LOCALLY LINEAR ESTIMATION

OF PARTIALLY LINEAR MODELS

by

Scott Arlen Hamilton

Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2133T

June 1994

LOCALLY LINEAR ESTIMATION OF PARTIALLY LINEAR MODELS

by

Scott Arlen Hamilton

A Dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1994

Approved by:

______________ Advisor

Paul Stewart ______________ Reader

______________ Reader

Scott Arlen Hamilton. Locally Linear Estimation of Partially Linear Models
(Under the direction of Dr. Young K. Truong)

ABSTRACT

In this work we propose using Speckman's (1988) two stage estimation approach of a partially linear model using locally linear smoothing [Fan (1992)]. The issue for estimation of the partially linear model is how to maintain the parametric rate of convergence of the linear estimates and the usual MSE properties of the nonparametric estimate. An important result by Rice (1986) indicated that when the covariates associated with the linear part of the model have a dependence relationship with the covariates associated with the nonparametric function, the parametric rate of convergence may not be obtained by existing methods of estimation. This negative result was addressed by Denby (1986) and Speckman (1988) who independently developed a two stage estimation procedure that achieved the desired outcome for kernel type estimation for fixed designs: root n convergence for the estimates of the linear coefficients and the usual MSE properties for the nonparametric estimate. These results motivate us to propose using Speckman's (1988) two stage estimation for random designs using locally linear smoothing. This approach had not been examined in the literature. Specifically, what had been done before required the assumption of a fixed design for the nonparametric function, where results only addressed MSE properties of estimating a univariate function, and no examination of locally linear estimation.

We find that using our proposed estimation method preserves the desireable results of Speckman's two stage approach as well as achieves the desireable MSE properties associated with locally linear estimation. It is well known within the field of nonparametric estimation that asymptotic results may not be relevant in

finite sample estimation. This is highly dependent upon the context. In many cases, it is instructive to examine the finite sample behavior of a proposed method. Since the sample size of our motivating example is 350 observations, we thought it necessary to examine the finite sample behavior of the proposed estimates of the model. The results indicate that the finite sample performance of the estimators is better than the asymptotic approximation would indicate for each sample size examined. We illustrate the proposed method by examination of the relationship between cholesterol and compliance to drug therapy in the LRC data.

ACKNOWLEDGEMENTS

I wish to express gratitude to my wife, Sharon Adele Dole, for her support, encouragement and inspiration during my graduate study. I thank my advisor, Professor Young K. Truong, for his guidance and especially for the support he provided in the final six months. I give a special thanks to Professor Stephen J. Marron for his encouragement and mentoring. I thank Professors Paul W. Stewart and C. M. Suchindran for their support. And to Professors Dana Quade and P.K. Sen for their advice and comments I give thanks. I wish to thank the students of the department of biostatistics, especially Benoit Masse, for their moral support. My parents deserve thanks for always being there when needed.

This research was supported in part by NIH grant number 5T32HD07237-10.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION TO THE PROBLEM AND THE MODEL

1.1 The Data

The Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) (1984) was a placebo-controlled double blind randomized clinical trial concerning the efficacy of cholestyramine for lowering cholesterol level and thereby reducing coronary heart disease. The patients were men aged 35 to 59 years with high initial cholesterol levels (total plasma cholesterol level greater than 265). The accrual phase consisted of four monthly measurements, then at the fifth visit patients were randomized to either the drug or a placebo. Thereafter, cholesterol measurements were taken every two months for an average of 7.4 years. All patients were advised on a cholesterol lowering diet. Baseline cholesterol was calculated as the average cholesterol of the third, fourth, and fifth visits. The specified dose was six 4-gram packets of cholestyramine daily. Cholestyramine is a resin that is unpleasant to swallow (gritty if swallowed promptly after mixing with water, slimy if swallowed later) and in many cases causes patients to experience intestinal cramps. Thus, many of the subjects of the LRC study were not able to adhere to the drug protocol 100%. In fact, the specified dose was reduced for some patients who could not tolerate the full dose. Others simply did not ingest all of the packets they were given.

Consequently, another variable called compliance was measured as the amount of unused drug returned at each visit. Figure 1 shows kernel density estimates

[Silverman (1986)] for the distribution of compliance over the entire study for the drug (solid line) and placebo (dashed line) groups. Also shown is median compliance at each visit for the drug = "-" and placebo = "+" groups. The bimodal kernel density estimates indicated that most subjects adhered to at least 80% of the specified dose, a smaller number had essentially zero compliance, with the remainder evenly spread between 20% and 80% compliance. The medians at each visit indicated a high level of compliance for both groups, which is a credit to the study management, and indicated that the randomization scheme effectively controlled for non-measurable factors that affected level of compliance.

The fact that cholestyramine could substantially reduce cholesterol if taken regularly was established before the LRC-CPPT (1984) study. A nice review and analysis of cholestyramine effectiveness using the compliance covariate is given in Efron and Feldman (1991) and the attached discussion. Therefore, the important issue here is whether high risk individuals would adhere to the drug regimen. In other words, given that cholestyramine reduces cholesterol, is it a *useful* drug to prescribe to high risk patients who vary in their adherence to the regimen?

For instance, consider the typical scenario where a physician regularly sees patients at risk for coronary heart disease and must make decisions about prescribing cholestyramine for elevated cholesterol. Typically, it can be expected that some patients will not comply with the full drug regimen. The physician would like to know if a reasonable reduction in cholesterol could still be achieved, or if alternative treatment should be sought. Or in some cases, after initially taking the prescribed dose the patient will begin to experience intolerable side effects and consequently reduce the dose ingested. The physician would like to know the predicted reduction in this patient's cholesterol given the reduced level of compliance. On the other hand, because of baseline characteristics some patients may not ingest the full dose at the onset, but after getting accustomed to the side effects they increase the dose ingested to the full dose prescribed.

All of these scenarios occurred in the LRC study. Thus by building a statistical model we can hopefully provide insight into the answers to our researach questions through analysis of the LRC data.

1.2 Development of the Model

We considered the data from the treatment arms of the Minnesota and Stanford sites for yearly visits, where visit 1 indicated baseline, up to year 9 because thereafter a large number of study participants dropped out due to myocardial infarction or death due to other causes.

Without priori knowledge of the pharmacokinetic relationship between cholesterol change and amount of cholestyramine ingested, we proceeded to narrow the field of possible models to ones for which the assumed structure was supported by the data. A plot of the trend in mean cholesterol over visit for the treatment and control groups from Grady (1992) showed the following. The mean cholesterol level for the treatment group lowered between visit 1 and 2, then increased mildly between 2 and 3, and remained flat thereafter. This behavior suggested that the mean effect of the drug occured within the first two years. Therefore, to capture the full effect of the drug with respect to compliance we examined the cholesterol outcome at visit 3.

To address our clinical questions regarding the relationship between cholesterol and compliance we constructed a regression model with total cholesterol at visit 3 as the outcome and compliance at visit 1 and 2 as predictors. Because baseline cholesterol is predictive of the total cholesterol trend, we controlled for its effect by also including it as a predictor. Hence, we defined the regression function of interest as

$$E(Cholesterol|Baseline\_Chol, Compliance). \tag{1.1}$$

A priori, we wished to specify and estimate (1.1) in such a manner as to minimize model bias. Nonparametric regression and its effect on model bias is discussed below. We started with graphical tools to get familiar with the structure of the data. To begin, we had to consider the issue of specification of a mathematical form to represent the regression function. Without prior pharmacokinetic knowledge of how the absorption of cholestyramine changes serum cholesterol we were left with an arbitrary choice, in which case a linear model is a typical starting point.

Figure 2 displays a scatterplot of compliance versus cholesterol after controlling for baseline, i.e. the points plotted on the vertical axis are the residuals of a linear model for cholesterol given baseline. Also depicted is a nonparametric smooth of each scatterplot. At first glance one may suggest that a line segment with a decreasing slope could be a reasonable approximation for parametrically describing the mean level of cholesterol as a function of compliance. However, considering the bimodal distribution of compliance (Figure 1) where mass is piled up at the extreme values of the covariate [0, 100] the slope of the line would be heavily determined by the cholesterol means at the extremes and behavior in-between would be ignored. The cost of this is illustrated in the subtle behavior observed in the nonparametric smooth. For instance, we observe the curve to be rather flat until approximately 65%-70% compliance, then it changes slope which could indicate that one should expect a certain level of compliance before a reduction in cholesterol occurs. This potentially useful information would be lost by approximating the regression function with a line segment or a quadratic curve.

The differences between the nonparametric regression estimate of (1.1) and using linear regression can be described in terms of model bias. Model bias occurs when the wrong mathematical form is specified to represent the conditional mean. Since almost no relationship between two biological variables occurring in nature is truly linear, linear regression almost always involves some amount of model bias. Since the model bias is unknown, it could be severe or negligible without any way to quantify it; however, some speculation of small model bias can sometimes be empirically justified with the data. In contrast, nonparametric regression involves no model bias in the following sense. It has been shown that the nonparametric estimate converges to the true conditional mean as the sample size increases and the bandwidth decreases (explained in Section 1.2.3).

Therefore, using nonparametric regression we are directly estimating the conditional mean of cholesterol as a function of compliance, thus the interpretation of the curve estimate is clear. Whereas, when a parametric function is specified for the conditional mean we interpret the curve estimate as an approximation of something which may or may not be an accurate form for the conditional mean. As shown by the compliance data, assuming a linear form for the regression function can force the estimate to mask possibly important patterns in the data. Whereas, nonparametric regression provides a more flexible estimate of the regression function (1.1) that is more sensitive to subtle fluctuations in the data. For these reasons, we believe that the flexibility and the low model bias of the nonparametric regression estimate gives us deeper insight into the behavior of the variables under study than specifying a parametric form for the conditional mean.

We have particular interest in the patterns of the curves near the boundaries of compliance, [0, 100], which means we want a kernel estimator of the conditional mean that is not adversely affected by estimation near the boundary. It is known that the bias or variance of the Nadaraya-Watson and Gasser-Müller kernel estimators are affected by estimation at the boundary, unless special kernels are used [Härdle (1989)]. In contrast, Fan (1992) demonstrated that the conditional bias and variance of the local linear estimate near the boundaries are of the same order of magnitude as in the interior [Fan and Gijbels (1992)].

The study of kernel regression has received a great deal of attention and refinement in the past fifteen years. It has been shown that in problems of higher dimension the kernel estimate of the regression function is vulnerable to slow convergence. This is explained in more detail in the section covering rates of convergence. One attempt to alleviate this problem is the partially linear model proposed in the mid eighties, which reduces the dimension of the nonparametric estimation by the number of linear coefficients added to the regression function. The question remains how to control for the effect of baseline cholesterol in the regression function after the compliance covariates are in the model. Figure 3 displays a scatter plot of cholesterol after adjustment for compliance (subracting off the locally linear estimate of regressing cholesterol on compliance alone) at visit one and two against baseline cholesterol also adjusted for the compliance covariates. Also shown is a locally linear smooth of the residuals, which strongly suggests that the relationship between baseline and cholesterol after controlling for compliance is roughly linear. Only in the high range of baseline representing less than 1% of the data, the curve strays from linearity. This suggests that cholesterol could be modeled adequately as a linear function of baseline, given information about compliance. Therefore, denoting cholesterol as Y, baseline as B, compliance at visit 1 as X_1, and compliance at 2 as X_2 we proposed the partially linear model

$$E(Y_i|B_i = b,\ X_{i1} = x_1,\ X_{i2} = x_2) = \alpha b + m(x_1,x_2)\ , \qquad i = 1,\ldots,n, \tag{1.2}$$

for the analysis of the LRC data.

Since the proposed method of estimation of (1.2) uses recent developments in the literature, the next chapter reviews kernel regression and the development of the partially linear model.

CHAPTER II: EXISTING METHODOLOGY

In this chapter, we review some of the major methods of nonparametric regression. This overview is provided to give the reader an understanding of the mechanics of smoothing and how the methods use the data to estimate the regression curve. Although several major methods exist, the question remains about which method to use. Among the methods of nonparametric regression, there is no clear choice of a superior method. Moreover, Silverman (1984) demonstrated the conditions for which the kernel and smoothing spline methods are equivalent, and Härdle (1989) gives the conditions under which kernel and k-nearest neighbor methods are equivalent. Additionally, all of the estimates obtained by the major nonparametric methods achieve the optimal rate of convergence for nonparametric regression. Therefore the choice becomes a consideration of several factors: Interpretability, assumptions about the marginal density, $f(x)$, the regression function, $m(x)$, the kernel function, $K(\cdot)$, behavior of the estimate near the boundary of the support of f, automated smoothing parameter selection and others. We shall consider some of these factors in our choice of method.

Smoothing splines is a popular approach to nonparametric regression which has a large amount of literature to support the methodology, for an overview see Eubank (1988). Another body of literature is devoted to the method of orthogonal series estimators which have mainly been applied in density estimation, for an overview see Härdle (1989). We have focused our discussion on the kernel and k-nearest neighbor methods for several reasons: the rich body of literature surrounding them, the desirable theoretical properties, and the ease of implementation and interpretation. The following paragraphs review the kernel and k-nearest neighbor methods. For ease of notation, the summary of these methods is focused on univariate regression in Sections 2.1 and 2.2.

2.1 Kernel Smoothing

In the one-dimensional case, kernel smoothing provides a conceptually simple approach to nonparametric regression. Additionally, the kernel regression estimator has computational advantages and optimal sampling properties. A sequence of kernel weights represented by $\{W_{ni}(x)\}_1^n$ is described by a density function with a scale parameter that adjusts the size and the form of the weights near x. This density function is referred to as the kernel function, $K(\cdot)$, which has the properties

$$\int K(u)du = 1,$$

$$\int K^2(u)du < \infty , \tag{2.1}$$

$$K(\cdot) \text{ is continuous and } \geq 0 .$$

The weight function $W_{ni}(x)$ takes the form

$$W_{ni}(x) = n^{-1}K_h(x - X_i)/\hat{f}_h(x) , \tag{2.2}$$

where the normalization of the weights by

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) , \tag{2.3}$$

the Rosenblatt-Parzen kernel density estimate of the marginal density of x [Rosenblatt (1956); Parzen (1962)] makes it possible to adapt to the local intensity of the x variables and assures the weights sum to 1, and where

$$K_h(u) = h^{-1} K(u/h) \ , \tag{2.4}$$

with h the scale factor. The weight function (2.2) was proposed by Nadaraya (1964) and Watson (1964) and therefore

$$\hat{m}_h(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i \ , \tag{2.5}$$

is referred to as the Nadaraya-Watson estimator. The many shapes that $W_{ni}(x)$ can take depend upon the choice of K and the smoothness of the regression curve depends upon the size of h, the bandwidth, and the chosen kernel function. Since one can achieve very similar results using different kernels simply by adjusting the bandwidth, choice of kernel is flexible.

Since the smoothness of the regression curve estimate depends on the chosen bandwidth, h, a useful way to understand the kernel estimator is to examine what happens when $h \to 0$ and when $h \to \infty$. As the bandwidth, h, shrinks to zero eventually there will be no observations in the window around x, and the estimate will be undefined. Therefore, let us for the moment suppose we are estimating m at X_i. Then as $h \to 0$, $\hat{m}_h(X_i) \to K(0)Y_i/K(0) = Y_i$. Thus, as the bandwidth shrinks, i.e. the smoothness decreases, the estimate reproduces the data points. Now, as $h \to \infty$, eventually the window of estimation for each x will include all of the data points. Also, as $h \to \infty$, $K\left(\frac{x - X_i}{h}\right) \to K(0)$; therefore,

$$\widehat{m}_h(x) \to \left[n^{-1}\sum_{i=1}^{n} K(0)Y_i\right]\left[n^{-1}\sum_{i=1}^{n} K(0)\right]^{-1} = n^{-1}\sum_{i=1}^{n} Y_i.$$

This shows that as the bandwidth gets large, i.e. the smoothness increases, the estimate becomes the mean of all the data. With this, one can get an idea of the trade-off between bias and variance. A very small bandwidth will create a rough curve that is highly variable. A large bandwidth will create a smooth curve that has low variance. The task is to find the bandwidth that will minimize the bias and the variance in such a way as to achieve optimal convergence in large samples to the true regression curve. Optimal convergence is discussed in Section 2.3.

The construction of the kernel estimate points out an important feature. The kernel estimate of $E(Y \mid X = x)$ includes information from Y_i's associated with X_i's within a neighborhood of x. Since the Y_i values in the neighborhood of x have mean close to but not equal to $E(Y \mid X = x)$, the estimate is biased. Therefore, the variance alone does not fully quantify the convergence of $\widehat{m}_h(x)$ to the true curve, $m(x)$. As a result, both the bias and the variance are quantified in order to examine asymptotic convergence of the estimate, and the convergence of a nonparametric estimate is measured in terms of mean squared error. For instance, the pointwise mean squared error is measured as

$$E[\widehat{m}_h(x) - m(x)]^2. \tag{2.6}$$

The form of the bias and variance are important features that characterize the advantages of different kernel estimators proposed in the literature. Since we choose to examine the general case where the X_i's are random, we will compare the bias and variance of several estimators for the random design. We examine the Nadaraya-Watson kernel estimator because of its historical significance and prevalence in the literature, the estimator such as that proposed by Gasser and Müller (1979), since it has received much interest [Hart and Wehrly (1986)], and the kernel weighted locally linear estimate of Fan (1992), because of its excellent MSE properties. Under the conditions that $h \to 0$ and $nh \to \infty$, the Nadaraya-Watson kernel estimate, from the weights defined in (2.2), has asymptotic

$$bias = h^2 \frac{(m''f + 2m'f')(x)}{2f(x)} d_K, \text{ and } variance = \frac{\sigma^2(x)}{nhf(x)} c_K \,, \tag{2.7}$$

where

$$d_K = \int u^2 K(u) du \ , c_K = \int K^2(u) du \ ,$$

$$\text{and } \sigma^2(x) = \int (y - m(x))^2 f(y \mid x) dy.$$

[Hārdle (1989)].

