{1/2}),a]$ where a is given initially in (5.9). Then we have shown using a remainder term of the fourth derivative in Taylor's expansion that

$$|\log[\delta^*_{x_i}(t)]+(\sigma^2_{x_i}/2)t^2-\mu_{x_i3}(it)^3| \le k_2t^4$$

for all $t \in (-\tau_1,\tau_1)$ where k_2 is a constant not dependent on i. It follows that for any positive constant ϵ' we can choose a Γ_1, where $0 < \Gamma_1 < \tau_1$, and Γ_1 depends upon ϵ' but not i, so that if $|t| < \Gamma_1$

$$|\log[\delta^*_{x_i}(t)]+(\sigma^2_{x_i}/2)t^2-\mu_{x_i3}(it)^3| \le \epsilon'|t|^3. \qquad (5.14)$$

We have to extend this bound to $\log[\delta^m_{nxf}(t/\sigma_m)]$.

The four term Taylor expansion $\log[\delta^m_{nxf}(t/\sigma_m)]$ about zero is given by

$$(1/2)(\Sigma_{i=1}^{\tilde{n}}(\beta_i^2(x)/\sigma_m^2)[\delta_{x_i}^{*''}(0)/\delta_{x_i}^{*}(0)])(it)^2$$

$$+ (1/6)(\Sigma_{i=1}^{\tilde{n}}(\beta_i^3(x)/\sigma_m^3)[\delta_{x_i}^{*'''}(0)/\delta_{x_i}^{*}(0)])(it)^3$$

$$= -t^2/2 + (1/6)(\mu_{m3}/\sigma_m^3)(it)^3.$$

Note that the expression (5.12a) which we are trying to bound is just the discrepancy between $\log[\delta_{nxf}^m(t/\sigma_m)]$ and its four term Taylor expansion. We can construct a bound on this difference. Note that

$$|\log[\delta_{nxf}^m(t/\sigma_m)]+t^2/2-(\mu_{m3}/6\sigma_m^3)(it)^3|$$

$$= |\Sigma_{j=1}^{\tilde{n}}\{\log[\delta_{x_j}^{*}(\beta_j(x)t/\sigma_m)]+t^2\beta_j^2(x)\sigma_{x_j}^2/2\sigma_m^2-\beta_j^3(x)\mu_{x_j3}(it)^3/6\sigma_m^3\}|$$

It now follows from (5.14) that for any $\epsilon' > 0$ we have that there exists a constant Γ_1 dependent on ϵ' such that if $t/\sigma_m < \Gamma_1$ then

$$|\log[\delta_{nxf}^m(t/\sigma_m)]+t^2/2-(\mu_{m3}/6\sigma_m^3)(it)^3| \le \tilde{n}\epsilon'(\sup_j\beta_j(x)/\sigma_m)^3|t|^3$$

$$\le \epsilon' k_3|t|^3/\tilde{n}^{1/2} \quad (5.15)$$

where k_3 is a positive constant and we have used the fact that $\beta_j(x) \le \Omega_2/(\Omega_1\tilde{n})$ and $\sigma_m^2 \ge \Omega_1^2/(\Omega_2^2\tilde{n})c$. The bound on this first term is complete.

Bound for (5.12b)

Next, we need a bound for the expression $|\log[\delta^m_{nxf}(t/\sigma_m)] + t^2/2|$ in (5.12b). Comparing $\log[\delta^*_{X_i}(t)] + \sigma^2_{X_i}t^2/2$ to its three term Taylor expansion about t=0 and, using the same arguments as above we have for all $t \in (-\tau_1,\tau_1)$

$$|\log[\delta^*_{X_i}(t)]+(\sigma^2_{X_i}/2)t^2| \le k_4|t|^3$$

where $\tau_1 \in (0,(2/C)^{1/2})$ and k_4 is a positive constant not dependent on i. Again using arguments posed in the proof for (5.12a) it follows that there exists a constant r_2, $0 < r_2 < \tau_1$, independent of n such that if $|t| < r_2$, then

$$|\log[\delta^m_{nxf}(t/\sigma_m)]+t^2/2| \le (1/4)t^2. \qquad (5.16)$$

Bound for (5.12c)

Finally, we need a bound for $|(\mu_{m3}/6\sigma^3_m)(it)^3|$. It is clear that there exists a constant r_3 independent of n whereby $t < r_3$ implies

$$|(\mu_{m3}/6\sigma^3_m)(it)^3| \le (1/4)t^2|\mu_{m3}|/\sigma^3_m$$
$$\le (1/4)t^2. \qquad (5.17)$$

We have now bounded (5.12a), (5.12b) and (5.12c)

We now have for $t < \min\{\Gamma_3, \Gamma_2\}$

$$\max\{|(\mu_{m3}/6\sigma_m^3)(it)^3|, |\log[\delta_{nxf}^m(t/\sigma_m)]+t^2/2|\} \le (1/4)t^2.$$

Putting these inequalities ((5.15), (5.16, (5.17)) together with (5.12) to get

$$\int_{r_3}\left|\frac{\delta_{nxf}^m(t/\sigma_m)-\exp(-t^2/2)[1+(\mu_{m3}/6\sigma_m^3)(it)^3]}{t}\right|dt$$

$$\le \int_{r_3}\exp(-t^2/4)\left(\frac{\epsilon' k_3 t^2}{\tilde{n}^{1/2}} + \frac{\mu_{m3}^2|t|^5}{72\sigma_m^6}\right)dt$$

when τ defining r_3 is chosen so that $\tau < \min\{\Gamma_1, \Gamma_2, \Gamma_3\}$.

To complete the proof we divide by $\sigma_m^3\tilde{n}$ which is equivalent to multiplying by $\tilde{n}^{1/2}k_5$ where k_5 is a positive constant since $\sigma_m^3\tilde{n}/\tilde{n}^{-1/2}$ is bounded for all $\tilde{n}$. This gives us

$$\int_{r_3}\exp(-t^2/4)[\epsilon' k_3 k_5 t^2 + k_6|t|^5/\tilde{n}^{1/2}]dt$$

$$= \epsilon'\int_{r_3}\exp(-t^2/4)k_3 k_5 t^2 dt$$

$$+ 1/\tilde{n}^{1/2}\int_{r_3}\exp(-t^2/4)k_6|t|^5 dt$$

where k_6 is another positive constant.

Both the integrals above are finite over any range. The second integral goes to zero a.s. as n --> ∞ and the first is zero for any $\tilde{n}$ since ϵ' can be made arbitrarily small. This completes the proof. ∎

To prove this next theorem we will need symbols for the moments and characteristic function of the random variables

$$(m_n^*(x)-\bar{m}_n(x)) = \Sigma_{i=1}^n \beta_i(x)[(Y_n \backslash X=x_i)-m_n(x_i)] \qquad (5.18)$$

where $\bar{m}_n(x) = \Sigma_{i=1}^n \beta_i(x) m_n(x_i)$. We will use similar symbols for these new variables as the last theorem wherever possible but now replacing the subscript f by g.

Let

$$\sigma_{mg}^2 = \Sigma_{i=1}^n \beta_i^2(x)\sigma_{x_i g}^2$$

$$\mu_{m3g} = \Sigma_{i=1}^n \beta_i^3(x)\mu_{x_i 3g} \quad \text{and}$$

$$\mu_{m4g} = \Sigma_{i=1}^n \beta_i^4(x)\mu_{x_i 4g}$$

be the second, third and fourth moments of the random variable (5.15) where $\mu_{x_i 3g}$ and $\mu_{x_i 4g}$ are the third and fourth moment respectively of $[(Y_n \backslash X=x_i)-m_n(x_i)]$. Similarly for absolute moments we have for example $\mu_{m|3|g} = \Sigma_{i=1}^n \beta_i^3(x)\mu_{x_i|3|g}$ where $\mu_{x_i|3|g}$ is the third absolute moment of $[(Y_n \backslash X=x_i)-m_n(x_i)]$. Also we write the characteristic function of the random sum (5.18) as

$$\delta_{nxg}^m(t) = \pi_{i=1}^n \delta_{nx_i}^*(\beta_i(x)t)$$

where $\delta_{nx_i}^*$ is the characteristic function of $[(Y_n \backslash X=x_i)-m_n(x_i)]$.

To simplify notation for this next theorem only, the subscript g will be dropped from σ_{mg}, μ_{m3g}, etc.

Theorem 5.2 Consider the random variables

$$(m_n^*(x)-m_n(x)) = \Sigma_{i=1}^n \beta_i(x)(Y_n \backslash X=x_i)-m_n(x) \tag{5.19}$$

where

$$\beta_i(x) = \underline{K}((x-X_i)/b_n)/\Sigma_{j=1}^n \underline{K}((x-X_j)/b_n).$$

Assume that the sequence $b_n = n^{-\alpha}$ where $1/2 < \alpha < 1$ and that the sequence $a_n = n^{-\alpha'}$ where $0 < \alpha' < 1$. Assume in addition that α and α' are such that these sequences satisfy the relationship that $nb_n^2/a_n^2 \to 0$ as $n \to \infty$. Let $K(x,y)$ be a two dimensional density function centered at $\mathbf{0}$ satisfying $K(x,y) = K_1(x)K_2(y)$ where $K_2(y)$ is bounded and has the characteristic function $\delta_2(t)$ and $K_1(x) = k_X(x)$ the marginal kernel.