Gasser and Müller (1979) proposed the weight sequence

$$W_{ni}(x) = n \int_{S_{i-1}}^{S_i} K_h(x-u) du, \tag{2.8}$$

where $X_{i-1} \leq S_{i-1} \leq X_i$ is chosen between the ordered data. The weights from (2.8) are then plugged into (2.5) to obtain the Gasser-Müller estimate $\widehat{m}_h(\cdot)$ of $m(\cdot)$. For the random design case, the Gasser-Müller estimate has asymptotic

$$bias = \frac{h^2}{2} m''(x) \, d_K \text{ and } variance = \frac{2\sigma^2(x)}{nhf(x)} c_K \,, \tag{2.9}$$

where d_K and c_K are as defined before [Härdle (1989)]. Fan (1992) examined a kernel estimator similar to the one discussed in Stone (1982), which is a kernel weighted locally linear estimate. Fan (1992) defined the estimate of $m(x)$ as $\widehat{a}$, where $[\widehat{a},\widehat{b}]$ minimize

$$\sum_{i=1}^{n}\left(Y_i - a - b(X_i - x)\right)^2 K\left(\frac{X_i - x}{h}\right) . \tag{2.10}$$

For the random design case, the kernel weighted locally linear estimate has asymptotic

$$bias = \frac{h^2}{2}\, m''(x)\, d_K \text{ and } variance = \frac{\sigma^2(x)}{nhf(x)}\, c_K\ , \tag{2.11}$$

where d_K and c_K are as defined before [Fan (1992)]. Ruppert and Wand (1994) extended these results of Fan (1992) to the d-dimensional multiple predictor case. Let the pairs $[(Y_1, \boldsymbol{X}_1),\ldots,(Y_n, \boldsymbol{X}_n)]$ be a random sample where the Y_i are scalar response variables and the $\boldsymbol{X}_i$ are $\Re^d$ valued covariate vectors having common density f with support in $\Re^d$. In their development, $\boldsymbol{H}$ is a d by d symmetric positive definite matrix depending on n called *the bandwidth matrix*, K is a bivariate kernel such that $\int K(\boldsymbol{u})d\boldsymbol{u} = 1$, and $K_H(\boldsymbol{u}) = |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}\boldsymbol{u})$. They assume that all odd order moments of K vanish, that is, $\int u_1^l u_2^k K(\boldsymbol{u})d\boldsymbol{u} = 0$ for all non-negative integers l and k such that their sum is odd. Also, $\int \boldsymbol{u}\boldsymbol{u}' K(\boldsymbol{u})d\boldsymbol{u} = \mu_2(K)\boldsymbol{I}_d$ where $\int u^2 K(\boldsymbol{u})du \neq 0$ independent of i and $\boldsymbol{I}_d$ is the d by d identity matrix. These conditions are satisfied by spherically symmetric kernels and product kernels based on symmetric univariate kernels, provided they have all moments existing. Finally, they assume that K is bounded. Ruppert and Wand (1994) found using locally linear smoothing for a d dimensional covariate vector the asymptotic

$$bias = \tfrac{1}{2}\mu_2(K)\mathrm{tr}\Big(\boldsymbol{H}\Gamma_m(\boldsymbol{x})\Big) \tag{2.12a}$$

and

$$variance = n^{-1}f^{-1}(\boldsymbol{x})\sigma^2(\boldsymbol{x})|\boldsymbol{H}|^{-1/2}\int K^2(\boldsymbol{v})d\boldsymbol{v}, \tag{2.12b}$$

where $\Gamma_m(\boldsymbol{x})$ is the d dimensional Hessian matrix with respect to the function $m(\cdot)$. The asymptotic conditional bias term, (2.12a) has an intuitively simple interpretation. The bias term is a sum of the measure of curvature in each direction multiplied by the amount of smoothing in that direction. Referring to the bandwidth matrices defined in Chapter III, when using $\boldsymbol{H}_1$ we smooth the same amount in the vertical and horizontal directions, using $\boldsymbol{H}_2$ we smooth by different amounts in the vertical and horizontal directions, and using $\boldsymbol{H}_3$ we allow for different amounts of smoothing in each direction on rotated axes. Thus, the intuitive idea of bias being increased when there is more curvature and more smoothing is easily seen in the asymptotic conditional bias terms.

Examination of (2.7), (2.12) and (2.11) gives some insight into the relative advantages and disadvantages of each procedure. When considering the Nadaraya-Watson kernel estimator, an additional assumption must be made of the existence of the first derivative of the marginal density, $f'(x)$, to calculate the bias. Also, the bias is a complicated function of $m''(x)$, $m'(x)$, $f'(x)$ and $f(x)$. In comparison, the bias of the Gasser-Müller estimator is only a function of $m''(x)$. It is more appealing to have the bias only a function of $m''(x)$ since we are estimating $m(x)$. It is not desireable to have the bias of our estimate of $m(x)$ in some way a function of the underlying design density. When considering the locally linear estimate we see that it has the appealing bias of the Gasser-Müller estimate, and the smaller variance of the Nadaraya-Watson estimate. When considering a general class of linear estimators, which includes kernel and spline estimators, Fan (1992) showed that the locally linear estimator reached the best possible rate of convergence of all estimators in this class, unlike the Gasser-Müller or Nadaraya-Watson estimators.

It is also interesting to compare the properties of these estimators at the boundaries of the support of f. Fan and Gijbels (1992) demonstrated in the univariate case that unlike the Nadarya-Watson and Gasser-Müller estimates, the kernel weighted locally linear estimate is not affected by estimation at the boundaries; further discussion of boundary effects is beyond the scope of this text, but for more details see Fan and Gijbels (1992). The conclusion drawn from the above arguments is that the best choice among the major existing kernel estimators is the kernel weighted locally linear estimate. In the next Section we consider the k-nearest neighbor estimate.

2.2 *K*-Nearest Neighbor Smoothing

The k-nearest neighbor estimate is in many ways similar to the kernel estimate. The neighborhood of points used in the kernel estimate was defined as those points falling within a fixed distance of x; whereas, the neighborhood of points used in the k-nearest neighbor $(k-NN)$ estimate is varying. Speaking in terms of Euclidian distance, the $k-NN$ neighborhood is the X_i's that are among the k closest observations to x. For the purpose of estimating the regression function, a uniform weight sequence $\{W_{ni}(x)\}_1^n$ is defined as

$$W_{ni}(x) = \begin{cases} 1/k, & \text{if } X_i \text{ is among the } k \text{ closest observations to } x\ , \\ 0 & \text{otherwise.} \end{cases} \tag{2.13}$$

The k-NN estimate, $\widehat{m}_k(x)$, of $m(x)$ is then constructed by plugging (2.13) into

(2.11). The smoothness of the curve estimate, $\widehat{m}_k(x)$, depends on the choice of k. For the $k-NN$ estimator, letting k grow larger than n is equivalent to letting h grow large in the kernel estimate which results in the mean of all the observations. Similarly, letting k go to 1 for the $k-NN$ estimator is equivalent to letting h go to 0 for the kernel estimator which results in reproducing the observations at each X_i with a jump between the observations.

Another feature that highlights the similarity of $k-NN$ and kernel estimation is to consider the case where the X values fall in an equidistant grid and let $k=2h$ in (2.13). In this case, the $k-NN$ weights are equivalent to the kernel weights when using a uniform kernel. To confirm the similarity in performance of the $k-NN$ and kernel estimators, we examine the asymptotic properties of the $k-NN$ estimate. Härdle (1989) states that Lai (1977) demonstrated the asymptotic variance of the $k-NN$ estimate when using weights (2.13). Under the conditions that $k \to \infty$ and $k/n \to 0$, the $k-NN$ estimate, $\widehat{m}_k(x)$, using weights (2.13) has asymptotic

$$bias = \left(\frac{k}{n}\right)^2 \frac{(m''f + 2m'f')(x)}{24 f^3(x)} \text{ and } variance = \frac{\sigma^2(x)}{k} \quad . \tag{2.14}$$

Naturally, other weight sequences for $k-NN$ estimation have been proposed. As described in Härdle (1989), Stone (1977) examined uniform, triangular, and quadratic weights. These weight sequences can be viewed in general using a kernel function, K,

$$W_{Di}(x) = \frac{K(x-X_i)}{\widehat{f}_D(x)} \quad , \tag{2.15}$$

where

$$\widehat{f}_D(x) = n^{-1} \sum_{i=1}^{n} K_D(x-X_i) \tag{2.16}$$

is a kernel density estimate of $f(x)$ with kernel sequence

$$K_D(u) = D^{-1}K(u/D) \quad , \tag{2.17}$$

and D is the Euclidian distance between x and its kth nearest neighbor. Looking at the $k-NN$ weights as (2.15), Härdle (1989) gives credit to Mack (1981) who calculated their asymptotic

$$bias = \left(\frac{k}{n}\right)^2 \frac{(m''f + 2m'f')(x)}{8f^3(x)} d_K \text{ and } variance = \frac{2\sigma^2(x)}{k} c_K \quad , \tag{2.18}$$

where d_K and c_K are defined as before. By letting $k = 2nhf(x)$, then other than a constant, the bias and variance of the $k-NN$ estimate are equivalent to that of the Nadaraya-Watson kernel estimate.

Looking at the form of the bias for the $k-NN$ estimate we notice several features. Firstly, the bias is a function of $f'(x)$, causing the necessity of extra smoothness conditions to be put on the marginal density, $f(x)$. Secondly, the cubed marginal density in the denominator causes the estimate to have larger bias than the kernel estimate in the support of the marginal distribution where the density is small, usually in the tails, whereas the variance of the kernel estimate is a function of $f^{-1}(x)$, causing it to be larger than the variance of the $k-NN$ estimate in areas where $f(x)$ is small. Thus, again we are faced with a trade-off between bias and variance when choosing between the kernel or $k-NN$ estimators.

By setting $k = 2nhf(x)$ in (2.18), we can directly compare the asymptotic bias and variance of the $k-NN$ estimate with that of the kernel estimates. In this case, other than a constant, the asymptotic bias and variance of the $k-NN$ estimate is equivalent to that of the Nadaraya-Watson kernel estimate. Recall that the conclusion of our previous discussion of the kernel estimates was that the kernel weighted locally linear estimator was preferred. Hence, the logical conclusion here is that the locally linear estimate is also preferred to the $k-NN$ estimator, based on comparisons of bias and variance.

2.3 Rates of Convergence

Based on the results given in Section 2.1 and 2.2, we can see that if the bandwidth (or the smoothing parameter, k, of the $k-NN$ estimate) is chosen as a function of the sample size, n, then all of the nonparametric estimates discussed converge to the true regression curve as n tends to infinity. In statistics, however, this is not the only measure of the performance of an estimator. We are also interested in the speed at which an estimate converges. In parametric regression, the form of the true regression curve is specified; therefore, we measure the speed at which the parameter estimates converge to the specified parameters. This is quantified as the speed at which the sample variance of the parameter estimate tends to zero, which is usually the square root of the sample size, n. However, if one considers how well the parametric specification resembles the true regression curve, one is left without an answer. One could claim that if the true form of the regression curve is unknown, then the parametric estimate of the regression curve will never converge to the true regression function.

One of the advantages of nonparametric regression is the ability to quantify how close the estimate is to the true regression curve and the rate at which it converges. This involves quantifying the convergence rate of the bias as well as the variance, or usually the summary statistic, pointwise mean squared error:

$$E[\widehat{m}(x) - m(x)]^2, \tag{2.19}$$

which equals the squared bias plus the variance. The conditional MSE, (2.19), only gives us the convergence properties at a single point, x. To measure the speed and accuracy of the entire regression curve estimate we must integrate over the support of f, which produces the mean integrated squared error (MISE)

$$\int \left(\widehat{m}(x) - m(x)\right)^2 f(x)dx \, . \tag{2.20}$$

Other measurements of distance for nonparametric estimates have been studied, for a review see Härdle (1989).

To illustrate how the speed of convergence of the MISE is measured, we will examine that of the kernel weighted locally linear estimator. In order to express the MISE in terms of the sample size, n, we must choose a bandwidth as a function of the sample size which minimizes the MISE. By using the expressions of asymptotic bias and variance we can find an approximate expression for the MISE

$$\tfrac{h^4}{4} d_K^2 \int \left(m''(x)\right)^2 dx + \tfrac{c_K}{nh} \int \frac{\sigma^2(x)}{f(x)} dx \quad , \tag{2.21}$$

where d_K, c_k, and $\sigma^2(x)$ are defined as before. By using simple calculus, we can show that from (2.21)

$$h_{opt} = \left[\frac{c_k \int \sigma^2(x)/f(x)dx}{d_k^2 \int \left(m''(x)\right)^2 dx} \right]^{1/5} n^{-1/5} \quad . \tag{2.22}$$

At this point, we observe that the optimal bandwidth itself will converge to zero as the sample size increases, at the rate of $n^{-1/5}$. Also, curves that have a large conditional variance, $\sigma^2(x)$, will have a larger optimal bandwidth to smooth out the roughness caused by higher variation. Substituting the form of h_{opt} from (2.22) back into (2.21) shows that the form of the approximate MISE when the bandwidth, h, is chosen to minimize MISE is

$$\tfrac{5}{4}\left[\left[c_k\int\sigma^2(x)/f(x)dx\right]^{4/5}+d_K^{2/5}\left[\int\left(m''(x)\right)^2dx\right]^{1/5}\right]n^{-4/5}\,. \tag{2.23}$$

Studying (2.23) reveals that the approximate MISE converges to zero at a rate of $n^{-4/5}$, when the bandwidth is suitably chosen, which implies that the nonparametric estimate of the regression curve converges at a rate of $n^{-2/5}$.

Stone (1980, 1982) demonstrated the optimal rate of convergence for a large class of nonparametric estimators of the regression mean, and the asymptotic properties of local polynomial estimators. The optimal rate is the speed of convergence for which a nonparametric estimate cannot converge any faster. Within the scope of the discussion here, when estimating a regression function nonparametrically, the speed at which the estimate converges depends upon 3 entities: 1) the smoothness assumptions given to m, where d is the dth derivative of m assumed to exist, 2) the order of the kernel, and 3) the dimension of X, p. Stone showed that the optimal rate of convergence is n^{-r} where $r = \frac{d}{2d+p}$. In the example, we assumed $m(\cdot)$ had a smooth second derivative, kernel of order 2, and the dimension of X was one. Hence, the kernel weighted locally linear estimate achieves the optimal rate of convergence for nonparametric estimation.

Truong (1989, 1992) extended the work of Stone to cover the median in independent and correlated data. Falk (1993) extended Stone's work to the estimation of a general class of regression functionals in independent data, which includes the mean, median and quantiles.

2.4 Confidence Intervals:

In the theoretical treatment of nonparametric regression, frequently the variation of the estimate is represented by a pointwise confidence interval at a fixed location of the covariates, for example, Chambers and Hastie (1991), Hastie and Tibshirani (1990). The justification is that pointwise confidence intervals are a descriptive technique to give an idea of the variability of the curve. However, this representation of the variability cannot be used for statistical inference because the simultaneous coverage probability of the set of intervals is well below the alpha level of the individual intervals. So, one must distinguish the situations where pointwise intervals are adequate from those where simultaneous coverage probability is important.

With the LRC data we may wish to plug in a set of clinical values and obtain a a set of predicted outcomes. We would like to have a smooth confidence belt surrounding the regression function over a finite region of the random covariates. A rigorous approach would be to find the distribution that the regression surface estimate converges to over the compact domain of the variable space. Central limit theory leads to the conjecture that this distribution would turn out to be a Gaussian process. Then we would find the distribution of the supremum of this Gaussian process over the domain of the conditioned variables in order to make statements about the variability of the predictions in the data analysis. Unfortunately, the application of this approach has met with disappointing results because the confidence belt is usually too large to make useful conclusions about the behavior of the regression function. These disappointing results suggest that the distribution of the supremum may attribute the curve estimate with too much variability (personal communication with J. S. Marron). The supremum of the stochastic process approach attempts to describe the variability of the curve in a space of infinite dimensionality. Indeed, the curve estimate belongs to the infinite dimensional space of smooth functions. However, this space does not include the large space of curves that are not smooth. Therefore, the infinite dimensional space of smooth functions is smaller than the infinite dimensional space that the distribution of supremum of the stochastic process describes.

Sample re-use methods have been proposed in the literature as an alternative to the analytic approach. An advantage to the sample-reuse approach is that it requires fewer assumptions about the distribution of the regression estimates. Essentially, the idea is to choose a suitable grid of points over the region of interest and approximate the joint distribution of the estimates at the chosen points with a bootstrap distribution. Although a continuous band around the regression estimate gives an appealing picture, for practical data analysis a suitably dense finite set of error bars loses little information and is much easier to compute and analyze [Härdle and Marron (1991)]. To obtain the correct simultaneous coverage probability for the error bars, one could use a conservative Bonferonni approach, or a more general approach detailed by Härdle and Marron (1991). For a reasonably dense grid of points, the bootstrap method is reputed to provide an adequate description of the variability of the regression function with little loss of information [Härdle and Marron (1991)].

For the case of the fixed design, Eubank and Speckman (1993) have developed a data driven confidence band estimating the distribution of the supremum of the estimate of the mean over a finite region of the covariate. The advantage that their method offers is that it is not dependent on the number of locations chosen to place the confidence bars, such as the Härdle-Marron approach. Their bias corrected method was shown to have the correct asymptotic coverage behavior. The methodology of Eubank and Speckman (1993) would need to be extended to the case of random designs in order to use it with the methodology developed in Chapter 3.

2.5 Nonparametric Regression for a Multivariate Response:

In Chapter I, we motivated the model by examining the third visit of the LRC data. However, the variables of interest were measured at multiple visits, i.e. 1,...,9, which casts a multivariate aspect to the problem. Although model (1.2) is not specifically designed to analyze a multivariate response and covariates, it can be used to examine data from multiple visits; and we do this in Chapter VII, where the LRC data is analyzed. Therefore, in the present section we review the smoothing literature which addresses a multivariate response.

Given a dataset with a multivariate response, the joint estimation of the regression function over all response variables using nonparametric estimation is a difficult problem that has not been solved. In the smoothing literature, only a few attempts have been made to analyze longitudinal data. The first nonparametric approach to the analysis of longitudinal data was by Gasser et al. (1984). Using the kernel function defined by the weights in (2.7), they nonparametrically estimated the growth curve, growth acceleration and velocity individually for 45 boys and 45 girls. A description of the relative merits of their study as well as their comparison to the popular parametric methods can be found in Appendix A. Although their study was an important step toward the development of nonparametric methods for approximating the response surface when successive observations have been taken on the same experimental unit, the primary criticism of their study was that they treated the serial observations from each child as independent in the selection of the optimal bandwidth. Hart and Wehrly (1986) showed that when correlation is accounted for, the optimal bandwidth can be larger or smaller than the optimal bandwidth assuming independence. Secondly, the variance of their growth curve estimate could be reduced by using the kernel weighted locally linear estimate.

Hart and Wehrly (1986) proposed a methodology for nonparametric smoothing of data with repeated measurements. In the perfectly balanced data fixed design setting, they proposed to take a simple average of the independent observations at each visit, which leaves a series of correlated mean values. The rest of the problem deals with how to account for the correlated structure when smoothing the means and how it affects the optimal bandwidth. If we denote the mean of the independent observations at each visit as $\overline{Y}_v$ the proposed model is

$$\overline{Y}_v = m(X_v) + \overline{\epsilon}_v, \quad v = 1, \ldots, V \quad , \tag{2.24}$$

where X_v is the value of the metameter or the controlled variable of interest, $\overline{\epsilon}_v = \frac{1}{n}\sum_{i=1}^{n} \epsilon_{iv}$, $m(\cdot)$ is a smooth function, and

$$\operatorname{cov}(\overline{\epsilon}_v, \overline{\epsilon}_{v'}) = \frac{\sigma^2}{n}\rho(X_v - X_{v'}) \quad , \tag{2.25}$$

The correlation function ρ is even with $\rho(0) = 1$ and $|\rho(\mu)| \le 1$ for all $\mu \in [-1,1]$. With (2.25) as the assumption of the correlation structure and using the Gasser-Müller kernel estimator, Hart and Wehrly showed that the optimal bandwidth is

$$h^*_{opt} = \left[\frac{\sigma^2\rho''(0)C_K}{d_K^2(m''(x))^2}\right]^{1/2} n^{-1/2} \quad , \tag{2.26}$$

where $C_K = 2\int_{-1}^{1}\int_{v}^{1}(u-v)K(u)K(v)dudv$. This compares with the optimum bandwidth assuming the observations are independent

$$h_{opt} = \left[\frac{\sigma^2 c_K}{d_K^2(m''(x))^2}\right]^{1/5} (nV)^{-1/5} . \tag{2.27}$$

Examination of (2.26) and (2.27) show that small V and large serial correlation could result in a smaller optimum bandwidth than with uncorrelated errors. This makes sense when considering that without accounting for the serial correlation, the kernel estimate attributes all of the smoothness in the curve to the mean function. However, since some of the smoothness in the data should be attributed to the correlation, the optimal bandwidth should be smaller resulting in a rougher estimate of the mean function. When V is large enough, the optimal bandwidth (2.26) could be larger than (2.27). Simulations showed that V must be greater than twenty for (2.26) to have an effect on the asymptotic relative efficiency of $\widehat{m}(x)$. One advantage to the Hart Wehrly method is that the usual requirement of $nh \to \infty$ is not necessary for consistency. However, criticism of the method can be made about the assumption of the correlation structure. Although Hart and Wehrly have not specified a form for the correlation function in their theoretical arguments, in order to estimate the optimal bandwidth in the numerical study they assume $\rho(v) = \exp(-\alpha\,|\,v\,|\,)$. This may imply that in practical situations, some parametric specification must be made about the correlation function, and then some estimation procedure used to estimate the parameters. This methodology leaves out any analysis of the individual characteristics of the regression curves. Also, for short series, such as the LRC data, their method would provide a highly variable estimate.

Additionally, as of yet no methodology has been introduced to evaluate the effect of covariates on the mean function. This suggests that Hart and Wehrly (1986) have successfully introduced a methodology for nonparametrically approximating the regression function of correlated means over time; however, not addressed one of the main objectives of data analysis, which is the effects of covariates. Also, their own methodology could be improved by use of the kernel weighted locally linear estimator which has better boundary conditions, bias and variance properties, as shown in Section 2.1.

Short sequences of measurements taken on a sample of individuals can be thought of as a collection of independent curves. Rice and Silverman (1991) assumed that $\mathrm{cov}\{Y_{iv}, Y_{iv'}\} = \gamma(v, v') = \sum_t \gamma_t \phi_t(v) \phi_t(v')$ is an orthogonal expansion of γ in terms of eigenfunctions. An individual's curve may then be represented as

$$Y_{iv} = m(v) + \sum_t \xi_t \phi_t(v) \quad , \; v = 1, \ldots, V \quad , \tag{2.28}$$

where the ξ_t are uncorrelated random variables with zero means and variance $E\xi_t^2 = \gamma_t$. Thus, the deviation of each observation from the mean is a sum of orthogonal curves multiplied by the uncorrelated random errors. Rice and Silverman (1991) follow the method of Hart and Wehrly (1986) of smoothing the means at each visit, v, to estimate $m(\cdot)$. However, Rice and Silverman (1991) use smoothing splines instead of the kernel procedure, and they use a cross-validated estimate of the optimal bandwidth proposed by Wehrly and Hart (1988) that does not depend upon the covariance. This cross-validated bandwidth procedure leaves out one experimental unit's set of observations at a time to be predicted by the remaining observations. They also obtain smooth estimates of the eigenfunctions using smoothing splines. Many of the same criticisms of Hart and Wehrly's (1988) approach apply here. Also, we cannot compare the methodology on a theoretical level since the asymptotic arguments were not developed.

Considering the methodology discussed in this Section, 2.5, it is evident that nonparametrically approximating the relationship of a multivariate response and covariates is a new topic (since 1984) that has much room for growth. Specifically, the effects of covariates have not been addressed in the literature. Additionally, the estimation of the regression function jointly over all responses has not been worked out and remains a difficult problem. As an intermediate approach in Chapter VII we use model (1.2) to examine the effects of compliance on total cholesterol in the LRC data at consecutive time points by examining the regression surface marginally at each visit.

2.6 Partially Linear Estimation

When examining the relationship between the response variable, Y, and some explanatory variables, $X_1,\ldots,X_d$, there may exist some prior knowledge about the relationship of certain variables that would lead to a partially parametric specification of the regression function. Several investigators have examined the properties of partially linear additive models of the form

$$Y_i = X_i'\beta + m(t_i) + \epsilon_i, \quad i = 1,\ldots,n \ , \tag{2.29}$$

where X_i is a vector of length d, β is a vector of regression coefficients, and m is a smooth function. If one is able to justify the linear part of the mean function, then one would hope that some advantage could be obtained by this added information about the regression function. In particular, we would hope that the estimate of β could achieve the parametric rate of convergence, $n^{-1/2}$, while at the same time achieving a faster rate of convergence for the estimate of m by reducing the dimensionality. Several authors have addressed this question using various estimation techniques. Estimates of β and m have been found by minimizing

$$n^{-1}\sum_{i=1}^{n}[Y_i - X_i\beta - m(t_i)]^2 + \lambda J(m) \tag{2.30}$$

with respect to β and m, where $J(m)$ is a functional chosen to penalize for roughness in the fitted m. Chen and Shiau (1991) mention that Wahba (1984) proposed a generalized cross-validation technique for the estimation of the smoothing parameter, λ.

Model (2.29) is sometimes called a partial spline model and it has been studied in the literature by Engle et al. (1986), Green et al. (1985), Shiau et al. (1986), Heckman (1986), Rice (1988), Speckman (1988), Chen an Shiau (1991), Eubank, Hart, and Speckman (1990) among others. In balanced cases of analysis of covariance, Heckman (1986) proved asymptotic normality of β and that the estimate of β had a convergence rate of $n^{-1/2}$ if X and t were not related to each other. However, Rice (1986) showed that when X and t were related to each other, the asymptotic bias of β dominated the variance and the root n rate could only be achieved if the estimate of m was undersmoothed. Therefore, the method of choosing λ proposed by Wahba (1984) would not be appropriate in that case. Chen and Shiau (1991) state that Eubank and Whitney (1989) reported similar results.

Motivated by the startling negative results above, several authors developed methods of estimation for (2.29) where parametric rates are achieved for β without undersmoothing m even when X and t are related. Denby (1986) and Speckman (1988) independently proposed partial regression estimates for β and m motivated by considering first a parametric representation of m as $W\phi$, where W is an $n \times q$

matrix of full rank and ϕ is an additional parameter. Model (2.29) then takes the form

$$Y_i = X_i\beta + W_i\phi + \epsilon_i \quad i = 1,\ldots,n. \tag{2.31}$$

It is easily shown that the normal equations for finding β and ϕ are

$$X'X\beta = X(Y - W\phi) \; , \tag{2.32.a}$$

$$W\phi = P_W(Y - X\beta) \; , \tag{2.32.b}$$

where $P_W = W(W'W)^{-1}W'$ denotes the projection onto the column space of W. As described in Speckman (1988), Green et al. (1985) proposed replacing P_W in (1.32) by a smoother M and simultaneously solving

$$\begin{aligned} \beta &= R(Y - m) \\ f &= M(Y - X\beta) \; , \end{aligned} \tag{2.33}$$

where R is an estimator of treatment effects. Taking M to be a smoother matrix K from kernel smoothing and letting $R = (X'X)^{-1}X'$, the Green-Jennison-Scheult (1985) estimators are defined as

$$\widehat{m}_{GJS} = K(Y - X\widehat{\beta}_{GJS}) \text{ , and} \tag{2.34.a}$$

$$\widehat{\beta}_{GJS} = (X'(I-K)X)^{-1}X'(I-K)Y \; . \tag{2.34.b}$$

Denby (1986) and Speckman (1988) considered a slightly different estimate motivated by partial residuals to estimate β and m by adjusting the response vector Y and the variables X for t, respectively, as $(I-K)Y$ and $(I-K)X$, where K is the smoother matrix defined in (2.34a,b). By denoting $\widetilde{Y} = (I-K)Y$ and $\widetilde{X} = (I-K)X$ the estimates of β and m are defined as

$$\widehat{\beta}_p = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{Y} \text{ and} \tag{2.35.a}$$

$$\widehat{m} = K(Y - X\widehat{\beta}_p) \ . \tag{2.35.b}$$

Note that the estimate of β_p (2.35.a) is formally the solution to the weighted least squares criterion

$$\min_{\beta} ||(I-K)(Y-X\beta)||^2 \ , \tag{2.36}$$

where $||\ ||$ denotes the Euclidian norm $||v||^2 = \sum v^2$ for $\mathbf{v} = (v_1,\ldots,v_n)' \in \mathbf{R}^n$. Speckman (1988) investigated the properties of $\widehat{\beta}_{GJS}$ and $\widehat{\beta}_p$ under a fixed design for t using kernel regression. He assumes the relationship $x_{ij} = g_j(t_i) + \eta_{ij}$, $i = 1,\ldots,n$, $j = 1,\ldots,d$, where x_{ij} is the ijth element of X, and η_{ij} are independent mean zero random variables independent of the ϵ_i. His results showed that the bias of $\widehat{\beta}_{GJS}$ had the nonparametric rate of convergence and dominated the variance; thus, $\widehat{\beta}_{GJS}$ did not achieve the parametric rate of convergence. However, the bias of $\widehat{\beta}_p$ was negligible compared to the variance in the limit, which allowed $\widehat{\beta}_p$ to achieve root n convergence. Speckman (1988) shows that the cross-validation and generalized cross-validation approach originally proposed by Craven and Wahba (1979) produces an unbiased estimate of the optimal bandwidth using the partial regression estimates proposed in his paper. Chen and Shiau (1991) essentially proposed the same method as Speckman's partial residual approach, however using smoothing splines to construct the smoothing matrix.

Although in some cases the structure of (2.31) may be viewed with skepticism, the estimate of the regression function given by (2.31) can be viewed as an approximation that is preferable in small sample sizes to a fully nonparametric estimate, since the model bias introduced by the parametric component may be offset by the reduction in variance. The relatively modest rates of convergence for multivariate kernel estimates makes the assumption of linearity on some of the variables attractive, particularly if there is sound scientific reasoning suggesting such an assumption. Additionally, if one of the primary goals is inference on a particular variable, say in the context of analysis of covariance, the partially linear model offers a flexible alternative to the fully parametric approach and a way to make inference that would otherwise be difficult in a fully nonparametric approach.

Using large sample arguments, it is apparent that the partial residual approach proposed by Denby (1986) and Speckman (1988) produces estimates with the most desirable analytic qualities. Speckman (1988) emphasized rates of convergence in his results without giving specific forms for bias and variance. In the proposed work we obtain specific forms for the bias and variance of the components of the regression function using the partial residual estimation approach proposed by Speckman (1988) combined with the kernel weighted locally linear method. By the arguments presented in Chapter II, the estimate of (1.2) using the proposed methodology should have the optimal rates of convergence as well as the most desireable forms for the bias and variance.

In Chapter III, we discuss the method of estimation of (1.2) combining the

CHAPTER III: THE METHOD

In this chapter we present the method of estimation of α and m in (1.2), which is in the spirit of Speckman (1988) described in Section 2.5. Following the notation of the previous chapter, let

$$(Y_1, B_1, X_{11}, X_{12})', \ldots, (Y_n, B_n, X_{n1}, X_{n2})'$$

be a set of independent and identically distributed random vectors. The compliance measures (X_{i1}, X_{i2}) have common density, f, bounded away from zero on support $[0,\ 100]^2$. Denote the conditional expectation of Y_i given $(X_{i1} = x_1, X_{i2} = x_2)$ as

$$g(x_1, x_2) = E(Y_i \mid X_{i1} = x_1, X_{i2} = x_2)\ ,\ i = 1, \ldots, n\ , \tag{3.1}$$

where $g(\,\cdot\,)$ is a smooth function. For ease of notation, denote the vector $(X_{i1}, X_{i2})'$ as $\boldsymbol{X}_i'$ and the vector $(x_1, x_2)'$ as $\boldsymbol{x}'$. Following the results of Chapter I, we shall use nonparametric regression to estimate the conditional means of cholesterol as a function of the compliance measures. The local linear nonparametric regression estimator is discussed in Section 2.1 where it is argued that the conditional asymptotic bias and variance are particularly appealing and appear superior to those of the Nadaraya-Watson or Gasser-Müller estimators. In particular, Fan (1992) showed that the local linear estimator achieves optimal minimax efficiency in the class of linear estimators given a suitable choice of kernel and bandwidth. The locally linear estimate of $g(\boldsymbol{x})$ is defined as $\widehat{g}(\boldsymbol{x}) = \widehat{c}_1$, where $[\widehat{c}_1, \widehat{\boldsymbol{d}}_1]$ minimize

$$\sum_{i=1}^{n}\Big(Y_i - c_1 - \boldsymbol{d}_1{}'(\boldsymbol{X}_i - \boldsymbol{x})\Big)^2 K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x}) \ . \tag{3.2}$$

The estimate $\widehat{g}(\boldsymbol{x})$ is found by fitting a plane to the data using weighted least squares. When using a Gaussian kernel and a bandwidth matrix $\boldsymbol{H}$, the weight given to a point $\boldsymbol{X}_i$ is the value of the $N(\boldsymbol{X}_i - \boldsymbol{x},\ \boldsymbol{H})$ density which has elliptical contours of the form $(\boldsymbol{X}_i - \boldsymbol{x})'\boldsymbol{H}^{-1}(\boldsymbol{X}_i - \boldsymbol{x}) = c, c > 0$. The $N(\boldsymbol{\mu}, \Sigma)$ is the Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. This Gaussian kernel allocates less weight to $\boldsymbol{X}_i$ the farther away it is from $\boldsymbol{x}$, but the weights also depend on the value of $\boldsymbol{H}$, which determines the length of the major and minor axes as well as their orientation at a given density level. If g has a high amount of curvature near $\boldsymbol{x}$ then the weights controlled by $\boldsymbol{H}$ should be such that more information is provided from nearby observations in order to reduce the bias. On the other hand, if g is nearly linear at $\boldsymbol{x}$ then the bias will be smaller, so we would want to reduce variance. In this case, we would want larger elliptical contours to get more observations included in the fitting process.