Assume further that for all x, any τ, and positive constants T, c, C, c*, C*, Ω_1, Ω_2 and q_τ which don't depend upon n, the following assumptions are satisfied:

i') $c < E(|Y \backslash X=x|^k) < C$ for $k=1,\ldots,4$;

ii') $\max_{t>\tau} |\delta_X^*(t)| < q_\tau$;

iii') The regression kernel $\underline{K}$ is a probability density centered at 0 which has finite support concentrated on an interval say $[-z,z]$ with $z > 0$ and satisfies $0 < \Omega_1 \le \underline{K}(x) \le \Omega_2 < \infty$ for all $x \in [-z,z]$;

iv') The kernel $K_2(.)$ has bounded absolute moments

$$0 < c^* \le \int |y|^k K_2(y)dy \le C^* < \infty, \ k=1,2,3,4;$$

v') For all x the random variable $(Y \backslash X=x) \le T < \infty$;

vi') The marginal kernel k_x has a bounded derivative and there is a bound S such that $k_x'(x)/k_x(x) < S$ for any x.

Then if $G_{nx}^m(y/\sigma_m)$ represents the distribution function of (5.19) after normalization and $\Phi(y)$ and $\phi(y)$ respectively represent the distribution function and density of the standard normal, we have with probability one

$$G_{nx}^m(y/\sigma_m)-\Phi(y)-(\mu_{m3}/6\sigma_m^3)(y^2-1)\phi(y) = o((nb_n)^{-\frac{1}{2}})$$

uniformly in y for any x which satisfies $h(x) > 0$.

Proof.

Following the outline to the proof of Theorem 5.1 we will first consider the relationship between the distribution functions of the random sums (5.18) and (5.19). Once this has been established, the proof will be completed using (5.18).

Consider $|G_{nx}^m(y/\sigma_m) - H_{nx}^{m'}(y/\sigma_m)|$ where $G_{nx}^m(.)$ is the distribution of (5.19) and $H_{nx}^{m'}(.)$ is the distribution of (5.18).

The bound to this difference (similar to that used in Theorem 5.1) is given by

$$kzb_n \sup_{x,y} g_{nx}(y) \sup_x m_n'(x)/\sigma_m$$

for some constant k.

Unfortunately, in this case, we cannot fix a bound independent of n for the last two terms above. We will show instead that $m_n'(x) < k_1/a_n$ and $g_{nx}(y) < k_2/a_n$ for each n and positive constants k_1 and k_2. The desired result will then follow from our assumptions concerning the sequences a_n and b_n. That is, $nb_n^2/a_n^2 \to 0$ as $n \to \infty$.

Now, consider $m_n(x)$. For the conditional kernel method, this may be interpreted as the expected value of $Y_n \backslash X=x$ at any point x. The kernel regression function needent be explicitly calculated with this method. So

$$\begin{aligned} m_n(x) &= \int y g_{nx}(y)dy \\ &= \Sigma_{i=1}^{n} a_n^{-1} \alpha_i(x) \int y K_2((y-Y_i)/a_n)dy \\ &= \Sigma_{i=1}^{n} \alpha_i(x) \int (a_n v + Y_i) K_2(v)dv \\ &= \Sigma_{i=1}^{n} \alpha_i(x) Y_i \end{aligned}$$

where as before,

$$\alpha_i(x) = K_1((x-X_i)/a_n)/\Sigma_{j=1}^{n} k_x((x-X_j)/a_n).$$

Now, for $m_n'(x)$

$$m_n'(x) = \frac{\Sigma_{i=1}^n (Y_i/a_n) K_1'((x-X_i)/a_n)}{\Sigma_{j=1}^n k_x((x-X_j)/a_n)}$$

$$- \frac{\Sigma_{i=1}^n (Y_i/a_n) K_1((x-X_i)/a_n) \Sigma_{j=1}^n k_x'((x-X_j)/a_n)}{(\Sigma_{j=1}^n k_x((x-X_j)/a_n))^2}.$$

Using assumptions v') and vi') and that $K_1(.) = k_x(.)$ we may conclude that $|\sup_x m_n'(x) a_n|$ is bounded for all n.

Next consider

$$g_{nx}(y/\sigma_m) = a_n^{-1} \frac{\Sigma_{i=1}^n K_1((x-X_i)/a_n) K_2((y/\sigma_m - Y_i)/a_n)}{\Sigma_{j=1}^n k_x((x-X_j)/a_n)}.$$

We use the fact that K_2 is bounded and that $K_1(.) = k_x(.)$ to conclude that $a_n \sup_{x,y} g_{nx}(y/\sigma_m)$ is bounded regardless of n. We have shown that

$$|G_{nx}^m(y/\sigma_m) - H_{nx}^{m'}(y/\sigma_m)| = o((nb_n)^{-1}).$$

In the remainder of the proof we consider only the sum (5.18).

To complete the proof of this theorem, we will show that under assumptions i') through iv') above, the random variables $(Y_n \backslash X=x_i)$ satisfy the properties i) and ii) of Theorem 5.1. Once this has been established, the arguments used in that theorem carry through here with $H_{nx}^{m'}(.)$ replacing $H_{nx}^m(.)$. Note that the conditions iv) and v) of Theorem 5.1 would be relaxed when using the current bandwidth parameters a_n and b_n. Here, with further restrictions on a_n and b_n we use v') and vi') instead.

We must show that for all x and any τ:

$$\text{a)} \quad c' < E(|Y_n\backslash X=x|^k) < C' \text{ for } k = 1,2,3,4$$

$$\text{b)} \quad \max_{t>\tau}|\delta^*_{nx}(t)| < q'_\tau \tag{5.20}$$

for some c', C', and $q'_\tau < 1$ positive constants which don't depend upon n.

To prove a), consider the kth absolute moment of $(Y_n\backslash X=x)$ for any x

$$E(|Y_n\backslash X=x|^k) = E[E(|Y_n\backslash X=x|^k)\backslash X_1=x_1,\ldots X_n=x_n]$$

where the outer expectation is taken with respect to the conditional probability measures associated with the random variables $(Y_1\backslash X_1=x_1),\ldots,(Y_n\backslash X_n=x_n)$.

We have

$$E(|Y_n\backslash X=x|^k) = \int\ldots\iint|u|^k g_{nx}(u)\,du\,dF(y_1\backslash X_1=x_1)\ldots dF(y_n\backslash X_n=x_n)$$

$$= \int\ldots\iint\Sigma_{i=1}^n|u|^k\alpha_i(x)a_n^{-1}K_2((u-y_i)/a_n)\,du$$

$$dF(y_1\backslash X_1=x_1)\ldots dF(y_n\backslash X_n=x_n)$$

$$\le \sup_i \iint|u|^k a_n^{-1}K_2((u-y_i)/a_n)\,du\,dF(y_i\backslash X=x_i)$$

$$= \sup_i \iint|a_n v+y_i|^k K_2(v)\,dv\,dF(y_i\backslash X=x_i)$$

$$= \sup_i \iint (|a_n^k v^k| + c_1|(a_n v)^{k-1} y_i| + \ldots + c_{k-1}|a_n v y_i^{k-1}| + |y_i^k|)$$
$$K_2(v)dvdF(y_i \backslash X = x_i)$$

for constants $c_1, \ldots, c_{k-1}$. Note that each of the integrals containing a_n will go to zero. For example,

$$\sup_i \iint |a_n v y_i^{k-1}| K_2(v)dvdF(y_i \backslash X_i = x_i)$$
$$\leq a_n C^* \int |y_i^{k-1}| dF(y_i \backslash X_i = x_i)$$
$$\leq a_n C^* C$$

which goes to zero as $n \rightarrow \infty$. C^* is from constraint iv') and C is from i'). This leaves

$$\sup_i \iint |y_i^k| K_2(v)dvdF(y_i \backslash X_i = x) = \sup_i \int |y_i^k| dF(y_i \backslash X_i = x_i)$$

which is bounded by constraint i').

Next, to show that b) in (5.20) holds, recall from the proof of ii) in determining the characteristic functions earlier in this chapter

$$\delta^*_{nx}(t) = \Sigma_{j=1}^n \exp(-itm_n(x))\alpha_j{}'x)\delta_y(a_n t)\delta_{x_j}(t)$$
$$\leq \Sigma_{j=1}^n \alpha_j(x)\delta_{x_j}(t)$$
$$\leq \sup_x \delta_x(t)$$
$$\leq q_\tau$$

for $t > \tau$. Note that the $\delta^*_{nx}(t)$ defined here is $\exp(-itm_n(x))\delta^*_{nx}(t)$ in the proof for the derivation of characteristic functions. This completes the proof. ■

Theorem 5.3 Consider the random variables $(m_n(x)-m(x))$ and $(m_n^*(x)-m_n(x))$ and their distribution functions $F_{nx}^m(y)$ and $G_{nx}^m(y)$ respectively. Assume that a_n and b_n are sequences as defined in Theorem 5.2 with the further assumption that $a_n^{-2}b_n \to 0$ as $n \to \infty$. Let the assumptions i'),through vi') of Theorem 5.2 be satisfied as well as iv) and vi) of Theorem 5.1. Finally, assume that first, second and third moments of Y as a function of x are differentiable with bounded derivatives. Then with probability one

$$F_{nx}^m(y/\sigma_{mf}) - G_{nx}^m(y/\sigma_{mg}) = o((nb_n)^{-1/2})$$

uniformly in y for any x satisfying $h(x) > 0$.

Proof.