Depending upon the problem, different types of bandwidth matrices, $\boldsymbol{H}$, are advantageous. In the context of bivariate density estimation, Wand and Jones (1993) examined the following three classes of smoothing matrices

$$\boldsymbol{H}_1 = \{h_1^2 I_2 : h_1 > 0\}, \quad \boldsymbol{H}_2 = \{\mathrm{diag}(h_1^2, h_2^2) : h_1, h_2 > 0,\},$$

and

$$\boldsymbol{H}_3 = \left\{ \begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix} : \ h_1, h_2 > 0,\ |h_{12}| \ < h_1 h_2 \right\} .$$

They also compared the performance of smoothing matrices such as $\boldsymbol{H}_2$ and $\boldsymbol{H}_3$ where the elements are determined as a scalar multiple of the covariance matrix of the X's. This is referred to as *sphering* or *scaling* the data. Using the sphering or scaling approach with entries in each element of the bandwidth matrix, the weights become the classical Mahalanobis' distance. Wand and Jones (1993) make a convincing argument that when the density of the variables has elliptical structure, the sphering and scaling approach will do well in terms of asymptotic mean integrated squared error, where doing well means coming close to what one could do with a fully flexible smoothing matrix. However, for cases when at least one of the marginal densities does not have a nice unimodal shape, the sphering/scaling approach does not perform well due to the fact that the entries in the sample covariance matrix do not take into account the curvature of the density. For instance, if one of the marginal densities is bimodal, the sample variance of that variable does not make a good measure for the the optimal amount of smoothing in that direction.

In practice, Wand and Jones (1993) conclude that $\boldsymbol{H}_1$ may be a start, but that in most cases one can do better with the flexibility of smoothing by different amounts in each direction given by $\boldsymbol{H}_2$. When the orientation of the density is not along the coordinate axes, there is much to be gained by including the off-diagonal orientation parameter of $\boldsymbol{H}_3$. Although, in the bivariate case, one can simply choose the optimal rotation by eye and then use $\boldsymbol{H}_2$ on the rotated data. The results of using each type of smoothing matrix will be incorporated into Section 3.1 regarding the asymptotic behavior of the estimates.

Returning to the next step in the estimation of (1.2), denote the conditional expectation of B_i given $(X_{i1} = x_1, X_{i2} = x_2)$ as

$$u(x_1, x_2) = E(B_i \mid X_{i1} = x_1, X_{i2} = x_2) \ , \ i = 1, \ldots, n \ , \tag{3.3}$$

where $u(\cdot)$ is a smooth function. The locally linear estimate of $u(\boldsymbol{x})$ is defined as $\widehat{u}(\boldsymbol{x}) = \widehat{c}_2$, where $[\widehat{c}_2, \widehat{\boldsymbol{d}}_2]$ minimize

$$\sum_{i=1}^{n}\left(B_i - c_2 - \boldsymbol{d}_2'(\boldsymbol{X}_i - \boldsymbol{x})\right)^2 K_H(\boldsymbol{X}_i - \boldsymbol{x}). \tag{3.4}$$

Let $\widetilde{Y}_i = Y_i - \widehat{g}(\boldsymbol{X}_i)$, and $\widetilde{B}_i = B_i - \widehat{u}(\boldsymbol{X}_i)$ denote the variables Y_i and B_i after "adjustment" for dependence on $\boldsymbol{X}_i$. The estimate for α in (1.2) is defined as the solution to

$$\min_{\alpha}\left[\sum_{i=1}^{n}\left(\widetilde{Y}_i - \alpha\widetilde{B}_i\right)^2\right]. \tag{3.5}$$

To estimate the function $m(\cdot)$ in (1.2) , define $m(\boldsymbol{x}) = \widehat{a}$, where $[\widehat{a}, \widehat{\boldsymbol{b}}]$ minimize

$$\sum_{i=1}^{n}\left((Y_i - \widehat{\alpha}B_i) - a - \boldsymbol{b}'(\boldsymbol{X}_i - \boldsymbol{x})\right)^2 K_H(\boldsymbol{X}_i - \boldsymbol{x}) . \tag{3.6}$$

<u>3.1 Identifiability</u>:

Model (1.2) is explicitly written without an intercept in order to make the elements of the model identifiable. Because there are no special constraints placed on $m(x_1, x_2)$, there is a free constant in the function. If we allowed an intercept in (1.2), we could arbitrarily change the intercept by adding a constant and subtracting that same constant from $m(x_1, x_2)$ without changing the value of the estimate of the regression function. Hence, this would make the elements of the model not identifiable. By adding and subtracting a constant to (1.2) without changing the value of the regression function we clarify that as written, model (1.2) is identifiable. We express (1.2) as

$$E(\boldsymbol{Y}|\boldsymbol{B},\, \boldsymbol{X}_1,\, \boldsymbol{X}_2) = \alpha\boldsymbol{B} + \boldsymbol{M}\ , \tag{3.7}$$

where $\boldsymbol{M} = [m(X_{11}, X_{12}), \ldots, m(X_{n1}, X_{n2})]'$. Since (1.2) is written without an intercept, we add and subtract a constant from the parametric and nonparametric components of the model, so that (3.7) becomes:

$$E(\boldsymbol{Y}|\boldsymbol{B},\, \boldsymbol{X}_1,\, \boldsymbol{X}_2) = (\alpha + c)\boldsymbol{B} + (\boldsymbol{M} - c\boldsymbol{B})\ . \tag{3.8}$$

There exists a smoothing matrix $\boldsymbol{K}$ (explicitly defined in Chapter 4), with respect to $\boldsymbol{X}$ such that $\boldsymbol{KY} = \widehat{\boldsymbol{G}}$, where $\widehat{\boldsymbol{G}} = [\widehat{g}(X_{11}, X_{12}), \ldots, \widehat{g}(X_{n1}, X_{n2})]'$. Then define $\widetilde{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{K})(\boldsymbol{Y} + c\boldsymbol{B})$ and $\widetilde{\boldsymbol{B}} = (\boldsymbol{I} - \boldsymbol{K})$. Then the solution to

$$\min_{\alpha + c} \| \widetilde{\boldsymbol{Y}} - (c + \alpha)\widetilde{\boldsymbol{B}} \|$$

are the normal equations

$$(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})(c + \alpha) = \widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{Y}},$$

which yield

$$\begin{aligned} c + \widehat{\alpha} &= (\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'(\boldsymbol{I} - \boldsymbol{K})(\boldsymbol{Y} + c\boldsymbol{B}) \\ &= (\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'(\boldsymbol{I} - \boldsymbol{K})\boldsymbol{Y} + c(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'(\boldsymbol{I} - \boldsymbol{K})\boldsymbol{B} \\ &= (\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{Y}} + c(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}} \end{aligned}$$

$$= (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'\widetilde{Y} + c\,.$$

Hence, $\hat{\alpha} = (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'\widetilde{Y}$, which is the same expression for the estimate as without adding and subtractin the constant c.

As evident by Chapter II, locally linear estimation and the partially linear model have been independently examined in the literature. However, the results of combining the two methodologies have not been considered, which leaves us with several issues to be addressed before analysis with the proposed method is undertaken. Firstly, do the appealing properties of bias and variance of the the locally linear estimate carryover to $\hat{m}$ in (1.2)? Since the partially linear model has not been explicitly examined in the multivariate smoothing case, under what conditions can we estimate model (1.2) using Speckman's (1988) two-stage procedure and still retain the root n rate of convergence for the linear parameter? What are the finite sample properties of the estimators of (1.2)? Lastly, what do the bias and variance of the estimates look like using the three types of bandwidth matrices? The next two chapters addresses these issues.

CHAPTER IV: ASYMPTOTIC BEHAVIOR

The theorems in this chapter give forms for the pointwise asymptotic bias, variance and normality of the proposed estimators. Proofs for the theorems are provided in Chapter V. For describing the variability of the regression curve these theorems are limited to a single point. In order to construct an appropriate statement about the variability of the regression curve over a compact range of the covariates, we must incorporate methodology for obtaining a simultaneous confidence band, or a suitably dense set of confidence bars, around the curve. A discussion of such methodology can be found in Section, 2.4. Combining the results of these theorems with the methodology discussed in Section 2.4 will provide appropriate statements about the variability of the regression curve over a compact region of the regression surface.

4.1 Definitions:

Because of the complexity of notation involved in the analytical computations, the dependence of random matrices on n and $\boldsymbol{H}$ is suppressed.

1) $\boldsymbol{Y} = (Y_1,\ldots,Y_n)' \qquad \boldsymbol{B} = (B_1,\ldots,B_n)'$.

2) $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$, $i = 1,\ldots,n;$ $\qquad \boldsymbol{X}^*_{i(\boldsymbol{x})} = \boldsymbol{H}^{-1/2}(\boldsymbol{X}_i - \boldsymbol{x})$, $i = 1,\ldots,n,$

where $\boldsymbol{H}$ is the positive-definite bandwidth matrix associated with regressing

Y_i on X_i $i = 1,\ldots,n$ when the linear parameter $\alpha = 0$. $H^{1/2}$ is defined as the matrix such that $H^{1/2}H^{1/2} = H$, and similarly for $H^{-1/2}$.

4) $X(x)$ is the n by 3 matrix equal to $\begin{bmatrix} 1 & X_1^{*\prime}(x) \\ \vdots & \vdots \\ 1 & X_n^{*\prime}(x) \end{bmatrix}$.

5) $W(x) = \text{diag}\{K_H(X_1 - x),\ldots,K_H(X_n - x)\}$, where
$K_H(u) = |H|^{-1/2}K(H^{-1/2}u)$.

6) $K^*(X_i;X_j) = e_1'(X'(x_j)W(x_j)X(x_j))^{-1}(1 \quad X_i^{*\prime}(x_j))'K_H(X_i - X_j)$,
where $e_1 = [1,0,0]'$.

7) The ith row of the matrix $K^* = [K^*(X_1;X_i),\ldots,K^*(X_n;X_i)]$.
Note that $\sum_{i=1}^{n} K^*(X_i;X_j) = 1$ and $\sum_{i=1}^{n} K^*(X_i;X_j)(X_i - X_j) = 0$ for all j.

8) $\widetilde{Y} = (I - K^*)Y \qquad \widetilde{B} = (I - K^*)B$

9) $\hat{m}(x) = e_1'(X'(x)W(x)X(x))^{-1}X'(x)W(x)[Y - \hat{\alpha}B] = \sum_{i=1}^{n} K^*(X_i;x)[Y_i - \hat{\alpha}B_i]$,

$\hat{u}(x) = e_1'(X'(x)W(x)X(x))^{-1}X'(x)W(x)B$ and $\hat{\alpha} = (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'\widetilde{Y}$.

10) $M = (m(X_1),\ldots,m(X_n))' \qquad \widehat{M} = K^*[Y - \hat{\alpha}B] = (\hat{m}(X_1),\ldots,\hat{m}(X_n))'$

11) $U = (u(X_1),\ldots,u(X_n))' \qquad \widehat{U} = K^*B = (\hat{u}(X_1),\ldots,\hat{u}(X_n))'$

12) The 2 by 2 Hessian matrices of the bivariate regression functions m and u at $\boldsymbol{x}$ are denoted by $\Gamma_m(\boldsymbol{x})$ and $\Gamma_u(\boldsymbol{x})$. $D_m(\boldsymbol{x})$ and $D_u(\boldsymbol{x})$ are 2 by 1 vectors of first order partial derivatives of the function m and u.

13) $\max\{\boldsymbol{H}\}$ is the scalar defined by $\max\{h_1^2,\ h_{12},\ h_2^2\}$.

14) $\| \boldsymbol{A} \|$ is defined as modulus of the largest singular value of $\boldsymbol{A}$. $\| \boldsymbol{A} \|_{\infty}$ is defined as $\max_i \sum_{j=1}^{n} |A_{ij}|$. $\| \boldsymbol{v} \| = \left(\sum_{i=1}^{n} v_i^2 \right)^{1/2}$ and $\| \boldsymbol{v} \|_{\infty} = \max_i | v_i |$.

4.2 Assumptions:

a) (i) $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ are random samples from a continuous density function f, with compact support $C_1 \in \Re^2$ with $0 < f(\boldsymbol{x}) < \infty$ for $\boldsymbol{x} \in C_1$.
(ii) $B_1, \ldots, B_n$ are random samples from a continuous density function f_B, with support $\in \Re$ with $0 < f_B(b) < \infty$ for $b \in \Re$.

b) $(i) \int (y - \alpha b) f(y \mid b, \boldsymbol{x}) dy = m(\boldsymbol{x})$, $\int \left(y - E(y \mid \boldsymbol{x}, b)\right)^2 f(y|\boldsymbol{x}, b) dy = \sigma^2 < \infty$ for all $\boldsymbol{x} \in C_1$ and $b \in C_2$,

$(ii) \int b f_B(b \mid \boldsymbol{x}) db = u(\boldsymbol{x})$, $0 < \int \left(b - u(\boldsymbol{x})\right)^2 f_B(b|\boldsymbol{x}) db = \sigma_B^2 < \infty$ for all $\boldsymbol{x} \in C_1$.

c) (i) Each element of $\Gamma_m(\boldsymbol{x})$ is continuous and bounded for $\boldsymbol{x} \in C_1$.
(ii) Each element of $\Gamma_u(\boldsymbol{x})$ is continuous and bounded for $\boldsymbol{x} \in C_1$.

d) The bandwidth matrix $\boldsymbol{H}$ is symmetric and positive definite.

e) The kernel function K is a bivariate continuous kernel with compact support such that $\int K(\boldsymbol{u})d\boldsymbol{u} = 1$. All odd order moments of K vanish, that is, $\int u_1^l u_2^k K(\boldsymbol{u})d\boldsymbol{u} = 0$ for all non-negative integers l and k such that their sum is odd. Also, $\int \boldsymbol{u}\boldsymbol{u}'K(\boldsymbol{u})d\boldsymbol{u} = \mu_2(K)I_2$ where $\int u_i^2 K(\boldsymbol{u})d\boldsymbol{u} \neq 0$ independent of i and I_2 is the 2 by 2 identity matrix. These conditions are satisfied by spherically symmetric kernels and product kernels based on symmetric univariate kernels, provided they have all moments existing.

f) For $0 < \delta < 1/2$, $\quad E(|B_1 - u(\boldsymbol{X}_1)|^{2+\delta} \mid \boldsymbol{X}_1) < \infty$.

g) For $0 < \delta < 1/2$, $E(|Y_1 - \alpha B_1 - m(\boldsymbol{X}_1)|^{2+\delta} \mid \boldsymbol{X}_1, B_1) < \infty$.

4.3 Theorems:

Theorem 1. Let assumptions (a) - (e) hold, and suppose that $n^{-1}|\boldsymbol{H}|^{-1/2}$ and each entry of $\boldsymbol{H}$ tends to zero as $n \to \infty$ with $\boldsymbol{H}$ remaining symmetric and positive definite, then

$$E\{\widehat{\alpha} - \alpha \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, B_1, \ldots, B_n\} = (\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'(\boldsymbol{I} - \boldsymbol{K}^*)\boldsymbol{M} \qquad (4.1a)$$

and

$$\operatorname{var}[\widehat{\alpha} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, B_1, \ldots, B_n] = \sigma^2(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'(\boldsymbol{I} - \boldsymbol{K}^*)^2\widetilde{\boldsymbol{B}}(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}. \qquad (4.1b)$$

Moreover, $(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'(\boldsymbol{I} - \boldsymbol{K}^*)\boldsymbol{M} = O_p\big((\max\{\boldsymbol{H}\})^2\big)$
and $\sigma^2(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1}\widetilde{\boldsymbol{B}}'(\boldsymbol{I} - \boldsymbol{K}^*)^2\widetilde{\boldsymbol{B}}(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}})^{-1} = \dfrac{\sigma^2}{n\sigma_B^2} + o_p(n^{-1})$.

This theorem shows that the asymptotic variance of the estimate of the linear component is of the order $1/n$. Thus, when using the optimal bandwidth

$\max\{\boldsymbol{H}\} \sim n^{-2/(4+d)}$ (here $d = 2$) and for $1 \leq d < 4$, the asymptotic bias $= O_p(\max\{\boldsymbol{H}\}^2) = O_p(n^{-\frac{4}{4+d}}) \ll n^{-1/2}$, so the squared bias is negligible compared with the variance. In this case, the estimate of the linear component in (1.2) achieves the parametric rate of convergence when using the kernel weighted locally linear estimate for the smooth part of (1.2). Additionally, by posing this theorem in the multivariate context for the smooth part, we demonstrate a new result that for $d > 4$ higher order kernels are necessary to preserve the parametric rate of convergence for the linear estimate.

Theorem 2. Under assumptions (a) - (e) and suppose that $n^{-1}|\boldsymbol{H}|^{-1/2}$ and each entry of $\boldsymbol{H}$ tends to zero as $n \to \infty$ with $\boldsymbol{H}$ remaining symmetric and positive definite, then the elements of $E\{\widehat{\boldsymbol{M}} - \boldsymbol{M} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, B_1, \ldots, B_n\}$ are $O_p(\max\{\boldsymbol{H}\})$ and the elements of $\mathrm{var}(\widehat{\boldsymbol{M}} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, B_1, \ldots, B_n\}$ are $O_p(n^{-1}|\boldsymbol{H}|^{-1/2})$.

Theorem 2 shows that by regressing $Y - \widehat{\alpha}B$ on $\boldsymbol{X}$, locally linear estimation of m behaves asymptotically like Theorem 2.1 of Ruppert and Wand (1994), even in the presence of the parametric term.