We established in the preceeding two theorems that

$$F_{nx}^m(y/\sigma_{mf})-\Phi(y)-(\mu_{m3f}/6\sigma_{mf}^3)(1-y^2)\phi(y) = o((nb_n)^{-1/2})$$

and

$$G_{nx}^m(y/\sigma_{mg})-\Phi(y)-(\mu_{m3g}/6\sigma_{mg}^3)(1-y^2)\phi(y) = o((nb_n)^{-1/2}).$$

It follows that we must show

$$[(\mu_{m3f}/6\sigma_{mf}^3) - (\mu_{m3g}/6\sigma_{mg}^3)](1-y^2)\phi(y) = o((nb_n)^{-1/2})$$

uniformly in y for all x where $h(x) > 0$. Since $(1/6)(1-y^2)\phi(y)$ is bounded we need

$$(nb_n)^{1/2} \left| \frac{\mu_{m3f}}{\sigma^3_{mf}} - \frac{\mu_{m3g}}{\sigma^3_{mg}} \right| \text{ --> } 0 \qquad \text{as } n \text{ --> } \infty.$$

The left hand side is equal to

$$(nb_n)^{1/2} \left| \frac{\Sigma^n_{i=1}\beta^3_i(x)\mu_{x_i 3f}}{(\Sigma^n_{j=1}\beta^2_j(x)\sigma^2_{x_j f})^{3/2}} - \frac{\Sigma^n_{i=1}\beta^3_i(x)\mu_{x_i 3g}}{(\Sigma^n_{j=1}\beta^2_j(x)\sigma^2_{x_j g})^{3/2}} \right|$$

$$\leq (nb_n)^{1/2} \left| \frac{\Sigma^n_{i=1}\beta^3_i(x)\mu_{x_i 3f}}{(\Sigma^n_{j=1}\beta^2_j(x)\sigma^2_{x_j f})^{3/2}} - \frac{\Sigma^n_{i=1}\beta^3_i(x)\mu_{x3f}}{(\Sigma^n_{j=1}\beta^2_j(x)\sigma^2_{xf})^{3/2}} \right| \tag{5.21}$$

$$+ (nb_n)^{1/2} \left| \frac{\Sigma^n_{i=1}\beta^3_i(x)\mu_{x_i 3g}}{(\Sigma^n_{j=1}\beta^2_j(x)\sigma^2_{x_j g})^{3/2}} - \frac{\Sigma^n_{i=1}\beta^3_i(x)\mu_{x3g}}{(\Sigma^n_{j=1}\beta^2_j(x)\sigma^2_{xg})^{3/2}} \right| \tag{5.22}$$

$$+ (nb_n)^{1/2} \left| \frac{\Sigma^n_{i=1}\beta^3_i(x)\mu_{x3f}}{(\Sigma^n_{j=1}\beta^2_j(x)\sigma^2_{xf})^{3/2}} - \frac{\Sigma^n_{i=1}\beta^3_i(x)\mu_{x3g}}{(\Sigma^n_{j=1}\beta^2_j(x)\sigma^2_{xg})^{3/2}} \right|. \tag{5.23}$$

Each summand (5.21), (5.22) and (5.23) goes to zero almost surely if the difference between numerators and the difference between denominators go to zero a.s. and the denominators are bounded away from zero. Consider first (5.21) after multiplying numerator and denominator by $\tilde{n}^{3/2}$. To show that the denominator is bounded away from zero note that

$$\tilde{n}^{3/2}(\Sigma_{i=1}^{n}\beta_i^2(x)\sigma_{X_i f}^2)^{3/2} \geq ((\Omega_1/\Omega_2)c)^{3/2}$$

$$\tilde{n}^{3/2}(\Sigma_{i=1}^{n}\beta_i^2(x)\sigma_{Xf}^2)^{3/2} \geq ((\Omega_1/\Omega_2)c)^{3/2} \qquad (5.24)$$

regardless of $\tilde{n}$. Next, consider the numerator of (5.21). We must show that

$$(nb_n)^{1/2}\ \tilde{n}^{3/2}\ \Sigma_{i=1}^{n}\beta_i^3(x)[\mu_{X_i 3f}-\mu_{X3f}] \ \text{-->}\ 0 \quad \text{a.s. as } n \ \text{-->}\ \infty$$

But

$$\tilde{n}^2\Sigma_{i=1}^{n}\beta_i^3(x)[\mu_{X_i 3f}-\mu_{X3f}] \leq (\Omega_2/\Omega_1)zb_n\sup_x\mu'_{X3f}$$

where: Ω_1, Ω_2 and z are from the constraint iii') and

$$\mu'_{X3f} = (d/dx)\mu_{X3f}.$$

The desired result now follows from our constraints on μ'_{X3f} and the fact that $nb_n/\tilde{n} = \mathbf{O}(1)$ a.s. A similar result holds for the denominator of (5.21).

Turning next to (5.22), we have corresponding to (5.24)

$$\tilde{n}^{3/2}(\Sigma_{i=1}^{n}\beta_i^2(x)\sigma_{X_1 g}^2)^{3/2} \geq ((\Omega_1/\Omega_2)c')^{3/2}$$

$$\tilde{n}^{3/2}(\Sigma_{i=1}^{n}\beta_i^2(x)\sigma_{Xg}^2)^{3/2} \geq ((\Omega_1/\Omega_2)c')^{3/2} \qquad (5.25)$$

for some c' regardless of $\tilde{n}$ using constraints iii') and i') and the results of theorem 5.2 (c' is a consequence of theorem 5.2). For the numerator of (5.22)(numerator and denomonator multiplied by $\tilde{n}^{3/2}$), we have that

$$\tilde{n}^2\Sigma_{i=1}^n\beta_i^3(x)[\mu_{x_i3g}-\mu_{x3g}] \le (\Omega_2/\Omega_1)zb_n\sup_x\mu'_{x3g}$$

where these bounds are now from iii') and

$$\mu'_{x3g} = \int y^3(d/dx)g_{nx}(y)dy$$

We must show that $b_n\sup_x\mu'_{x3g}$ --> 0 as n --> ∞. Note that

$$\mu'_{x3g} = \Sigma_{i=1}^n\int y^3K_2((y-Y_i)/a_n)dy$$

$$\cdot\; a_n^{-2}\left|\frac{K_1'((x-X_i)/a_n)}{\Sigma_{j=1}^nK_1((x-X_j)/a_n)} + \frac{K_1((x-X_i)/a_n)\Sigma_{j=1}^nK_1'((x-X_j)/a_n)}{(\Sigma_{j=1}^nK_1((x-X_j)/a_n))^2}\right|$$

The third moment of K_2 is bounded by our assumptions and so is the ratio $K_1'(.)/K_1(.)$. Our assumptions concerning the sequences a_n and b_n now shows that the numerator in (5.22) goes to zero. A similar argument shows that the difference between denominators also goes to zero.

Finally, consider (5.23). That the denominator (with numerator and denominator multiplied by $\tilde{n}^{3/2}$) is bounded away from zero follows from (5.24) and (5.25). The difference in numerators in (5.23) is given by

$$\tilde{n}^{3/2}(nb_n)^{1/2}\Sigma_{i=1}^n\beta_i^3(x)[\mu_{x3f} - \int\cdots\int y^3\Sigma_{j=1}^na_n^{-1}\alpha_j(x)K_2((y-u)/a_n)dyf_{x_j}(u)du]$$

but note that

$$\tilde{n}^2\Sigma_{i=1}^n\beta_i^3(x)[\mu_{x3f} - \int\cdots\int y^3\Sigma_{j=1}^na_n^{-1}\alpha_j(x)K_2((y-u)/a_n)dyf_{x_j}(u)du]$$

$$\le \Omega_2/\Omega_1[\mu_{x3f} - \Sigma_{j=1}^n\alpha_j(x)a_n^{-1}\iint y^3K_2((y-u)/a_n)dyf_{x_j}(u)du].$$

We need to show that the sum goes to μ_{x3f} as $n \to \infty$. Toward this end, note that

$$\iint a_n^{-1} y^3 K_2((y-u)/a_n)\,dy f_{x_j}(u)du$$

$$= \iint (a_n v+u)^3 K_2(v)dv f_{x_j}(u)du$$

Each of the terms containing a_n will go to zero as $n \to \infty$ leaving us with

$$\int u^3 df_{x_j}(u) = \mu_{3x_j f}.$$

The conditions given in this theorem on the continuity of conditional moments as well as the fact that $\tilde{n}/na_n = O(1)$ now give us the desired result. A similar argument with the denominators in (5.23) completes the proof. ∎

6. Computer Studies

Introduction

The purpose of this chapter is to report the results of several computer studies designed to provide some indication of the advantages and disadvantages of the estimators and methods presented in this dissertation.

The second section of this chapter describes the computer studies and the purpose of each. The third section gives a brief summary of the methods used to establish various parameters in the experiments. The final three sections give the results of the experiments with some concluding remarks.

The Experiments

Three separate experiments were run each with the intention of shedding some light on how the estimators compare to one another or how the methods for computing the bootstrap distribution discussed earlier compare. The first pretains to the methods.

The first experiment compares the two methods of constructing the bootstrap distribution of $m_n(x)$ from the conditional distribution of Y. We examine the Monte Carlo procedures and the numerical convolution method both discussed in Chapter 4. The results presented here give a report of both accuracy and computer time.

The second experiment compares each of the following estimators:

1) normal approximation I;

2) normal approximation II (with moments estimated locally);

3) conditional kernel estimator (method 1);

4) averaged kernel estimator (method 2);

5) bootstrap estimator (method 3).

The specific question we wish to address with this simulation study is how do these estimators compare when tested using different distributions of (Y\X=x) and different models m(x). We generate data for three different distributions: exponential, uniform and normal. We also consider a linear model, a step function and a piecewise linear approximation to a smooth curve for m(x). Other factors are kept constant. This includes:

a) Range of X = (1,10);

b) Distribution of X = uniform;

c) Sample size = 20.

The data recorded for each of the estimators above is summarized from 27 computer runs. Three runs are generated for each model-distribution combination.

The final computer study is designed to demonstrate how the conditional kernel method performs with hetroscedastic data compared with the other methods used in the previous experiment.

Bandwidth Selection and Other Procedural Considerations

Several parameters had to be determined for each of the runs. This section describes the methods used in choosing these parameters.

First, several kernel bandwidth parameters were required. The bandwidth parameter for the kernel regression estimator was selected by using cross validation. Here one data pair at a time, say (x_i, y_i), is eliminated and the kernel regression function is estimated. The value of the regression function estimate at x_i is compared to y_i and the squared difference is recorded. Repeating this procedure for $i=1,\ldots,n$ and summing the squared differences gives a measure of error for a fixed bandwidth parameter. The bandwidth size which yields the smallest sum of squared differences is selected. Possible values of the bandwidth parameter are selected by a line search procedure.

The bandwidth parameter for density estimation is chosen by the method of maximum likelihood. Here again, one data value or pair is eliminated at a time from consideration (say x_i in the case of a univariate density) and the density is estimated. The estimated density function value at the eliminated point is computed and the products of these values, where i ranges from 1 to n, are compared in search of a maximum. The candidate values for the bandwidth parameter are again selected by a line search procedure.

Several parameters require estimation for each run. n estimates of h(x) are needed for normal approximation I. A univariate kernel density estimate is used. The maximum likelihood method as described above is used to choose from several candidates for the bandwidth parameter.

Estimation of the moment σ_x^2 differs for the two normal approximations. For normal approximation I:

$$\sigma_{xI}^2 = 1/19\Sigma_{i=1}^{20}(Y_i - m_n(x_i))^2.$$

For Normal Approximation II:

$$\sigma_{xII}^2 = 1/4\Sigma(Y_i - \bar{Y}_5)^2$$

where the sum is over the set of 5 nearest neighbors in the x direction to the point x and $\bar{Y}_5$ is the average of the 5 corresponding Y coordinates.

To get the variance for normal approximation I, we compute

$$\sigma_{mnx}^2 = \sigma_{xI}^2 \int \underline{K}^2(u)du/(h(x)na_n). \tag{6.1}$$

To get the variance for normal approximations II, we compute

$$\sigma_{mnx}^2 = \Sigma_{i=1}^n \underline{K}^2((x-X_i)/a_n)\sigma_{x_i II}^2/(\Sigma_{j=1}^n \underline{K}((x-X_j)/a_n))^2. \tag{6.2}$$

The computation of the third moment was required for the approximation discussed in Chapter 5 based on the three term Edgeworth expansion. This expansion was computed for each of the runs reported in study 2 but it was only in vary rare instances that the third moment term made any difference at all and then it was not very significant. For this reason only the usual normal approximation results are reported for study 2.

One other concern of the experiments is the handling of points on the outer extremes of the sample space where the bias of weighting samples to one side of a given point is greatest. Procedures for selecting a different bandwidth for that region have been suggested in the literature. This problem has been reduced here by eliminating from consideration all parts of the regression interval except the middle 80 percent between first and last X values in the sample.

Study 1

This first study is designed to compare methods for computing the bootstrap distribution once conditional density estimates have been calculated. The two methods to be compared are the Monte Carlo procedure and the numerical convolution procedure described in Chapter 4. Since neither of the codes for these procedures is professionally designed or tested, this study is included here only to justify the procedures used in the remaining experiments.

The convolution procedure was dependent upon the grid used to compute the approximate conditional densities, so the grid used for the conditional densities must vary along with the grid used for the convolution procedure. The Monte Carlo procedure doesn't have this same dependence but for the experiment described here the conditional density grid was changed to match the convolution procedure as shown in Table 6.1. The following table and Figure 6.1 summarize the results.

Table 6.1 Methods for Constructing the Bootstrap Distribution

	Cond. Density	Monte Carlo	Convolution
	num. of grid pts.	num. of resamples	num. of grid pts.
dotted line	30	10	30
dashed line	50	50	50
solid line	150	100	150

In Figure 6.1 which follows, note that the solid lines give almost identical bounds. These bounds are almost identical to the true bounds.

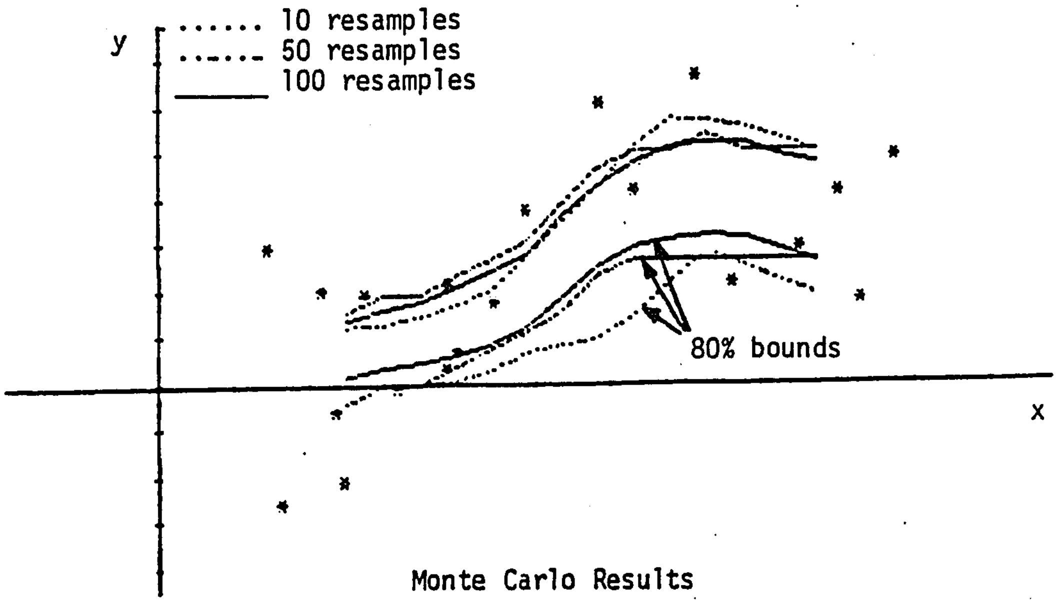
Computing time considerations lead to the use of the Monte Carlo procedure as the degree of accuracy appears almost twice as good for a unit of C.P.U. time.

The experiments of Study 2 and Study 3 use the Monte Carlo procedure with 250 bootstrap replications.

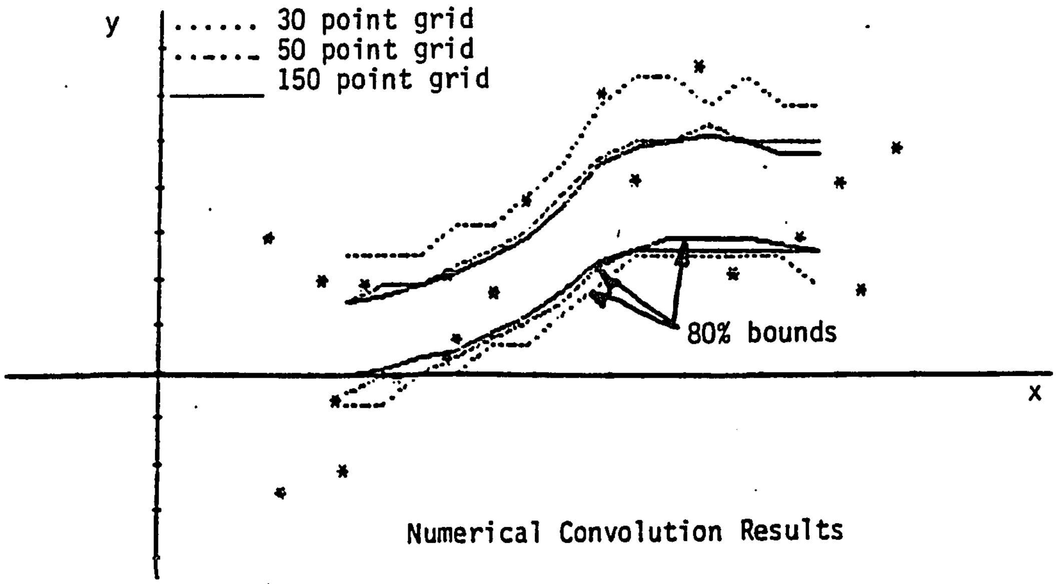
Study 2

In this section, we compare the actual performance of the estimators on 27 generated data sets. Figures 6.2a, 6.3a and 6.4a illustrate the three models and examples of the data sets used in the study.

The first model (6.2a) is piecewise linear approximation of a smooth curve. The data for this run is exponential with variance 1.



Monte Carlo Results



Numerical Convolution Results

Figure 6.1. Constructing the Bootstrap Density

The mean of the exponential is placed at the model value m(x) for any given x.

The next model (6.3a) is linear. Data for this run was generated from a uniform distribution with variance 2. The distribution was centered at the model value m(x) for any given value of x.

The last model (6.4a) is a step function. The normal distribution centered at m(x) with variance 2 was used to generate this data.