Theorem 3. Under assumptions (a) - (f) and suppose that $|\boldsymbol{H}| = O(n^{-2/3})$ [*cf.* p. 66 remark (*iv*)] and each entry of $\boldsymbol{H}$ tend to zero as $n \to \infty$ with $\boldsymbol{H}$ remaining symmetric and positive definite, then

$$P\left[n^{1/2}\frac{(\widehat{\alpha} - E\widehat{\alpha})}{(\sigma/\sigma_B)} \leq t \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, B_1, \ldots, B_n\right] = \Phi(t) + o_p(1). \tag{4.1}$$

Theorem 3 shows that the estimate of the linear coefficient in model (1.2) achieves the desireable root n rate of convergence, even when jointly estimating a nonparametric term. For the statement of the next theorem, let $\widehat{m}_o(\boldsymbol{x})$ denote the locally linear estimator of $m(\boldsymbol{x})$ if α were known exactly, i.e. $\widehat{m}_o(\boldsymbol{x}) = a$ where $[a, \boldsymbol{b}']'$

minimize

$$\sum_{i=1}^{n}\Big((Y_i-\alpha B_i)-a-\boldsymbol{b}'(\boldsymbol{X}_i-\boldsymbol{x})\Big)^2 K_H(\boldsymbol{X}_i-\boldsymbol{x}) .$$

Theorem 4. Under assumptions (a) - (e) and (g) and suppose that $n^{-1}|\boldsymbol{H}|^{-1/2}$ and each entry of $\boldsymbol{H}$ tends to zero as $n\to\infty$ with $\boldsymbol{H}$ remaining symmetric and positive definite, then

$$(i)\quad E(\widehat{m}(\boldsymbol{x})-m(\boldsymbol{x})\mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,B_1,\ldots,B_n)$$
$$= E(\widehat{m}_o(\boldsymbol{x})-m(\boldsymbol{x})\mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,B_1,\ldots,B_n)(1+o_p(1)),$$

and

$$(ii)\quad \mathrm{var}(\widehat{m}(\boldsymbol{x})\mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,B_1,\ldots,B_n)$$
$$= \mathrm{var}(\widehat{m}_o(\boldsymbol{x})\mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,B_1,\ldots,B_n)(1+o_p(1)).$$

Also,

$$(iii)\ P\left[\sqrt{n|\boldsymbol{H}|^{1/2}}\left(\frac{\widehat{m}(\boldsymbol{x})-E\widehat{m}(\boldsymbol{x})}{\tau(\boldsymbol{x})}\right)\le t\mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,B_1,\ldots,B_n\right]=\Phi(t)+o_p(1),$$

where $\tau^2(\boldsymbol{x})=f^{-1}(\boldsymbol{x})\sigma^2\int K^2(\boldsymbol{v})d\boldsymbol{v}$.

The asymptotic results of this chapter demonstrate that the appealing bias and variance properties of local linear smoothing [Fan (1992)] hold as well for model (1.2), and that $\widehat{\alpha}$ maintains the root n rate of convergence under the method of estimation. Also, we investigated the properties of the estimates when using smoothing matrices. Proofs for the theorems in this chapter are provided in the next chapter.

CHAPTER V: PROOFS OF THEOREMS

The proofs of Theorems 1 - 4 detail the asymptotic results for the estimation of a two dimensional smooth function. These results are exactly analogous for higher dimensions. Before we begin the calculations for the proof of Theorem 1, we obtain some preliminary results in the form of lemmas.

Lemma 1. Let assumptions (a(i)), (d) and (e) hold and suppose $n^{-1}|\boldsymbol{H}|^{-1/2}$ and each entry of $\boldsymbol{H}$ tends to zero as $n \to \infty$ with $\boldsymbol{H}$ remaining symmetric and positive definite. Then there is a positive constant c_1 such that

$$\lim_n P\Big[n(\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})\boldsymbol{X}(\boldsymbol{x}))_{jj}^{-1} \leq c_1 \ for\ \boldsymbol{x} \in C_1 and\ j = 1, 2, 3\Big] = 1.$$

Comment. Lemma 1 points out that with the normalizaton of the $\boldsymbol{X}(\boldsymbol{x})$ matrix with $\boldsymbol{H}^{-1/2}$, similar to Stone (1982), the inverse of the matrix $(\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})\boldsymbol{X}(\boldsymbol{x}))$ exists in probability. Without the normalization, as in Ruppert and Wand (1994), the inverse does not exist in probability, although the forms for the asymptotic bias and variance remain the same.

Proof. It turns out that $n^{-1}(\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})\boldsymbol{X}(\boldsymbol{x})) =$

$$\begin{bmatrix} n^{-1}\sum_{i=1}^{n} K_H(\boldsymbol{X}_i - \boldsymbol{x}) & n^{-1}\sum_{i=1}^{n} K_H(\boldsymbol{X}_i - \boldsymbol{x})\boldsymbol{X}_i^{*\prime}(\boldsymbol{x}) \\ n^{-1}\sum_{i=1}^{n} K_H(\boldsymbol{X}_i - \boldsymbol{x})\boldsymbol{X}_i^{*}(\boldsymbol{x}) & n^{-1}\sum_{i=1}^{n} K_H(\boldsymbol{X}_i - \boldsymbol{x})\boldsymbol{X}_i^{*}(\boldsymbol{x})\boldsymbol{X}_i^{*\prime}(\boldsymbol{x}) \end{bmatrix}.$$

The arguments used for taking the expectation of the elements of $n^{-1}(X'_{(x)}W_{(x)}X_{(x)})$ are standard in the literature and follow from assumptions (a(i)) and (e). The matrix $n^{-1}(X'_{(x)}W_{(x)}X_{(x)})$ is partitioned into a scalar in the (1,1) entry, a 1 by 2 vector in the (1,2) entry, a 2 by 1 vector in the (2,1) entry and a 2 by 2 matrix in the (2,2) entry. Taking expectation of the scalar in the (1,1) position of $n^{-1}(X'_{(x)}W_{(x)}X_{(x)})$ yields

$$
\begin{aligned}
E\left[n^{-1}\sum_{i=1}^{n} K_H(X_i - x)\right] &= \int |H|^{-1/2} K\left(H^{-1/2}(u-x)\right) f(u)\,du \\
&= \int K(v) f(H^{1/2}v + x)\,dv \text{ , by change of variable} \\
&= \int K(v)\left(f(x) + [f(H^{1/2}v + x) - f(x)]\right)dv. \qquad (5.1)
\end{aligned}
$$

By assumption, K is continuous and has compact support, thus bounded. So by Lebesgue's Dominated Convergence Theorem, (5.1) is dominated by the quantity in the large parentheses. Since the integration, $\int K(v)dv$, over the support of K is unity, (5.1) $= f(x)(1 + o(1))$.

Taking the expectation of the vector in the (2,1) position of $n^{-1}(X'_{(x)}W_{(x)}X_{(x)})$ yields

$$
E\left[n^{-1}\sum_{i=1}^{n} K_H(X_i - x)X^*_{i(x)}\right]
$$

$$
= \int |H|^{-1/2} K\left(H^{-1/2}(u-x)\right)\left(H^{-1/2}(u-x)\right) f(u)\,du
$$

$$= \int K(\boldsymbol{v})\boldsymbol{v}f(\boldsymbol{H}^{1/2}\boldsymbol{v}+\boldsymbol{x})d\boldsymbol{v}$$

$$= \int \boldsymbol{v}K(\boldsymbol{v})\Big(f(\boldsymbol{x}) + [f(\boldsymbol{H}^{1/2}\boldsymbol{v}+\boldsymbol{x}) - f(\boldsymbol{x})]\Big)d\boldsymbol{v} = o(1).$$

Taking the expectation of the vector in the (2,2) partition of $n^{-1}(\boldsymbol{X}'_{(\boldsymbol{x})}\boldsymbol{W}_{(\boldsymbol{x})}\boldsymbol{X}_{(\boldsymbol{x})})$ yields

$$E\left[n^{-1}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{X}_i - \boldsymbol{x})\boldsymbol{X}^*_{i(\boldsymbol{x})}\boldsymbol{X}^{*\prime}_{i(\boldsymbol{x})}\right]$$

$$= \int |\boldsymbol{H}|^{-1/2}K\Big(\boldsymbol{H}^{-1/2}(\boldsymbol{u}-\boldsymbol{x})\Big)(\boldsymbol{H}^{-1/2}(\boldsymbol{u}-\boldsymbol{x}))(\boldsymbol{H}^{-1/2}(\boldsymbol{u}-\boldsymbol{x}))'f(\boldsymbol{u})d\boldsymbol{u}$$

$$= \int K(\boldsymbol{v})\boldsymbol{v}\boldsymbol{v}'f(\boldsymbol{H}^{1/2}\boldsymbol{v}+\boldsymbol{x})d\boldsymbol{v}$$

$$= \int \boldsymbol{v}\boldsymbol{v}'K(\boldsymbol{v})\Big(f(\boldsymbol{x}) + [f(\boldsymbol{H}^{1/2}\boldsymbol{v}+\boldsymbol{x}) - f(\boldsymbol{x})]\Big)d\boldsymbol{v}$$

$$= f(\boldsymbol{x})\mu_2(K)\boldsymbol{I}_2(1+o(1)).$$

These results indicate that by the law of large numbers

$$n^{-1}(\boldsymbol{X}'_{(\boldsymbol{x})}\boldsymbol{W}_{(\boldsymbol{x})}\boldsymbol{X}_{(\boldsymbol{x})}) \xrightarrow{p} \begin{bmatrix} f(\boldsymbol{x}) & \boldsymbol{0}' \\ \boldsymbol{0} & f(\boldsymbol{x})\mu_2(K)\boldsymbol{I}_2 \end{bmatrix}.$$

By assumptions (a(i)) and (e) the scalar in the (1,1) entry and the diagonals of the matrix in the (2,2) entry are bounded away from zero. Hence, by these arguments, the result stated in Lemma 1 holds for a fixed $\boldsymbol{x}$. By applying the argument used in Truong and Stone (1992) and using Hoeffding's (1963) inequality the result holds uniformly for $x \in C_1$. $\square$

Lemma 2. Let assumptions (a(i)), (d) and (e) hold and suppose $n^{-1}|H|^{-1/2}$ and each entry of H tends to zero as $n \to \infty$ with H remaining symmetric and positive definite. Then there are positive constants c_2 and $c_{2'}$ such that

$$\lim_n P\left[\sum_{i=1}^{n}[K^*(X_i;x)]^2 \le c_2 n^{-1}|H|^{-1/2},\ \max_i|K^*(X_i;x)| \le c_{2'}n^{-1}|H|^{-1/2} \text{ for } x \in C_1\right] = 1.$$

Proof. For the first assertion, we note that

$$\sum_{i=1}^{n}[K^*(X_i;x)]^2$$

$$= e_1'(X'(x)W(x)X(x))^{-1}X'(x)W^2(x)X(x)(X'(x)W(x)X(x))^{-1}e_1$$

$$\le e_1'(X'(x)W(x)X(x))^{-1}\Big(|H|^{-1/2}\max_v K(v)\Big)X'(x)W(x)X(x)(X'(x)W(x)X(x))^{-1}e_1$$

$$= e_1'(X'(x)W(x)X(x))^{-1}\Big(|H|^{-1/2}\max_v K(v)\Big)I_3 e_1 \tag{5.2}$$

$= ne_1'(X'(x)W(x)X(x))^{-1}\Big(n^{-1}|H|^{-1/2}\max_v K(v)\Big)I_3e_1$. By Lemma 1 the leading term is bounded by a constant and K is bounded by 1. Hence, (5.2) $= O_p(n^{-1}|H|^{-1/2})$ uniformly in x. For the second assertion, we note that by Lemma 1

$$\max_i|K^*(X_i;x)|$$

$$= \max_i|e_1'(X'(x)W(x)X(x))^{-1}(1\ \ H^{-1/2}(X_i - x)')'|H|^{-1/2}K(H^{-1/2}(X_i - x))|$$

$$\leq n^{-1}c_1 e_1' I_3 \max_i |(1 \;\; H^{-1/2}(X_i - x)')' |H|^{-1/2} K(H^{-1/2}(X_i - x))|. \tag{5.3}$$

In the support of K, $(X_i - x)$ is bounded by $\max\{H\}^{1/2}\mathbf{1}$, where $\mathbf{1} = (1,1)'$. Thus (5.3) is

$$\leq n^{-1}c_1 e_1' I_3 (1 \;\; H^{-1/2}\max\{H\}^{1/2}\mathbf{1}')' |H|^{-1/2} K(H^{-1/2}\max\{H\}^{1/2}\mathbf{1}). \tag{5.4}$$

K is bounded by 1, thus (5.4) is

$$\leq n^{-1}c_1 e_1' I_3 (1 \;\; H^{-1/2}\max\{H\}^{1/2}\mathbf{1}')' |H|^{-1/2} = c_1 n^{-1}|H|^{-1/2}. \;\square$$

Lemma 3. Let assumptions (a), (b(*ii*)), (d) and (e) hold and suppose that $n^{-1}|H|^{-1/2}$ and each entry of H tends to zero as $n \to \infty$ with H remaining symmetric and positive definite. Then

$$\| K^*(B - U) \|^2 = O_p(|H|^{-1/2}).$$

Proof. $E\Big(\| K^*(B-U) \|^2 \mid X_1, \ldots, X_n \Big) = \mathrm{tr}[K^{*\prime}K^*]\sigma_B^2 = O_p(|H|^{-1/2})$, by Lemma 2. Thus, by Chebyshev's inequality the result follows. $\square$

Before we state the result of the next lemma we note that under assumption (a(i)) as $n^{-1}|H|^{-1/2}$ and each entry of H tends to zero as $n \to \infty$, then there are positive constants c_3 and $c_{3'}$ such that

$$\lim_n P\Big(c_3 \max\{H^{1/2}\} \leq \tfrac{1}{n}\sum_{i=1}^{n} \mathrm{I}(|X_i - x| \leq \max\{H^{1/2}\}) \leq c_{3'} \max\{H^{1/2}\} \text{ for } x \in C_1 \Big) = 1, \tag{5.5}$$

where $\mathrm{I}(\cdot)$ is the indicator function.

Lemma 4. Let assumptions (a), (b(*ii*)), (c(*ii*)), (d) and (e) hold and suppose that $n^{-1}|\boldsymbol{H}|^{-1/2}$ and each entry of $\boldsymbol{H}$ tends to zero as $n \to \infty$ with $\boldsymbol{H}$ remaining symmetric and positive definite. Then

$$\tfrac{1}{n}(\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}}) \xrightarrow{p} \sigma_B^2.$$

Proof. Expanding $\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}}$ we obtain

$$\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}} = \boldsymbol{B}'(\boldsymbol{I}-\boldsymbol{K}^*)'(\boldsymbol{I}-\boldsymbol{K}^*)\boldsymbol{B} = [\boldsymbol{B}-\widehat{\boldsymbol{U}}]' \cdot [\boldsymbol{B}-\widehat{\boldsymbol{U}}] .$$

Hence, we can re-express $n^{-1}\widetilde{\boldsymbol{B}}'\widetilde{\boldsymbol{B}}$ as

$$n^{-1}\sum_{i=1}^{n}\left(B_i - \hat{u}(\boldsymbol{X}_i)\right)^2$$

$$= n^{-1}\sum_{i=1}^{n}\left[\left(B_i - u(\boldsymbol{X}_i)\right) + \left(\hat{u}(\boldsymbol{X}_i) - u(\boldsymbol{X}_i)\right)\right]^2$$

$$= \tfrac{1}{n}\sum_{i=1}^{n}\left(B_i - u(\boldsymbol{X}_i)\right)^2 - \tfrac{2}{n}\sum_{i=1}^{n}\left(B_i - u(\boldsymbol{X}_i)\right)\left(\hat{u}(\boldsymbol{X}_i) - u(\boldsymbol{X}_i)\right) + \tfrac{1}{n}\sum_{i=1}^{n}\left(\hat{u}(\boldsymbol{X}_i) - u(\boldsymbol{X}_i)\right)^2. \tag{5.6}$$

By assumptions (a(ii)) and (b(ii)) and the law of large numbers, the first term converges to σ_B^2, which is bounded for all $\boldsymbol{x} \in C_1$. For the third term we use the following argument. Recall that $\hat{u}(\boldsymbol{X}_i) = \sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)B_j$. Define

$u^*(\boldsymbol{X}_j) = u(\boldsymbol{X}_i) + [\boldsymbol{H}^{1/2}(\boldsymbol{X}_j - \boldsymbol{X}_i)]' D_u(\boldsymbol{X}_i)$. Notice that

$\sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)u^*(\boldsymbol{X}_j) = u(\boldsymbol{X}_i)$. Thus,

$\hat{u}(\boldsymbol{X}_i) - u(\boldsymbol{X}_i)$

$$= \widehat{u}(\boldsymbol{X}_i) - \sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i) u^*(\boldsymbol{X}_j)$$

$$= \sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i) B_j - \sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i) u^*(\boldsymbol{X}_j) = \sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[B_j - u^*(\boldsymbol{X}_j)]$$

$$= \sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[B_j - u(\boldsymbol{X}_j)] + \sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[u(\boldsymbol{X}_j) - u^*(\boldsymbol{X}_j)]. \tag{5.7}$$

Squaring (5.7) we get that $n^{-1}\sum_{i=1}^{n}\left(\widehat{u}(\boldsymbol{X}_i) - u(\boldsymbol{X}_i)\right)^2$ is equal to

$$n^{-1}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[B_j - u(\boldsymbol{X}_j)]\right]^2$$

$$+ 2n^{-1}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[B_j - u(\boldsymbol{X}_j)]\sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[u(\boldsymbol{X}_j) - u^*(\boldsymbol{X}_j)]\right]$$

$$+ n^{-1}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[u(\boldsymbol{X}_j) - u^*(\boldsymbol{X}_j)]\right]^2,$$ which is bounded by

$$2n^{-1}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[B_j - u(\boldsymbol{X}_j)]\right]^2 + 2n^{-1}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[u(\boldsymbol{X}_j) - u^*(\boldsymbol{X}_j)]\right]^2.$$

The first term converges to zero by Lemma 3. For the second term, recall that in the support of K, $(\boldsymbol{X}_j - \boldsymbol{X}_i)$ is bounded by $\max\{H^{1/2}\}\mathbf{1}$. By using a Taylor's expansion of $u(\boldsymbol{X}_j)$ around $\boldsymbol{X}_i$, i.e.