These figures show only a sample of the runs which were performed (run 1, run 14 and run 27). Each distribution was run with each model three times. Note also that in each case the x coordinates were selected randomly from a uniform distribution with the support [1,10].

The series of figures illustrates the performance of each method for the three sample runs described above. For each example 6 drawings are given:

Figure a is the model and kernel regression estimate;

Figure b is for the normal approximation I;

Figure c is for the normal approximation II;

Figure d is for the conditional kernel method;

Figure e is for the averaged kernel method;

Figure f is for the bootstrap method.

Each figure (b-f) contains the data, the true model (dotted line) and 80% confidence bounds (solid lines) computed by each method as specified. The confidence coefficient is interpreted as a pointwise probability of the bounds including the value of the model m(x).

The output retained from each run consisted of the average confidence bound width and the percent of the time the model was within the bounds. These quantities were computed by comparing the values at fixed points in x. A table giving a run by run account of these values is included as an appendix.

The data is summarized in three 3 ways: overall, by model and by distribution. These summaries appear next in Table 6.2.

Table 6.2 Bound Width and Percent Coverage Summary (20 pts.)

	Norm I		Norm II		Meth 1		Meth 2		Meth 3	
	%	wth.	%	wth.	%	wth.	%	wth.	%	wth.
overal.	.524	.716	.7	1.13	.810	1.40	.787	1.29	.663	1.00
curve	.554	.849	.772	1.40	.968	1.67	.960	1.61	.776	1.21
linear	.591	.558	.757	.8	.797	1.08	.757	.983	.666	.821
step	.429	.742	.577	1.19	.663	1.44	.644	1.29	.546	.978
expon.	.41	.637	.602	.936	.716	1.23	.681	1.22	.554	.875
unif.	.559	.650	.667	1.08	.836	1.35	.812	1.22	.679	.933
normal	.605	.862	.698	1.37	.877	1.62	.868	1.44	.754	1.20

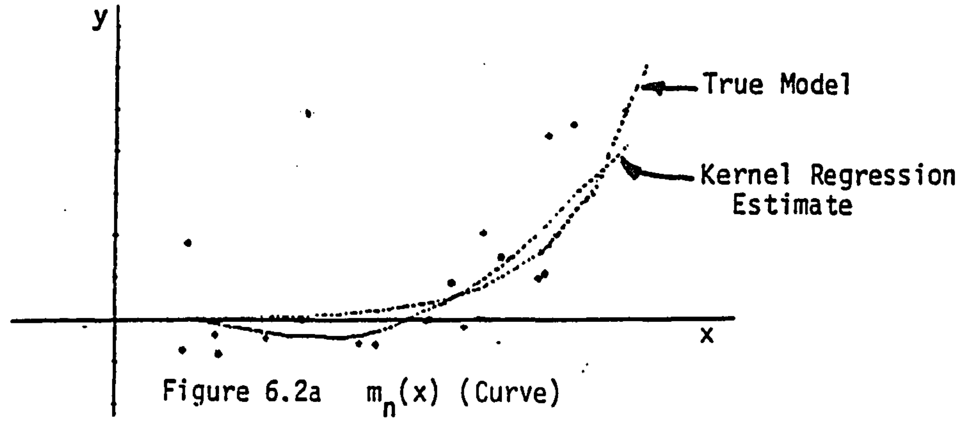


Figure 6.2a $m_n(x)$ (Curve)

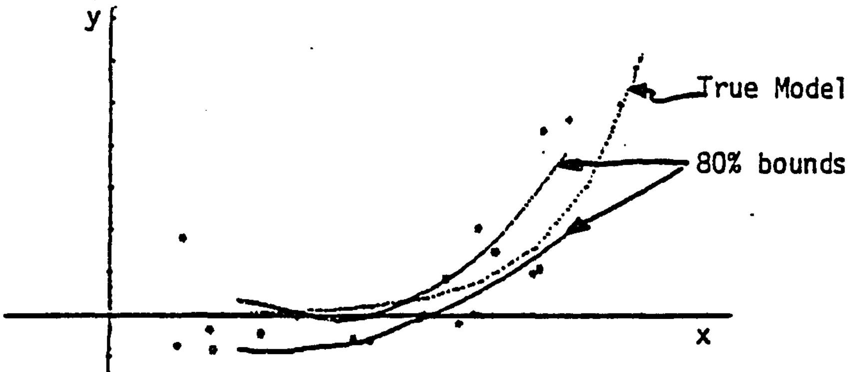


Figure 6.2b Normal Approximation I (Curve)

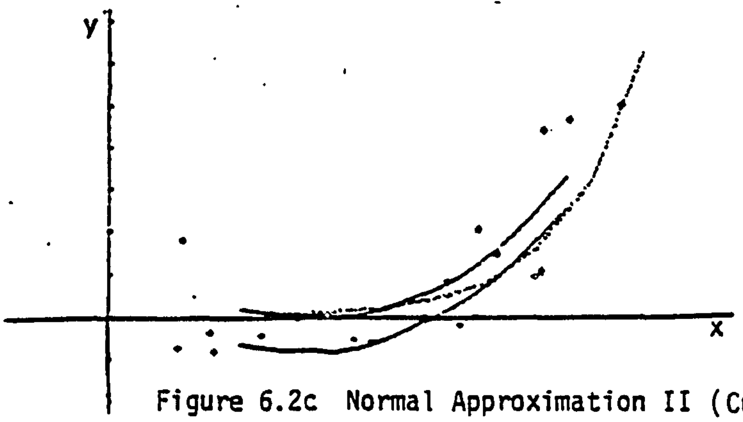


Figure 6.2c Normal Approximation II (Curve)

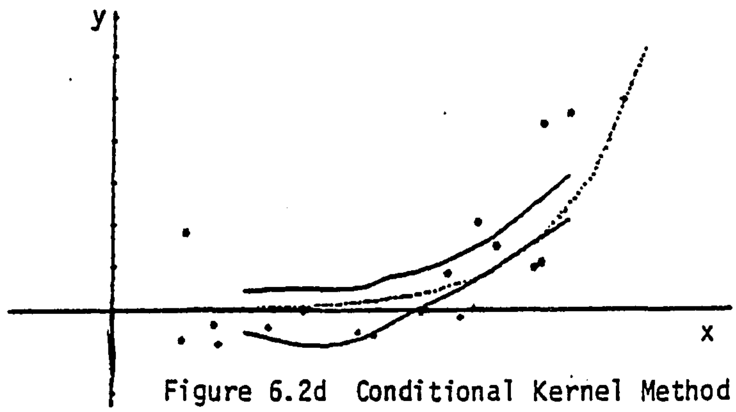


Figure 6.2d Conditional Kernel Method (Curve)

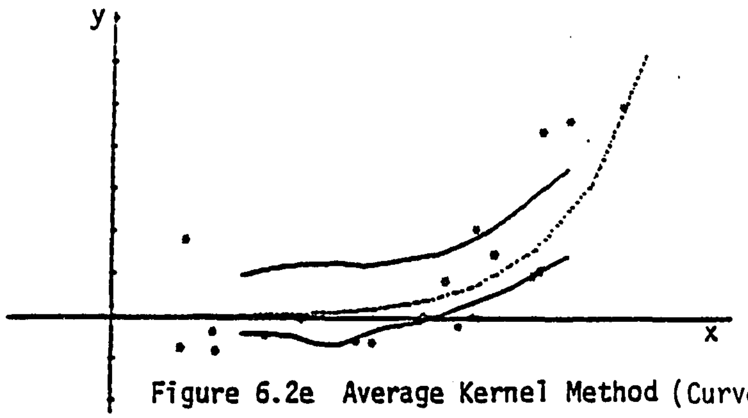


Figure 6.2e Average Kernel Method (Curve)

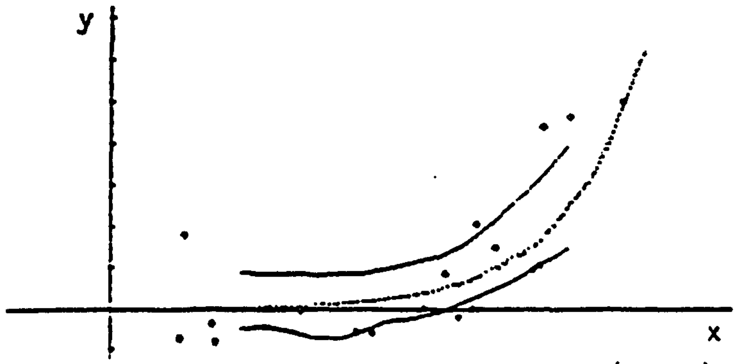


Figure 6.2f Bootstrap Method (Curve)

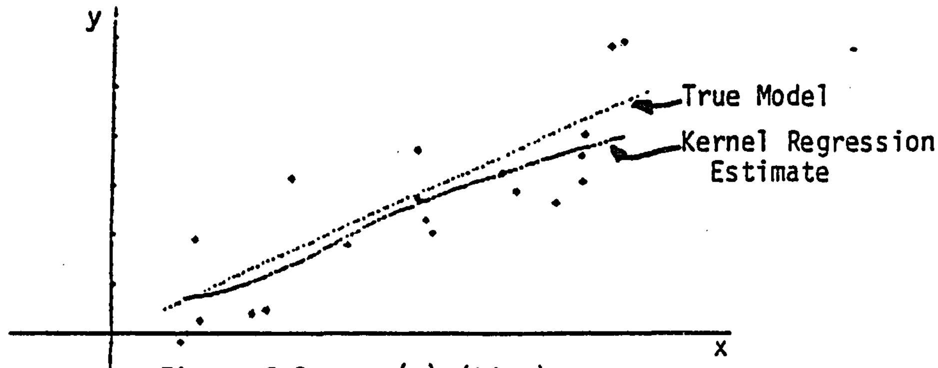


Figure 6.3a $m_n(x)$ (Line)

Figure 6.3b Normal Approximation I (Line)

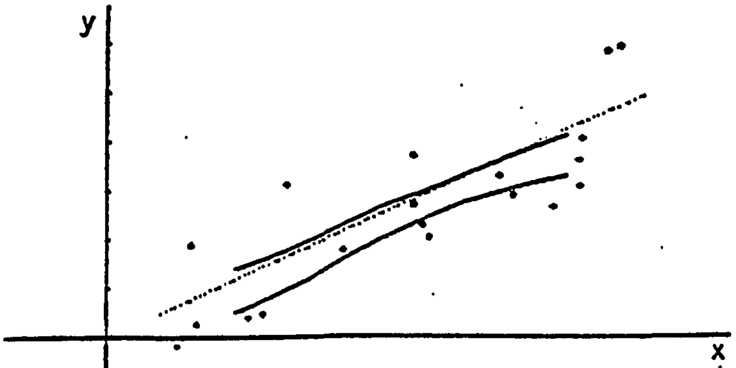


Figure 6.3c Normal Approximation II (Line)

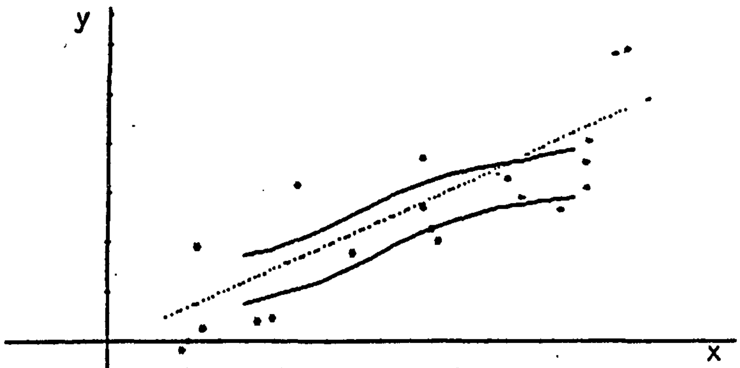


Figure 6.3d Conditional Kernel Method (Line)

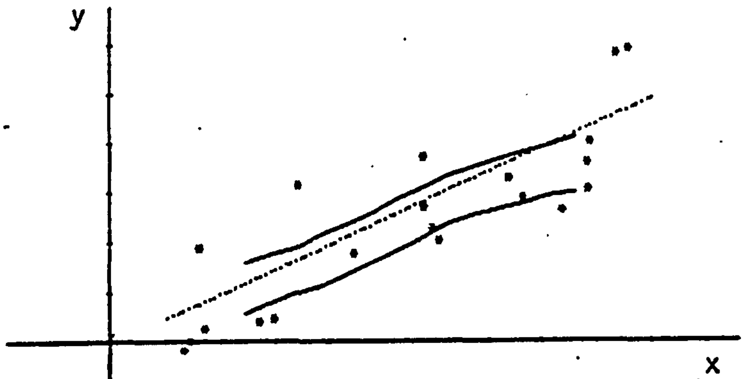


Figure 6.3e Average Kernel Method (Line)

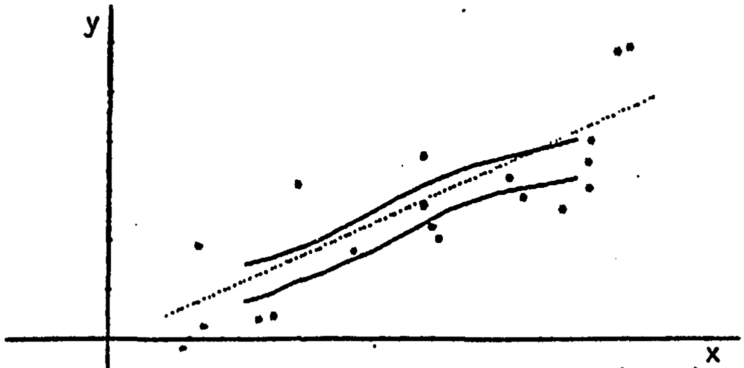


Figure 6.3f Bootstrap Method (Line)

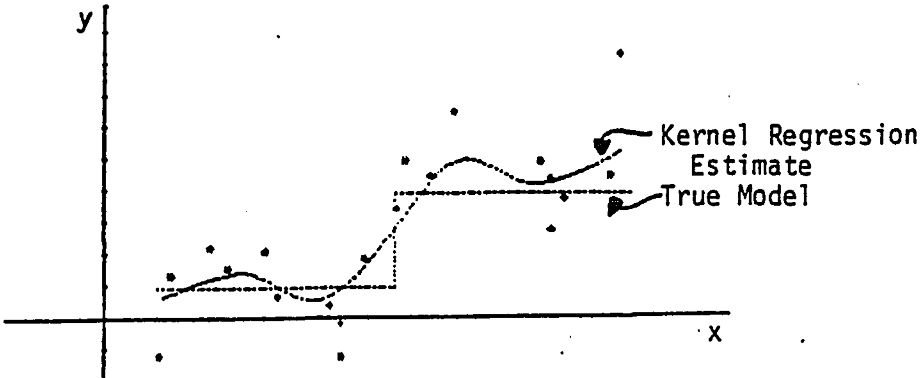


Figure 6.4a $m_n(x)$ (Step)

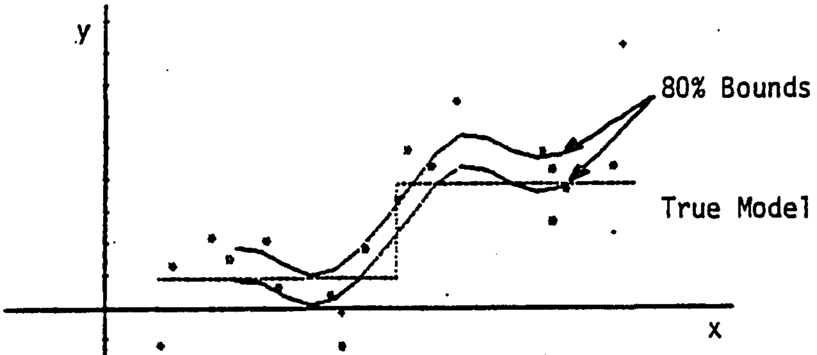


Figure 6.4b Normal Approximation I (Step)

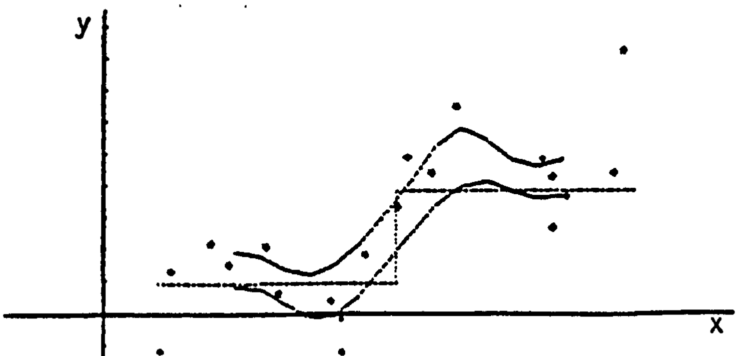


Figure 6.4c Normal Approximation II (Step)

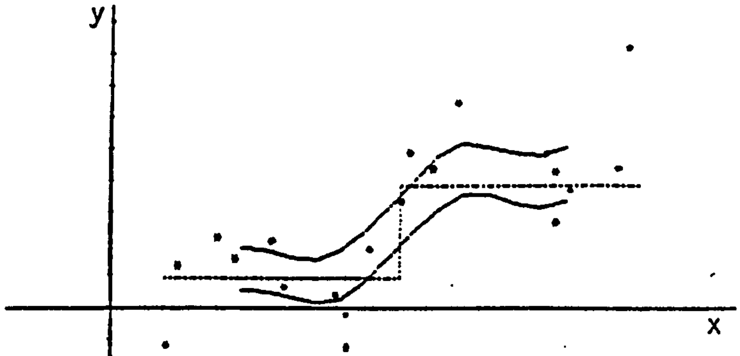


Figure 6.4d Conditional Kernel Method (Step)

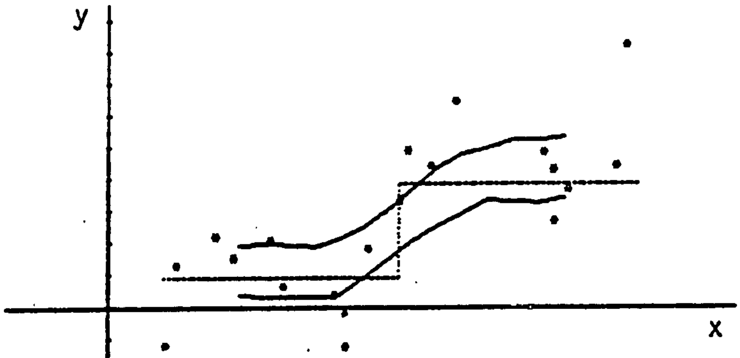


Figure 6.4e Average Kernel Method (Step)

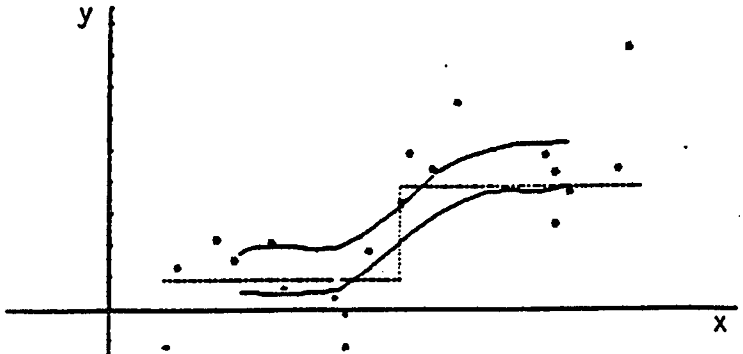


Figure 6.4f Bootstrap Method (Step)

It is difficult from a study of this size to draw any firm conclusions. The summary data of Table 6.2 does indicate that some of the estimators do a reasonable job of constructing confidence bounds.