$u(\boldsymbol{X}_j) = u(\boldsymbol{X}_i) + (\boldsymbol{X}_j - \boldsymbol{X}_i)' D_u(\boldsymbol{X}_i) + (\boldsymbol{X}_j - \boldsymbol{X}_i)' \Gamma_u(\xi_i)(\boldsymbol{X}_j - \boldsymbol{X}_i)$ and Lemma 2,

$$\sum_{j=1}^{n} K^*(\boldsymbol{X}_j; \boldsymbol{X}_i)[u(\boldsymbol{X}_j) - u^*(\boldsymbol{X}_j)]$$

$$\leq \sum_{\substack{j \\ |X_i - X_j| \leq \max\{H^{1/2}\}}} c_{2'} n^{-1} |H|^{-1/2} \left[\| \Gamma_u(\xi_i) \| \max\{H\} \right]$$

$$\leq c_4 |H|^{-1/2} \max\{H\} \tfrac{1}{n} \sum_{i=1}^{n} \mathrm{I}(|X_i - x| \leq \max\{H^{1/2}\}). \tag{5.8}$$

By (5.5), (5.8) is $O_p\left(|H|^{-1/2}\max\{H\}^2\right)$. Using the bandwidth matrix that minimizes the integrated asymptotic mean squared error (optimal bandwidth) implies $\max\{H\} \sim n^{-1/3} \Rightarrow |H| \sim n^{-2/3} \Rightarrow |H|^{1/2} \sim \max\{H\}$, hence $O_p\left(|H|^{-1/2}\max\{H\}^2\right) = O_p(\max\{H\})$. Thus,

$2n^{-1}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} K^*(X_j; X_i)[u(X_j) - u^*(X_j)]\right]^2 = O_p(\max\{H\}^2)$, which takes care of the third term in (5.6). For the second term in (5.6), note that by the Cauchy-Schwartz inequality

$$n^{-1}\sum_{i=1}^{n}\left(B_i - u(X_i)\right)\left(\hat{u}(X_i) - u(X_i)\right)$$

$$\leq \left[n^{-1}\sum_{i=1}^{n}\left(B_i - u(X_i)\right)^2\right]^{1/2} \cdot \left[n^{-1}\sum_{i=1}^{n}\left(\hat{u}(X_i) - u(X_i)\right)^2\right]^{1/2}. \tag{5.9}$$

The first term in (5.9) goes to σ_B, which is finite and the second term is the same as the third term in (5.6), which goes to zero. Hence the only term in (5.6) not tending to zero is the constant σ_B^2. □

Lemma 5. Let assumptions (a(i)), (c(i)), (d) and (e) hold and suppose $n^{-1}|H|^{-1/2}$ and each entry of H tends to zero as $n \to \infty$ with H remaining symmetric and positive definite. Then

$$(I - K^*)M = O_p\big(\max\{H\}\big)\mathbf{1}\ .$$

Proof. We examine the ith element of $(I - K^*)M$ and argue that the result holds for $i = 1,\ldots,n$. Define $m^*(X_j) = m(X_i) + [H^{1/2}(X_j - X_i)]' D_m(X_i)$. Notice that $\sum_{j=1}^{n} K^*(X_j; X_i) m^*(X_j) = m(X_i)$. So the ith element of $(I - K^*)M$, $m(X_i) - \sum_{j=1}^{n} K^*(X_j; X_i) m(X_j)$, can be written as

$$\sum_{j=1}^{n} K^*(X_j; X_i) m^*(X_j) - \sum_{j=1}^{n} K^*(X_j; X_i) m(X_j)$$

$= \sum_{j=1}^{n} K^*(X_j; X_i)[m^*(X_j) - m(X_j)] = O_p(\max\{H\})$, by using Taylor's theorem as in the argument used in the proof to Lemma 4. □

Proof of Theorem 1:

To begin the calculations, we express $\hat{\alpha}$ as $(\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'\widetilde{Y} = (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)Y$.

$$E\{\hat{\alpha} \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\}$$

$$= (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)[\alpha B + M]$$

$$= (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)\alpha B + (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)M$$

$$= (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'\widetilde{B}\alpha + (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)M$$

$$= \alpha + (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)M\ .$$

From Lemma 5, the ith element of $(I - K^*)M = O_p\big(\max\{H\}\big)$. So

$\widetilde{B}'(I - K^*)M$

$= O_p(\max\{H\}) \cdot \sum_{i=1}^{n}\big(B_i - \hat{u}(X_i)\big)$

$= O_p(\max\{H\})\left[\sum_{i=1}^{n}\big(B_i - u(X_i)\big) + \sum_{i=1}^{n}\big(\hat{u}(X_i) - u(X_i)\big)\right]$. It follows from the argument used in the proof of Lemma 4 that

$n^{-1}\sum_{i=1}^{n}\big(\hat{u}(X_i) - u(X_i)\big) = O_p(\max\{H\})$, and by the Central Limit Theorem,

$n^{-1}\sum_{i=1}^{n}\big(B_i - u(X_i)\big)$ is $o_p(n^{-1/2})$. Hence,

$$n^{-1}\widetilde{B}'(I - K^*)M = O_p\big((\max\{H\})^2\big) + o_p(\max\{H\}n^{-1/2}), \tag{5.10}$$

and by Lemma 4 the result follows.

We now consider the variance. We establish a result to be used in the following proof. From the arguments used to show the convergence of the third term in (5.6), it follows that $\frac{1}{n}\| U - K^*B \|^2 = O_p\big((\max\{H\})^2\big) + O_p\big(n^{-1}|H|^{-1/2}\big)$. Thus, there exists a bounding function

$$b(n) = c_5\big(\max\{H\}\big)^2 + c_6(n^{-1}|H|^{-1/2}) \tag{5.11}$$

with finite constants c_5 and c_6 such that $\frac{1}{n}\| U - K^*B \|^2 \le b(n)$ uniformly in n. Expressing $\hat{\alpha}$ as $(\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)Y$ we have

$$\mathrm{var}(\hat{\alpha} \mid X_1,\ldots,X_n,\, B_1,\ldots,B_n) = (\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)V(I - K^*)'\widetilde{B}(\widetilde{B}'\widetilde{B})^{-1},$$

where $V = \sigma^2 I_n$. Thus,

$$\mathrm{var}(\hat{\alpha} \mid X_1,\ldots,X_n,\, B_1,\ldots,B_n)$$

$$= \sigma^2(\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I - K^*)(I - K^*)'\widetilde{B}(\widetilde{B}'\widetilde{B})^{-1}$$

$$= \sigma^2(\widetilde{B}'\widetilde{B})^{-1} + \sigma^2(\widetilde{B}'\widetilde{B})^{-1}[\widetilde{B}'K^{*2}\widetilde{B} - 2\widetilde{B}'K^*\widetilde{B}](\widetilde{B}'\widetilde{B})^{-1}.$$

Using Lemma 4, it will suffice to show that the terms in the brackets are componentwise $o_p(n)$. From Lemma 2, $\| K^* \| \le [\mathrm{tr}(K^{*\prime}K^*)]^{1/2} = O_p(|H|^{-1/4})$. It follows from Lemma 3 and (5.11) that

$$\begin{aligned} \| K^*\widetilde{B} \| &= \| K^*(B - U) + K^*U - K^{*2}B \| \\ &\le \| K^*(B - U) \| + \| K^*(U - K^*B) \| \end{aligned}$$

$$\leq \| K^*(B-U) \| + \| K^* \| \; \| (U - K^*B) \|$$
$$= O_p(|H|^{-1/4}) + O_p(|H|^{-1/4}[nb(n)]^{1/2})$$
$$= O_p[nb(n)] \; .$$

Since using the optimal bandwidth implies $\max\{H\} \sim n^{-1/3} \Rightarrow |H| \sim n^{-2/3}$. Using the results from Lemma 4, $\| \widetilde{B} \| = O_p(n^{1/2})$, hence
$|\widetilde{B}' K^* \widetilde{B}| \leq \| \widetilde{B}' K^* \widetilde{B} \| \leq \| \widetilde{B}' K^* \| \; \| \widetilde{B} \| = O_p(n^{3/2} b(n))$, and
$|\widetilde{B}' K^{*2} \widetilde{B}| \leq \| \widetilde{B}' K^* \| \; \| K^* \widetilde{B} \| = O_p(n^2 b^2(n))$. Using Lemma 4, we have

$$\mathrm{var}(\widehat{\alpha} \mid X_1, \ldots, X_n, B_1, \ldots, B_n) = \frac{\sigma^2}{n\sigma_B^2} + \frac{\sigma^2}{n\sigma_B^2}\Big(O_p(n^{3/2}b(n)) + O_p(n^2 b^2(n))\Big)\frac{\sigma^2}{n\sigma_B^2}$$

$$= \frac{\sigma^2}{n\sigma_B^2} + n^{-1}\Big(O_p(n^{1/2}b(n)) + O_p(nb^2(n))\Big).$$

Since $b(n) \to 0$ and $nb^2(n) \to 0$ as $n|H|^{1/2} \to \infty$ and the elements of $H \to \mathbf{0}$, the last term is $o_p(n^{-1})$. □

Proof of Theorem 2:

We begin by expressing $\widehat{M}$ as

$$K^*(Y - \widehat{\alpha}B) = K^*[(Y - \alpha B) + (\widehat{\alpha} - \alpha)B] \; . \tag{5.12}$$

Taking the conditional expectation of the difference between (5.12) and M we have

$$E\Big(K^*[(Y - \alpha B) + (\widehat{\alpha} - \alpha)B] - M \mid X_1, \ldots, X_n, B_1, \ldots, B_n\Big)$$

$$= E\Big(K^*(Y-\alpha B) - M + K^*(\hat{\alpha}-\alpha)B \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big)$$

$$= E\Big(K^*(Y-\alpha B) - M \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big)$$

$$+ K^*BE\Big(\hat{\alpha}-\alpha \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big).$$

$$= E\Big(K^*(Y-\alpha B) - M \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big)$$

$$+ \hat{U}E\Big(\hat{\alpha}-\alpha \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big). \tag{5.13}$$

By assumption, u is a continuous function with bounded support, thus bounded. Hence, $\hat{u}$ is bounded. So by the results of Theorem 1, (5.13) is

$$= K^*((M+\alpha B) - \alpha B) - M + O_p\Big((\max\{H\})^2\Big)\mathbf{1}$$

$$= (K^*-I)M + O_p\Big((\max\{H\})^2\Big)\mathbf{1}, \tag{5.14}$$

where $\mathbf{1}$ is the n by 1 vector of ones. Applying Lemma 5, the elements of $(K^*-I)M$ are $O_p(\max\{H\})$.

We now look at the variance.

$$\mathrm{var}[K^*[(Y-\alpha B) + (\hat{\alpha}-\alpha)B] \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n]$$

$$= \mathrm{var}[K^*[(Y-\alpha B)] \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n] \tag{5.15a}$$

$$+\operatorname{var}[K^*(\hat{\alpha}-\alpha)B] \mid X_1,\ldots,X_n,\, B_1,\ldots,B_n] \tag{5.15b}$$

$$+\,2\operatorname{cov}[K^*[(Y-\alpha B),\, K^*(\hat{\alpha}-\alpha)B] \mid X_1,\ldots,X_n,\, B_1,\ldots,B_n]\,. \tag{5.15c}$$

By Lemma 2, the elements of (5.15a) are $O_p(n^{-1}|H|^{-1/2})$. The second term, (5.15b), is equal to $\frac{\sigma^2}{n\sigma_B^2}K^*B(K^*B)' = \frac{\sigma^2}{n\sigma_B^2}\widehat{U}\widehat{U}'$. Since $\hat{u}$ is a bounded function, the elements of $\widehat{U}\widehat{U}'$ are bounded. Hence the elements of (5.15b) are $O_p(n^{-1})$. We use the Cauchy-Schwartz inequality for (5.15c) and the results of the second part of Theorem 2 hold. □

Proof of Theorem 3 :

From the results of Theorem 1, we show that

$$\lim_{n\to\infty} P\Big[\sqrt{n^{-1}(\widetilde{B}'\widetilde{B})}\,\frac{(\hat{\alpha}-\alpha)}{\sigma} \le t \mid X_1,\ldots,X_n,\, B_1,\ldots,B_n\Big] = \Phi(t) + o_p(1). \tag{5.16}$$

The two-stage estimation procedure involves the bandwidth matrix $\boldsymbol{H}$, which is dependent on n. Because for each different n we have a different $\boldsymbol{H}$, we also have a different sequence of random variables. In this case, the Central Limit Theorem argument used must apply to double arrays. Chung (1974) shows that if the conditions for the Lindeberg-Feller theorem hold, then the Central Limit Theorem for triangular arrays will also hold. In this proof, we follow the approach of Speckman (1988) and show that the Hájek-Šidak condition [Sen and Singer (1993)] holds, which implies the conditions for the Lindeberg-Feller theorem [Chung

(1974)] hold. We include some details that show that the Hájek-Šidak condition implies the condition for the Lindeberg-Feller theorem [Chung (1974)].

Recall that $\widehat{\alpha}$ is defined as the solution to the minimization problem

$$\min_{\alpha}\left[\sum_{i=1}^{n}\left(\widetilde{Y}_i-\alpha\widetilde{B}_i\right)^2\right],$$

which induces the linear equation $\widetilde{B}'\widetilde{B}\widehat{\alpha}=\widetilde{B}'\widetilde{Y}$. Adding and subtracting $(\widetilde{B}'\widetilde{B})\alpha$ to the left side and rearranging we obtain $\sqrt{n}(\widehat{\alpha}-\alpha)=\sqrt{n}(\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I-K^*)(Y-B\alpha)$. We can re-write the expression on the right hand side as $\sqrt{n}(\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I-K^*)[(Y-B\alpha-M)+M]$. In view of Lemma 4 and (5.10), $n(\widetilde{B}'\widetilde{B})^{-1}\to\sigma^2$ and $n^{-1/2}(\widetilde{B}'\widetilde{B})^{-1}\widetilde{B}'(I-K^*)M=o_p(1)$. Thus, $n^{-1/2}\widetilde{B}'(I-K^*)[(Y-B\alpha-M)+M]$ will have the same asymptotic distribution as $n^{-1/2}\widetilde{B}'(I-K^*)(Y-B\alpha-M)$. Considering the results of Lemma 4 we define $c'=\widetilde{B}'(I-K^*)=(c_1,\ldots,c_n)$, then according to the probabilistic version of the Hájek-Šidak condition, if

$$\frac{\max_i c_i^2}{\sum_{i=1}^{n} c_i^2}\xrightarrow{p}0\,, \tag{5.17}$$

then (5.16) will hold. Expanding $c'c$ we get $\widetilde{B}'(I-K^*)'(I-K^*)\widetilde{B}=\widetilde{B}'(I-2K^*+K^{*\prime}K^*)\widetilde{B}$. From Lemma 4 and the second part of the proof to Theorem 1, $n^{-1}\widetilde{B}'(I-2K^*+K^{*\prime}K^*)\widetilde{B}\xrightarrow{p}\sigma_B^2$. Thus, (5.17) will hold if

$$\max_i c_i^2 = o_p(n). \tag{5.18}$$

Next we use the fact that

$$\begin{aligned}\max_i c_i &= \| \boldsymbol{c} \|_\infty \\ &= \| (\boldsymbol{I}-\boldsymbol{K}^*)'(\boldsymbol{I}-\boldsymbol{K}^*)\boldsymbol{B} \|_\infty \\ &\leq (\| \boldsymbol{I} \|_\infty + \| \boldsymbol{K}^* \|_\infty)'(\| \boldsymbol{I} \|_\infty + \| \boldsymbol{K}^* \|_\infty) \| \boldsymbol{B} \|_\infty \\ &= (1 + \| \boldsymbol{K}^{*\prime} \|_\infty)(1 + \| \boldsymbol{K}^* \|_\infty) \| \boldsymbol{B} \|_\infty \, .\end{aligned}$$

By definition, $\| \boldsymbol{K}^* \|_\infty = 1$. Thus if the components of $\boldsymbol{B}$ are uniformly bounded, it follows that $\| \boldsymbol{c} \|_\infty = O_p(1)$. To show that the components of $\boldsymbol{B}$ are uniformly bounded it suffices to show that

$$\| \boldsymbol{B} - \boldsymbol{U} \|_\infty = o_p(n^{1/2}) \, . \tag{5.19}$$

From assumption (f), the Markov inequality yields:

$$P\Big[\max_i | B_i - u(\boldsymbol{X}_i) | \geq m \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\Big]$$

$$\leq \sum_{i=1}^n P\Big(| B_i - u(\boldsymbol{X}_i) | \geq m \mid \boldsymbol{X}_i\Big) \leq \frac{nC}{m^{2+\delta}} \, ,$$

for any constant m. If we choose m as $\frac{n^{1/2}}{\log n}$ then

$$P(\| \boldsymbol{B} - \boldsymbol{U} \|_\infty \geq \frac{n^{1/2}}{\log n} \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \to 0$$

and (5.19) holds, hence (5.18) holds. The following shows that (5.17) implies that the Lindeberg condition [Chung (1974)] holds. To simplify the presentation of the next part of the proof, we define $\mu_i = B_i\alpha + m(X_i)$. The conditional Lindeberg Condition for the Central Limit Theorem is

$$\frac{1}{s_n^2}\sum_{i=1}^{n} E\Big\{c_i^2(Y_i - \mu_i)^2 I\{|c_i(Y_i - \mu_i)| > \eta s_n\} \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big\} \xrightarrow{P} 0, \tag{5.20}$$

where $s_n^2 = \sum_{i=1}^{n} \mathrm{var}\Big(c_i(Y_i - \mu_i) \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big) = \sigma^2\sum_{i=1}^{n} c_i^2.$