Following Table 6.3, we summarize the results for each of the estimators. Table 6.3 has been included here to support these summaries. Run by run information used to compute the table has been included as an appendix. Table 6.3 gives percent coverage and bound width information for a study of 6 runs of 40 data points each. Data pairs for the run were generated the same way as for the earlier study. The Y component is exponential with mean 1 and the model is piecewise linear

$$m(x) = 3 + .25(x-5) \quad \text{when} \quad 1 < x \leq 5 \quad \text{and}$$
$$m(x) = 4 + .6(x-5) \quad \text{when} \quad 5 < x \leq 10 \ .$$

Table 6.3 Bound Width and Percent Coverage Summary (40 pts.)

Norm I		Norm II		Meth 1		Meth 2		Meth 3	
%	wth.	%	wth.	%	wth.	%	wth.	%	wth.
.54	.463	.688	.719	.824	.929	.8	.934	.692	.572

We now conclude this section with a brief summary of the performances of the estimators.

Normal approximation I overestimates the coverage probabilities and has the narrowest average bound width of the 5 estimators. At least three factors appear to contribute to this error, each by their influence on the variance component.

First, the residual variance is used to estimate the variance of Y where the residuals are computed as the difference between the data values y_i and the regression approximation $m_n(x_i)$. This is a biased estimate of the variance and we would expect it to be low since the data values y_i have a significant effect on the regression approximation at the point x_i.

A second factor which contributes to this problem is in the denominator of (6.1). The estimate of h(x) made by a kernel density estimate will give low values at the end points and counter balance this with slightly high estimates in the interior (since it is computed to integrate to 1). These slightly high estimates in the denominator also reduce the variance estimate.

The final contributer to this low variance estimate is the quantity $\int \underline{K}^2(u)du$ in the numerator of (6.1). Experimental results indicate that this approximation also leads on average to underestimating the true variance for finite sample sizes.

One would expect the effect of these biases to be reduced as n gets larger. Note that the normal approximation I percent coverage in Table 6.3 is only slightly higher than the overall catagory of Table 6.2. When compared to the exponential catagory of Table 6.2, however, there is a significant improvement (recall that Table 6.3 has been compiled using exponential data).

The normal approximation II using the variance estimate (6.2) provides a closer approximation to the desired coverage probabililty .8. Comparing this estimator to method 1 which follows, we note that the coverage probability is low by .11 but the bound width is slightly narrower also. This method of computing the variance is the one suggested in Yakowitz and Szidarovszky (1985).

Method 1, the conditional kernel method gave a very accurate coverage probability but also the widest bound width of the five estimators. While this estimator seemed to provide good results in both the 20 point and 40 point studies, it was difficult to determine by the results of this study, what if any, the effect of the third moment component discussed in Chapter 5 had. Comparing these results again with normal approximation II in the row for exponential data in Table 6.2 we note that the difference in coverage probability is about the same as in the overall catagory.

Method 2 the average kernel method gave results very similar to method 1. It would seem on a basis of these results that method 1 would be preferable as it is easier to compute and applicable to hetroscedastic data. It is possible, however, that for different model and data situations, the averaging procedure of method 2 would provide an advantage.

Finally method 3 the bootstrap method showed some of the same problems as the normal approximation I. The residual values are biased and hence the confidence coefficient is to large. Turning to Table 6.3, we see a great improvement in the percentage coverage for the 40 point exponential data when compared to the 20 point exponential runs of Table 6.2.

Study 3

This last computer study was conducted to demonstrate the advantages of those estimators which are valid for homoscedastic data. Figures showing the performance of the 5 estimators (in the same order as the last computer study) are given in Figure 6.5. Here a linear model is used and the data is generated from a normal distribution where the variance ranging from .5 at x = 1 to 8 at x = 9. Figure 6.5 shows how only the kernel conditional estimator and the normal approximation II, with locally estimated moments respond to the change in variance. The bound width and percent coverage of the estimators is summarized in the table which follows.

	Percent Coverage	Interval Width
Normal Approx. I	.769	.901
Normal Approx. II	.923	.948
Conditional kernel Est.	1.	1.227
Averaged Kernel Est.	.846	1.3
Bootstrap Est.	.923	1.23

Table 6.4 Coverage and Width Summary Using Hetroscedastic Data

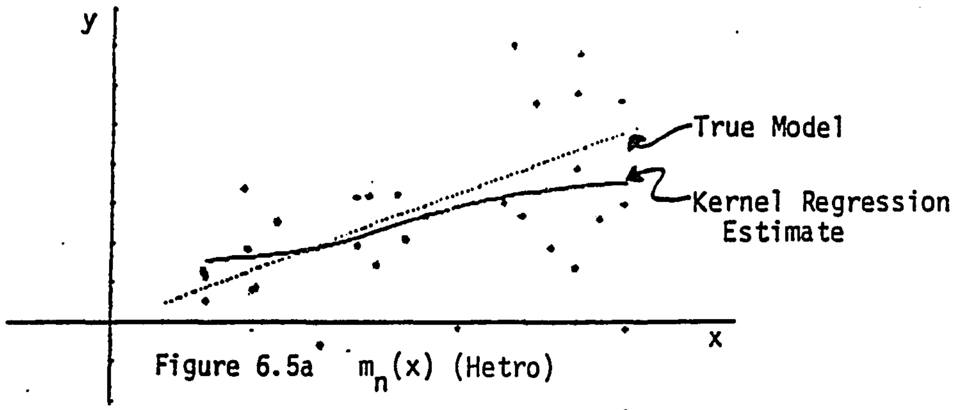


Figure 6.5a $m_n(x)$ (Hetro)

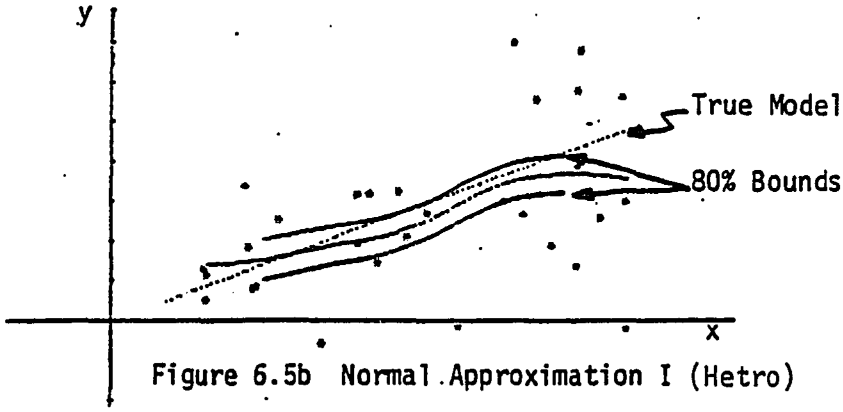


Figure 6.5b Normal Approximation I (Hetro)

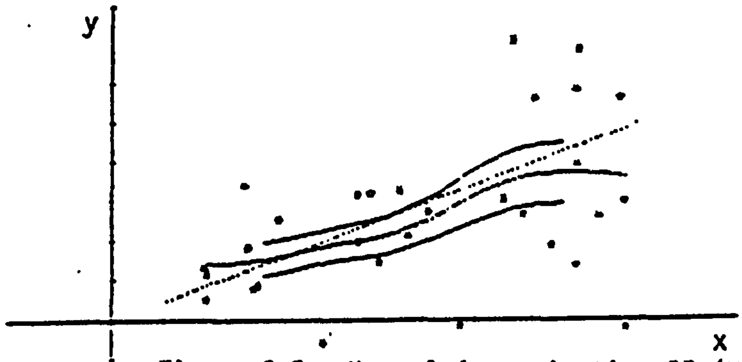


Figure 6.5c Normal Approximation II (Hetro)

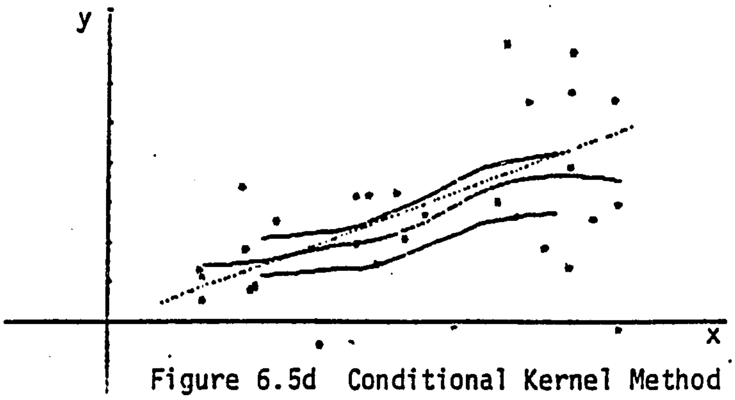


Figure 6.5d Conditional Kernel Method (Hetro)

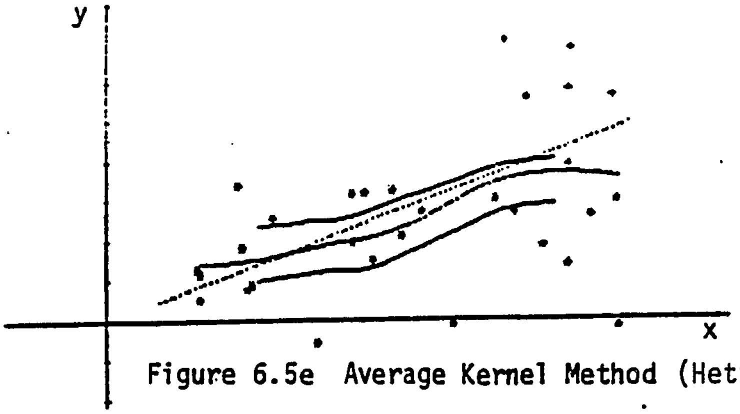


Figure 6.5e Average Kernel Method (Hetro)

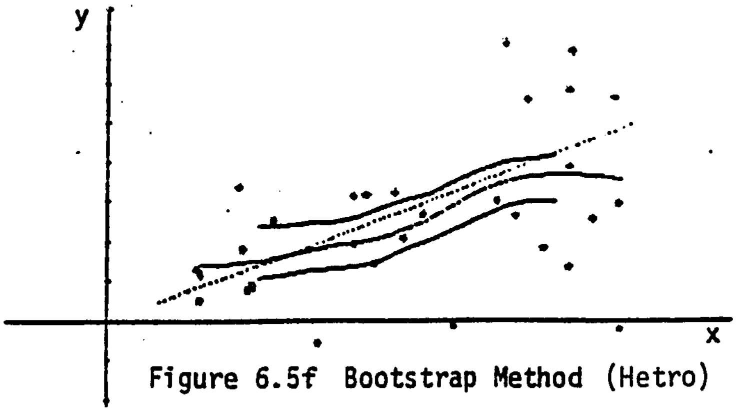


Figure 6.5f Bootstrap Method (Hetro)

APPENDIX

Run by Run Report

Run #	Norm I		Norm II		Meth 1		Meth 2		Meth 3	
	%	wth.	%	wth.	%	wth.	%	wth.	%	wth.
					(20 points)					
1	.429	.671	.714	.963	1.0	1.61	1.0	1.67	.928	1.05
2	.429	.844	.714	1.149	1.0	1.54	1.0	2.01	.786	1.16
3	.214	.733	.5	1.529	1.0	1.56	.857	1.613	.5	.659
4	.643	.609	.786	.943	.857	1.15	.929	1.01	.643	.829
5	.333	.956	.733	1.893	1.0	1.93	1.0	1.87	.60	1.33
6	.938	.831	1.	1.616	1.0	1.80	1.0	1.65	.813	1.08
7	1.0	1.028	1.	1.608	1.0	1.80	1.0	1.57	.80	1.35
8	.286	.73	.571	1.096	1.0	1.20	.929	1.20	.929	1.08
9	.714	1.14	.929	1.804	.929	2.17	.929	2.09	.786	1.62
10	.642	.638	.928	.950	1.0	1.23	.857	1.03	.786	.807
11	.154	.482	.307	.564	.385	.861	.308	.901	.308	.658
12	1.0	.496	1.	.711	1.0	1.09	1.0	.916	.857	.673
13	.467	.448	.733	.665	.733	.867	.733	.832	.80	.619
14	.714	.566	.714	.729	.786	.989	.929	.998	.857	.753
15	.5	.617	.75	.881	.583	1.16	.583	1.03	.667	.859
16	.692	.66	.692	.727	.692	1.13	.692	1.12	.615	1.03

Run #	Norm I		Norm II		Meth 1		Meth 2		Meth 3	
	%	wth.	%	wth.	%	wth.	%	wth.	%	wth.
17	.769	.549	1.	1.04	1.0	1.29	1.0	1.22	1.0	1.03
18	.385	.567	.692	.929	.923	1.08	.923	.833	.615	.799
19	.077	.64	.231	.940	.154	1.10	.077	1.07	.077	.899
20	.6	.549	.667	.821	.333	.843	.533	.996	.533	.690
21	.143	.584	.357	.795	.643	1.24	.50	1.27	.143	.848
22	.462	.587	.692	.895	.846	1.27	.692	1.05	.769	.85
23	.308	.733	.610	1.513	.846	1.61	.923	1.55	.615	.954
24	.667	.5	.667	.614	.80	.982	.733	.762	.733	.689
25	.428	1.038	.571	1.66	.571	1.53	.643	1.46	.643	1.14
26	.461	1.042	.615	2.05	.846	2.25	.769	1.83	.615	1.40
27	.714	1.006	.786	1.41	.857	1.44	.786	1.61	.714	1.25
				(40 points)						
1	.571	.482	.714	.700	1.	.779	1.	.7	1.	.542
2	.286	.542	.357	.662	.429	.910	.357	.923	.286	.688
3	.429	.440	.643	.517	1.	1.31	1.	1.365	.857	.543
4	.6	.423	1.	1.04	.8	.687	.8	.67	.866	.538
5	.857	.421	.769	.712	.857	.752	.857	.731	.643	.532
6	.5	.467	.642	.686	.857	1.13	.786	1.13	.5	.59

REFERENCES

Bickel, P.J. and Freedman, D.A. (1981) "Some Asymptotic Theory for the Bootstrap", Ann. Statis. **9**, 1196-1217.

Cacoullos, T. (1966). "Estimation of a Multivariate Density". Ann. Inst. Statis. Math. **18**, 179-189.

Chung, K. L. (1974) "A Course in Probability Theory", Academic Press Inc., London.

Devroye, L. (1982) "Necessary and Sufficient Conditions for the Pointwise Convergence of Nearest Neighbor Regression Function Estimates", Z. Wahrscheinlichkeitstheorie Verw Gekiete **61**, 467-481.

Efron, B. (1979) "Bootstrap Methods: Another Look at the Jackknife", Ann. Statis. **7**, 1-26.

Efron, B. (1981a) "Nonparametric Estimates of Standard Error: The Jackknife, The Bootstrap and Other Methods", Biometrika 68, **3**, 589-599.

Efron, B. (1981b) "Nonparametric Standard Errors and Confidence Intervals", The Canadian Journal of Statis. **9**, 139-172.

Efron, B. (1982) "The Jackknife the Bootstrap and Other Resampling Plans", J.W. Arrowsmith Ltd. England.

Feller, W. (1966) "An Introduction to Probability Theory and its Applications (Vol II)", John Wiley and Sons Inc., New York.

Freedman, D. A. (1981) "Bootstraping Regression Models", Ann. Statis. **9**, 1218-1228.

Hall, P. (1981) "Laws of the Iterated Logarithm for Nonparametric Density Estimators", Z. Wahrscheinlichkeitstheorie Verw Gekiete **56**, 47-61.

Hardel, W. (1984) "A Law of the Iterated Logarithm for Nonparametric Regression Function Estimators", Ann. Statis. **12**, 624-635.

Johnston, G. (1979) "Smooth Nonparametric Regression Analysis", Inst. Statis. Mimeo Series 1253, University of N. Carolina.

Mack, Y. P. and Silverman, B. W. (1982) "Weak and Strong Uniform Consistency of Kernel Regression Estimates", Z. Wahrscheinlichkeitstheorie Verw Gebiete **61**, 405-415.

Noda, K. (1976). "Estimation of a Regression Function by the Parzen Kernel Type Density Estimators". Ann. Inst. Math. Statis. **28**, 221-234.

Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode". Ann. Math. Statis., **33**, 1065-1076.

Rosenblatt, M. (1956) "Remarks on Some Nonparametric Estimates of a Density Function", Ann. Math. Statis. **27**, 832-837.

Schuster, E. F. (1972) "Joint Asymptotic Distribution of the Estimated Regression Function at a Finite Numer of Discrete Points", Ann. Math. Statis. **43**, 84-88.

Schuster, E. and Yakowitz, S. (1979) "Contributions to the Theory of Nonparametric Regression, with Application to System Identification", Ann. Statis. **7**, 139-149.

Singh, K. (1981) "On the Asymptotic Accuracy of Efron's Bootstrap", Ann. Statis. **9**, 1187-1195.

Stute, W. (1982) "The Law of the Logarithm for Kernel Density Estimators", Ann. Prob. **10**, 414-422.

Van Ryzin, J. (1969) "On Strong Consistency of Density Estimates", Ann. Math. Statis. **40**, 1765-1772.

Yakowitz, S. J. (1985) "Nonparametric Density Estimation, Prediction, and Regression for Markov Sequences", J. Amer. Statis. Assoc. **80**, 215-221.

Yakowitz, S. J. and Szidarovszky, F. (1985) "A Comparison of Kriging with Nonparametric Regression Mehtods", J. Multivariable Anal. **16**, 21-53.

Watson, G. S. (1964). "Smooth Regression Analysis". Sankhyâ A, 26, 359-372.