Observe that the left hand side of (5.20) is less than

$$\frac{1}{s_n^2}\sum_{i=1}^{n} c_i^2 E\Big\{(Y_i - \mu_i)^2 I\{|Y_i - \mu_i|^2 > \eta^2 s_n^2/\max_i|c_i|^2\} \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big\}$$

$$= \frac{1}{\sigma^2}E\Big\{(Y_1 - \mu_1)^2 I\{|Y_1 - \mu_1|^2 > \eta^2 s_n^2/\max_i|c_i|^2\} \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\Big\}.$$

Set $\Omega_n = \Big\{\max_i c_i^2/\sum_{i=1}^{n} c_i^2 > \epsilon\Big\}$. We showed that $\lim_n P(\Omega_n) = 0$. Since (5.17) holds, we can state that

$$P\Big\{E[(Y_1 - \mu_1)^2 I\{(Y_1 - \mu_1)^2 > \eta^2\sigma^2/\max_i c_i^2\} \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n] \geq \delta\ ;\ \Omega_n^c\Big\}$$

$$\leq P\Big\{E[(Y_1 - \mu_1)^2 I\{(Y_1 - \mu_1)^2 > \eta^2\sigma^2/\epsilon\} \mid X_1,\ B_1] \geq \delta\ ;\ \Omega_n^c\Big\}$$

$$\leq P\Big\{E[(Y_1 - \mu_1)^2 I\{(Y_1 - \mu_1)^2 > \eta^2\sigma^2/\epsilon\} \mid X_1,\ B_1] \geq \delta\Big\}$$

$$\leq \frac{1}{\delta}\, E[(Y_1 - \mu_1)^2 I\{(Y_1 - \mu_1)^2 > \eta^2\sigma^2/\epsilon\}] \tag{5.21}$$

The last inequality follows from the Markov inequality. Since $\lim_n P(\Omega_n) = 0$, we can make $\eta^2\sigma^2/\epsilon$ arbitrarily large by choosing large enough n. By assumption $E(Y_1 - \mu_1)^2$ is finite, therefore (5.21) $\to 0$ as $\eta^2\sigma^2/\epsilon \to \infty$ by Lebesgue's Dominated Convergence theorem, thus (5.20) holds.

Proof of Theorem 4 :

Recall that $\widehat{m}_o(\boldsymbol{x}) \equiv a$ where $[a, \boldsymbol{b}']$ minimize

$$\sum_{i=1}^{n}\Big((Y_i - \alpha B_i) - a - \boldsymbol{b}'(\boldsymbol{X}_i - \boldsymbol{x})\Big)^2 K_H(\boldsymbol{X}_i - \boldsymbol{x}) \ . \tag{5.22}$$

For (i) recall that $\widehat{m}(\boldsymbol{x}) = e_1'(\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})\boldsymbol{X}(\boldsymbol{x}))^{-1}\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})[\boldsymbol{Y} - \widehat{\alpha}\boldsymbol{B}]$
$= e_1'(\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})\boldsymbol{X}(\boldsymbol{x}))^{-1}\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})[\boldsymbol{Y} - \alpha\boldsymbol{B}] + (\alpha - \widehat{\alpha})e_1'(\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})\boldsymbol{X}(\boldsymbol{x}))^{-1}\boldsymbol{X}'(\boldsymbol{x})\boldsymbol{W}(\boldsymbol{x})\boldsymbol{B}.$
Thus, $E(\widehat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n)$
$= E(\widehat{m}_o(\boldsymbol{x}) - m(\boldsymbol{x}) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n) + \widehat{u}(\boldsymbol{x})E(\widehat{\alpha} - \alpha \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n)$
$= E(\widehat{m}_o(\boldsymbol{x}) - m(\boldsymbol{x}) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n)(1 + o_p(1))$, since $\widehat{u}$ is bounded and using the results of Theorem 1. For (ii),
$\mathrm{var}(\widehat{m}(\boldsymbol{x}) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n) = \mathrm{var}(\widehat{m}_o(\boldsymbol{x}) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n)$
$+ \widehat{u}^2(\boldsymbol{x})\mathrm{var}(\widehat{\alpha} \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n) + \widehat{u}(\boldsymbol{x})\mathrm{cov}(\widehat{m}_o(\boldsymbol{x}),\ \widehat{\alpha} \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n).$
Since $\widehat{u}$ is bounded, $\widehat{u}^2$ is also bounded. Therefore, using the results of the second part of Theorem 1 and the Cauchy-Schwartz inequality,
$\mathrm{var}(\widehat{m}(\boldsymbol{x}) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n, B_1,\ldots,B_n)$
$= \mathrm{var}(\widehat{m}_o(\boldsymbol{x}) \mid \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,\ B_1,\ldots,B_n)(1 + o_p(1)).$

To show asymptotic normality for $\widehat{m}(x)$, we must first show asymptotic normality for $\widehat{m}_o(x)$. The solution to the set of linear equations induced by (5.22) can be expressed as

$$X'(x)W(x)Y^* - X'(x)W(x)X(x)\widehat{\beta} = \mathbf{0}\ , \tag{5.23}$$

where $Y^* = Y - \alpha B$. Recall $a_o \equiv a_o(x)$ minimizes $E\left((Y^* - a)^2 \mid X_1,\ldots,X_n,B_1,\ldots,B_n\right)$ and $b_o \equiv b_o(x) = \left[\frac{\partial a_o(x)}{\partial x_1}, \frac{\partial a_o(x)}{\partial x_2}\right]'$. It follows from Lemma 1 of Fan, Hu and Truong (1991) that $a_o - \widehat{a} = o_p(1)$ and $b_o - \widehat{b} = o_p(1)$. Adding and subtracting $\beta_o = [a_o, b_o]'$ to left side of (5.23) yields

$$X'(x)W(x)Y^* - X'(x)W(x)X(x)(\widehat{\beta} - \beta_o) - X'(x)W(x)X(x)\beta_o = \mathbf{0}\ . \tag{5.24}$$

Then $\widehat{\beta} - \beta_o = (X'(x)W(x)X(x))^{-1}(X'(x)W(x)Y^* - X'(x)W(x)X(x)\beta_o)$, and

$$\widehat{a} - a_o = e_1'(X'(x)W(x)X(x))^{-1}(X'(x)W(x)Y^* - X'(x)W(x)X(x)\beta_o).$$

Let $S(x) = X'(x)W(x)Y^* - X'(x)W(x)X(x)\beta_o = X'(x)W(x)(Y^* - X(x)\beta_o)$. Then

$$\widehat{a} - a_o = e_1'(X'(x)W(x)X(x))^{-1}S(x)\ .$$

We can express $e_1'(X'(x)W(x)X(x))^{-1}S(x)$ as

$$\sum_{i=1}^{n}\left(Y_i^* - a_o - b_o'(X_i - x)\right)K^*(X_i;x)\ .$$

To show asymptotic normality for $\widehat{m}_o(x)$, it suffices to show that the conditional version of Liapounov's condition holds

$$\rho = \frac{\sum_{i=1}^{n} E\left(\left| \left(Y_i^* - a_o - b_o'(X_i - x)\right)K^*(X_i;x) \right|^{2+\delta} \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\right)}{\left[\operatorname{var}\left(e_1'(X'(x)W(x)X(x))^{-1}S(x) \mid X_1,\ldots,X_n,\ B_1,\ldots,B_n\right)^{1/2}\right]^{2+\delta}} \to 0\ ,$$

for some $\delta > 0$. This results from the second part of Theorem 2 and the following.

$$\sum_{i=1}^{n} E\left| \left(Y_i^* - a_o - b_o'(X_i - x)\right)K^*(X_i;x) \right|^{2+\delta}$$

$$= n\int\int\int \left| (y^* - a_o - b_o'(u - x))K^*(u;x) \right|^{2+\delta} f(y,u,b)\,dy\,du\,db$$

$$= n\int\int\int \left| y^* - a_o - b_o'(u - x) \right|^{2+\delta} K^*(u;x)^{2+\delta} f(y,u,b)\,dy\,du\,db$$

$$= n\int\int\int \left| y^* - a_o - b_o'(u - x) \right|^{2+\delta} f(y \mid u,b)\,dy\,K^*(u;x)^{2+\delta} f(u,b)\,du\,db$$

$$= n\int\int\int \left| y^* - a_o - b_o'H^{1/2}v \right|^{2+\delta} f(y \mid H^{1/2}v + x, b)\,dy$$

$$\cdot K^*(H^{1/2}v + x;x)^{2+\delta} f(H^{1/2}v + x)\,dv\,db|H|^{1/2}$$

$$\leq n\int\int\int \left| y^* - a_o - b_o'H^{1/2}v \right|^{2+\delta} f(y \mid H^{1/2}v + x, b)\,dy$$

$$\cdot (c_{2'}n^{-1}|H|^{-1/2})^{2+\delta} f(H^{1/2}v + x)dvdb|H|^{1/2}, \text{ by Lemma 2}$$

$$= O\left(n^{-(1+\delta)}|H|^{-\frac{1+\delta}{2}}\right), \text{ by assumption (g). This implies that}$$

$$\sum_{i=1}^{n} E\left(|\left(Y_i^* - a_o - b_o'(X_i - x)\right)K^*(X_i;x)|^{2+\delta} | X_1,\ldots,X_n, B_1,\ldots,B_n\right)$$

$$= O_p\left[n^{-(1+\delta)}|H|^{-\frac{1+\delta}{2}}\right].$$

By the results of Theorem 2,

$$\mathrm{var}\left(e_1'(X'(x)W(x)X(x))^{-1}S(x) \mid X_1,\ldots,X_n, B_1,\ldots,B_n\right) = O_p(n^{-1}|H|^{-1/2}),$$

which implies

$$\left[\mathrm{var}\left(e_1'(X'(x)W(x)X(x))^{-1}S(x) \mid X_1,\ldots,X_n, B_1,\ldots,B_n\right)^{1/2}\right]^{2+\delta} = O_p\left[n^{-(1+\delta/2)}|H|^{-\frac{2+\delta}{4}}\right].$$

Therefore,

$$\rho = \frac{O_p\left[n^{-(1+\delta)}|H|^{-\frac{1+\delta}{2}}\right]}{O_p\left[n^{-(1+\delta/2)}|H|^{-\frac{2+\delta}{4}}\right]} = O_p\left[\left(n^{-1}|H|^{-1/2}\right)^{\delta/2}\right],$$

hence asymptotic normality holds. By the results of Theorem 3,

$$\widehat{m}(\boldsymbol{x}) - \widehat{m}_o(\boldsymbol{x}) = \sum_{i=1}^{n}\Big(Y_i - \widehat{\alpha}B_i\Big)K^*(\boldsymbol{X}_i;\boldsymbol{x}) - \sum_{i=1}^{n}\Big(Y_i - \alpha B_i\Big)K^*(\boldsymbol{X}_i;\boldsymbol{x})$$

$= (\widehat{\alpha} - \alpha)\sum_{i=1}^{n} B_i K^*(\boldsymbol{X}_i;\boldsymbol{x}) = (\widehat{\alpha} - \alpha)\widehat{u}(\boldsymbol{x}) = O_p(n^{-1/2})$. Note that $n \gg n|\boldsymbol{H}|^{1/2}$; hence, the last result of Theorem 4, (*iii*), follows from (*i*), (*ii*) and the asymptotic normality for $\widehat{m}_o(\boldsymbol{x})$. □

Remarks:

i) The conditional distribution of Y_i given $(B_i, \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ depends on $B_i, \boldsymbol{X}_{i1}$ and $\boldsymbol{X}_{i2}$, and the Y_i are independent but not necessarily identically distributed.

iia) Suppose the conditional distribution of Y_i given $(B_i, \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ is specified by

$$Y_i = m(B_i, \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}) + \epsilon_i,$$

where $m(\cdot)$ is an arbitrary smooth function and the ϵ_i are independent and identically distributed with mean zero and variance σ^2 (or higher moment if necessary). Moreover, ϵ_i is independent of $(B_i, \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ for all i.

b) The conditional distribution of B_i given $(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ is specified by

$$B_i = u(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}) + \eta_i,$$

where $u(\cdot)$ is an arbitrary smooth function and the η_i are independent and identically distributed with mean zero and variance σ_B^2. Also, η_i is independent of $(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ for all i.

c) The joint distribution of the $(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ is concentrated on the unit square.

d) The joint distribution of $(Y_i, B_i, \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ is the product of the above conditional distributions. Note that the $m(\cdot)$ and $u(\cdot)$ are arbitrary smooth (bounded) functions and the distributions of ϵ_i and η_i are arbitrary. Therefore, there are an infinite number of multivariate distributions that satisfy our conditions.

iii) An important topic of future research already underway at the time of this writing is to relax the assumption of homogeneity of variance. A few technical details are required to relax assumption (*b*) to where the conditional variance of B_i given $(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ is a function of the realized values of $\boldsymbol{X}_{i1}$ and $\boldsymbol{X}_{i2}$. The same can be done for the conditional variance of Y_i given $(B_i, \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$.

iv) A bandwidth matrix can be termed optimal in the sense that it minimizes the Asymptotic Mean Integrated Squared Error. Using the Optimal Bandwidth implies that $\max\{\boldsymbol{H}\} = O(n^{-1/3})$, which implies $|\boldsymbol{H}| = O(n^{-2/3})$.

CHAPTER VI: PRACTICAL PERFORMANCE

6.1 The Design:

We conducted a simulation experiment to evaluate the estimation of the components of (1.2) for fixed samples sizes, design densities, and various points of the curve. Such a study is necessarily restrictive since we had to choose from an infinite set of possible combinations of the regression function, design density, error density, sample size, and noise level. However, we feel that our results give us useful insights into the practical performance of our estimators.

For our study, we specified the standard deviation of the errors at $\sqrt{13.7}$, which was approximately $\frac{1}{2}(\max m - \min m)$. The ith elements of the (X_1, X_2) vector were chosen to be independent uniform deviates on [0,1], and independent beta deviates with parameters 0.95 and 0.9, respectively. The parameters of the beta distribution were chosen to resemble the marginal distribution of the compliance covariate. The B_i's were generated as $1 + \eta_i$, where η_i was generated from a Gaussian distribution with zero mean and variance equal to 13.7. The linear component of (1.2) was set at 0.6 and the smooth component at:

$$m(x_1, x_2) = 1 - 48(x_1 + x_2) + 218(x_1^2 + x_2^2) - 315(x_1^3 + x_2^3) + 145(x_1^4 + x_2^4). \quad (6.1)$$

We generated the errors as Gaussian throughout. Figure 4 displays a marginal plot of (6.1) with X_2 fixed at 0.5. To study the mean squared error properties of the estimators, we chose three points of the curve to estimate: $\max\big(m(x_1, x_2)\big)$, $\min\big(m(x_1, x_2)\big)$ and the first inflection point. We set the sample size equal to 100, 500 and 2500. For the smoothing matrix we chose $\boldsymbol{H}_1$ where the diagonal elements of $\boldsymbol{H}_1$ were the asymptotically optimal bandwidth, and the average squared error (ASE) optimal bandwith. The ASE optimal bandwidth was defined as

$$\min_h \left[n^{-1} \sum_{i=1}^{n} \left(\hat{m}(X_{i1}, X_{i2}, h) - m(X_{i1}, X_{i2}) \right)^2 \right]. \tag{6.2}$$

To estimate (6.2), we fit the model over a grid of bandwidths equally spaced from $\log(\frac{1}{3}h_{opt})$ to $\log(3h_{opt})$, then chose the bandwidth that minimized the ASE. Hence, each sample produced it's own ASE optimal bandwidth. We experimented with an additional definition of the ASE optimal bandwidth:

$$\min_h \left[n^{-1} \sum_{i=1}^{n} \left(\hat{\alpha}(h)B_i + \hat{m}(X_{i1}, X_{i2}, h) - \alpha B_i - m(X_{i1}, X_{i2}) \right)^2 \right]. \tag{6.3}$$

However, preliminary results were essentially indistinguishable from those using (6.2), as expected from the results in Chapter V.

Figure 5 shows the geometrically increasing structure of the grid spacing between possible bandwidths. The derivation of the asymptotically optimal bandwidth for the case where H_1 follows in the next section.

6.2 Optimal Bandwidth

In the design of this experiment, we use bandwidth matrix $\boldsymbol{H}_1$. In the following arguments, expectation is taken with respect to the distribution of Y conditional upon the baseline measures, $\boldsymbol{B}$, and the compliance measures, $\boldsymbol{X}$. Let $\theta(h_1, \boldsymbol{x}) = E[\alpha b + m(\boldsymbol{x}) - \hat{\alpha} b - \hat{m}(\boldsymbol{x})]^2$

$$= \mathrm{var}[\widehat{\alpha}b + \widehat{m}(\boldsymbol{x})] + [\mathrm{bias}(\widehat{\alpha}b + \widehat{m}(\boldsymbol{x}))]^2$$

$$= b^2\mathrm{var}\widehat{\alpha} + \mathrm{var}\widehat{m}(\boldsymbol{x}) + 2b\mathrm{cov}(\widehat{\alpha},\widehat{m}(\boldsymbol{x})) + [b\mathrm{bias}\widehat{\alpha} + \mathrm{bias}\widehat{m}(\boldsymbol{x})]^2$$

$$= b^2\mathrm{var}\widehat{\alpha} + \mathrm{var}\widehat{m}(\boldsymbol{x}) + 2b\mathrm{cov}(\widehat{\alpha},\widehat{m}(\boldsymbol{x}))$$

$$+ b^2(\mathrm{bias}\widehat{\alpha})^2 + 2b\mathrm{bias}\widehat{\alpha}\mathrm{bias}\widehat{m}(\boldsymbol{x}) + (\mathrm{bias}\widehat{m}(\boldsymbol{x}))^2.$$

From the results of Theorems 1 and 2, define

$$\mathrm{bias}\widehat{m}(\boldsymbol{x}) = \tfrac{h_1^2}{2}\sum_{k=1}^{2}\frac{\partial^2 m(\boldsymbol{x})}{\partial x_k^2}\, C_2(K) = \tfrac{h_1^2}{2}D_1(\boldsymbol{x})C_2(K)$$

and

$$\mathrm{var}\widehat{m}(\boldsymbol{x}) = [nh_1^2 f(\boldsymbol{x})]^{-1}\sigma^2 C_1(K)\ ,$$

where $C_1(K) = \int K^2(v)dv$, and $C_2(K) = \int v^2K(v)dv$.

$$\mathrm{bias}\widehat{\alpha} = \tfrac{h_1^4}{4}\left[\sum_{k=1}^{2}\frac{\partial^2 m(\boldsymbol{x})}{\partial x_k^2}\right]\left[\sum_{k=1}^{2}\frac{\partial^2 u(\boldsymbol{x})}{\partial x_k^2}\right]C_2^2(K)\sigma_B^{-2} = \tfrac{h_1^4}{4}D_1(\boldsymbol{x})D_2(\boldsymbol{x})C_2^2(K)\sigma_B^{-2}$$

and

$$\mathrm{var}\widehat{\alpha} = \tfrac{1}{n}\left[\frac{\sigma^2}{\sigma_B^2}\right].$$

Thus, it remains to calculate $\mathrm{cov}(\widehat{\alpha},\widehat{m}(\boldsymbol{x}))$ in order to have all of the elements of the asymptotic $\theta(h_1,\boldsymbol{x})$. From the results of Chapter 5, we know that the MSE of $\widehat{\alpha}$ is dominated by the variance, which is of order root n. Therefore, the components of $\theta(h_1,\boldsymbol{x})$ that are multiplied by the bias or variance of $\widehat{\alpha}$ have a faster rate of convergence than the components of $\theta(h_1, x)$ that are functions of the bias and variance of $\widehat{m}(x)$ alone. Furthermore, by the Cauchy-Schwartz inequality cov($\widehat{\alpha}$,$\widehat{m}(x)$) will have a faster rate of convergence than the components of $\theta(h_1, x)$ that are functions of the bias and variance of $\widehat{m}(x)$ alone. Thus, asymptotically $\theta(h_1, x)$ will be dominated by the bias^2 and variance of $\widehat{m}(x)$. This means that the result for the asymptotically optimal bandwidth for the partially linear model using locally linear regression is the same as estimating without the linear coefficient in the model. Therefore, we can write $\theta(h_1, x)$ as

$$[nh_1^2 f(x)]^{-1}\sigma^2(x)C_1(K) + \frac{h_1^4}{4}D_1^2(x)C_2^2(K) + O(h_1^6) + O(h_1^8) + o_p(n^{-1}).$$

Taking the derivative of $\theta(h_1, x)$with respect to h_1 we obtain

$$\theta'(h_1, x) \approx -2[nh_1^3 f(x)]^{-1}\sigma^2 C_1(K) + h_1^3 D_1^2(x)C_2^2(K).$$

Integrating over the distribution of x we obtain

$$\int \theta'(h_1, x)dxdb = -2\frac{C_1(K)}{nh^3}\int \sigma^2 f^{-1}(x)dx + h_1^3 C_2^2(K)\int D_1^2(x)dx.$$

Setting this equal to zero and solving for h_1 yields

$$h_1^{opt} = \left[\frac{2C_1(K)\int \sigma^2 f^{-1}(x)dx}{C_2^2(K)\int D_1^2(x)dx}\right]^{1/6} n^{-1/6}.$$

6.3 Computational Issues:

As detailed in Chapter III, the calculation of $\hat{\alpha}$ and $\hat{m}$ in (1.2) involved a two stage procedure: First separately estimating α from the residuals after removing the estimated "dependence" of the B_i's and the Y_i's on the (X_{i1}, X_{i2})'s, then estimating m from the residuals of the $\hat{\alpha}B_i$'s subtracted from the Y_i's. Using a direct method to calculate all of the estimates required 3 x $O(n^2)$ number of operations. Thus, the implementation of the estimation has the potential to require a great deal of computer time. For example, using transparent programming in Fortran and compiling utilizing parallel processing on the Convex Supercomputer at University of North Carolina at Chapel Hill, the calculation of one point of the regression surface used 58.88 seconds of cpu time for a sample of size 500. For our experiment, this would have required some of the programs to run for three weeks, which could be a long time or a short time depending on the research situation. For us, this was a long time. So we investigated fast implementations of smoothing techniques.

Fan and Marron (1993) describe two fast implementations of the local linear smoother in the univariate regression setting. The first, called "binning", requires the calculation of the regression estimate over an equally spaced grid. The essential idea involves binning the (X_i, Y_i)'s to obtain gridcounts at each gridpoint. Then at each gridpoint using the gridcounts and a weighted average of the Y_i's to calculate the regression estimate. Taking advantage of the equal spaces between the gridpoints to reduce the number of kernel evaluations reduces the order of operations from $O(n^2)$ to $O(g)$, where g is the number of gridpoints. Our method, however, required us to calculate the regression estimate at the observed data.

Therefore, we could not exploit the advantages of binning over an evenly spaced grid.

The second fast implementation procedure, called "updating", requires one sort of the data on the covariate. For an intuitive understanding of this method consider estimation at an equally spaced grid of points using a uniform kernel. Use the Nadaraya-Watson estimate of data $\{(1,Y_1),\ldots,(n,Y_n)\}$, to be evaluated at the design points $\{1,\ldots,n\}$. For $j=1,\ldots,n$, each $\hat{m}_h(j)$ is an ordinary average of $\{Y_{j-i_h},\ldots,Y_{j+i_h}\}$, for some i_h determined by the bandwidth. After $\hat{m}_h(j)$ has been computed, calculation of $\hat{m}_h(j+1)$ does not require recalculation of the entire average, but only deletion Y_{j-i_h} and insertion of Y_{j+i_h+1}. For evaluation at the entire grid this only requires $O(n)$ operations, compared to $O(n^2h)$ for a direct implementation.

Two immediate problems of this method are that it requires estimation at an equally spaced grid and the use of a uniform kernel, which has graphical drawbacks. However, updating ideas have been adapted for polynomial kernels at nonequally spaced data by the Heidelberg school [Gasser and Kneip (1989)]. The drawback in these developments for us was that the methods were developed for univariate regression which has closed forms for the locally linear estimate, unlike the form for the estimate in the multivariate setting. Therefore, we developed a version of updating based on the moving window ideas that extends to the multiple regression setting and multivariate function estimation in general. The heart of the method involves keeping track of indices on the sorted and unsorted data. In comparison to the calculation using the direct method in the example above, our updating program used .293 seconds of cpu time, which is over 200 times faster. The updating programs allowed us to run all of our simulations over one weekend, in contrast to months using direct methods.

6.4 Results:

Figures 6 - 14 give a graphical summary of the results. In each figure, the plot on the left shows the results for the experiments where the (X_{i1}, X_{i2})'s were chosen from the uniform distribution, and the plot on the right from the beta. Figure 6 shows the results of estimating $m(x_1, x_2)$ at the maximum of m for sample sizes 100 (solid), 500 (dashed) and 2500 (dot-dashed), using the asymptotically optimal bandwidth. The curves are Rosenblätt-Parzen kernel density estimates of the distribution of the local linear estimates of m from the 500 simulated datasets. The solid vertical bar is the value of m at the maximum and the dashed vertical bars represent m plus the asymptotic bias for the respective sample size, calculated using the forms in Chapter V. The farthest dashed vertical bar from the solid vertical bar goes with the smallest sample size. We immediately notice that for all sample sizes, the bias of the estimate is generally smaller than the asymptotic bias. Judging by the visual appearance of the density estimates, there seems to be no difference in the results from the two design densities.

Tables 1 and 2 show monte-carlo estimates of the bias over the 500 samples at the three points of estimation, the three sample sizes, the two design densities and the two bandwidths (h_1 indicating the asymptotically optimal bandwidth, h_2 indicating the ASE optimal bandwidth). Although analytically the bias is not an explicit function of the design density, it is indirectly affected through the bandwidth. The integral of the conditonal variance over the design density appears in the numerator of the asymptotically optimal bandwidth. In our example, the conditional variance was constant. Thus, the inverse of the design density integrated over it's support affected the size of the bandwidth for a fixed sample size. In our example, the asymptotically optimal bandwidth for the uniform density was smaller than for the beta density. Correspondingly, we observed greater bias for the beta density at smaller samples and a more rapid reduction in bias for the beta density as the sample size increased.

Theoretically, for h_1 we should expect a reduction in bias by a factor of 1.31 at each sample size increment. Concentrating on min$m$ and max$m$, when the sample sized was increased from 100 to 500, we observed a reduction in bias of factors ranging from 1.85 to 1.43. When the sample size was increased from 500 to 2500, we observed changes in the bias from an increase by factors ranging from 1.88 to 1.33. The increase in bias at the inflection point may have reflected bias that was so small that monte-carlo variation was dominating. 95% confidence intervals for the monte-carlo estimates of bias at the inflection point at the beta design for sample sizes 100 and 500 were $[-0.264, 0.260]$ and $[-0.097, 0.027]$.

In Figure 6 we can visually confirm the reduction in variance of the locally linear estimate as the sample size increases. Theoretically, for h_1 the standard deviation should halve for each sample size increase. Tables 3 and 4 display the standard deviations of the estimates over the 500 samples for the three points of estimation, the three samples sizes, the two design densities and the two bandwidths. When the sample size was increased from 100 to 500, we observed a reduction in standard deviation of factors ranging from 1.91 to 2.33, the smaller reductions generally occurring at the inflection point. When the sample size was increased from 500 to 2500, we observed reductions of factors ranging from 1.51 to 1.81.

Figure 7 displays the results of locally linear estimation of $m(x_1, x_2)$ at the inflection point. Since the curvature is zero at the inflection point, theoretically we should expect to see no bias in the estimate. The curves in Figure 7 confirm that at every sample size the estimate had very little bias; usually less than one fifth of the size of the bias at the minimum or the maximum. 95% confidence intervals for the monte carlo estimates of bias show no difference from zero. Tables 1 and 2 indicate that the bias when using h_1 was smaller than using h_2. This was expected since the ASE optimal bandwidth was almost always larger than the asymptotically optimal bandwidth.

Figure 8 displays the results of locally linear estimation of $m(x_1, x_2)$ at the minimum. The results are very similar to estimation at the maximum with the sign of the bias reversed.

Figure 9 shows the results for estimating the linear coefficient of (1.2) using the asymptotically optimal bandwidth. As expected with this simulation design, we observed practically no bias in the estimate. Table 7 gives monte-carlo estimates of the bias of the estimate of the linear coefficient for the three sample sizes, the two design densities and the two bandwidths. The results in Table 7 confirm the observations from Figure 9 about the bias of the estimate. Table 8 displays the standard deviation of the estimate of the linear coefficient for the three sample sizes, the two design densities and the two bandwidths. Theoretically, we should expect to see a reduction in variance by a factor of 2.24 for each increment in sample size. When the sample size was increased from 100 to 500 we observed reductions ranging from 2.29 to 2.56. When the sample size was increased from 500 to 2500 we observed reductions ranging from 2.30 to 2.42.

The plots in Figure 10 give the results of selecting the ASE optimal bandwidth for each sample. The curves are density estimates of the distributions of log base three of the ASE optimal bandwidths divided by the asymptotically optimal bandwidth for samples of size 100 (solid), 500 (dashed) and 2500 (dot-dashed). It is immediately clear that the ASE optimal bandwidths were always larger than the asymptotically optimal bandwidth. Since the ASE optimal bandwidths were larger than the asymptotically optimal bandwidths, we should expect to see more bias and less variation in the estimates using the ASE optimal bandwidths.

The results in Tables 1 through 4 confirm that the bias was larger and the standard deviation smaller for the estimates using the ASE optimal bandwith compared with those using the asymptotically optimal bandwidth. Tables 5 and 6 display monte-carlo estimates of the MSE of the locally linear estimate at the three points of estimation, the three sample sizes, the two design densities and the two bandwidths. There is no clear trend in MSE. However, we can state that h_1 did a better job when estimating the minimum, otherwise h_2 usually produced smaller MSE. Usually, we are interested in an estimate of the integrated mean squared error over the entire curve when comparing bandwidth choices. Since we restricted estimation to three points, we cannot make such comparisons; however, these results do suggest that if one is interested in the minimum of the curve, or the maximum, or the inflection point, using different bandwidths may provide better answers than if the same bandwidth were used at each point.

Tables 7 and 8 give monte-carlo estimates of the bias and variance of the linear coefficient estimate for the three sample sizes, the two design densities and the two bandwidths. The results in tables 7 and 8 indicate that there was essentially no difference in the results between the two bandwidths.

The plots in Figures 11 through 13 show the same results as in Figures 6 through 8, however using the ASE optimal bandwidth instead of the asymptotically optimal bandwidth. These plots echo what was noted earlier, that the distribution of estimates using the ASE optimal bandwidths generally reflect higher bias and smaller variance in comparison to Figures 6 through 8. Figure 14 shows the results for estimating the linear coefficient in (1.2) using the ASE optimal bandwidth. Because the bias was so small when using h_1 or h_2, Figures 9 and 14 appear almost identical.

VII. EXAMPLE

In this chapter we describe the process of estimating the components of (1.2) using the LRC data, and we examine the results in light of the clinical questions posed in Chapter I. We eschew any discussion of the medical implications of these results due to a lack of formal training in this area.

7.1 The Model

As Chapter II illustrated, the partially linear model has been examined in the literature; however, software to implement the procedure has not been available for several reasons. The two most obvious ones being the partially linear model has not been estimated using locally linear regression, or with a multivariate smooth function. Therefore, In order to estimate the linear coefficient and the smooth function in (1.2), using the methodology detailed in Chapter III, we had to develop software.

Data analysis is often an iterative procedure that involves fitting a model or several models many times. In our case, estimation of the smoothing matrix, $\boldsymbol{H}$, involves trying several and choosing the one that provides the most pleasing fit to the eye. As the research in this area develops, we hope to have an automated bandwidth matrix selection procedure; but right now, we only have our subjective judgement to assess the quality of the fit. This process is complicated when the dimension is greater than one, as is our case. One way to simplify this situation is to estimate the curve in one direction while holding the values of the covariates in the other directions fixed. As an example of this, Figure 15 displays locally linear estimates of the regression function of the cholesterol, after adjustment for baseline, on the most recent compliance measure while holding the previous compliance measure fixed at 95%. We usually try to undersmooth slightly, as the eye can always smooth out rough features, but cannot recover features that have been smoothed out.

From the set of curves in Figure 15, we selected the one with bandwidth equal to 32. The estimate of α from (1.2) corresponding to this bandwidth was 0.71. We used a bootstrap method, which is discussed later, to calculate a 95% confidence interval of $\widehat{\alpha}$: (0.59, 0.83).

Efron (1991) modelled the treatment arm of the Stanford branch of the LRC data with the quadratic curve:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \ , \ i = 1, \ldots, n \ .$$

with the usual assumptions applied to ϵ_i. Figure 16 displays the locally linear fit from Figure 15 with a quadratic fit overlaid. It is apparent that the bimodal shape of the design density is causing the fit of the quadratic to be such that the shape ignores the behavior between compliance of 0 and 100. By noticing the distance between the locally linear curve and the quadratic we can better understand what we term "model bias."

As mentioned in Chapter II, depending upon the research question, it is sometimes desireable to obtain a confidence belt surrounding the estimated curve. For reasons detailed in Chapter II, we addressed this problem by constructing a set of simultaneous confidence bars over a reasonably dense finite grid. We followed Härdle and Marron's (1991) comment that "for a reasonably dense collection *of error bars* there is little information lost and the bar approach is much easier to compute and also to analyze (italics indicate added text)." We chose the set of design points [0, 10, 20, ... , 100] as the values of compliance from which to construct the confidence set.

As the results of the Theorems in Chapter IV indicate, the variance of the regression function at any point on the curve away from the boundaries is a function of the design density, the bandwidth, the conditional variance and the kernel. Therefore, to estimate the variance, we would have to also estimate the design density and the conditional variance. Alternatively, we constructed a bootstrapped distribution of the curve estimates by resampling the data vectors, $(Y_i,\ B_i,\ X_{i1},\ X_{i2})$, with replacement and re-estimating the curve. Ditka (1988) showed that up to a bias term, this type of bootstrap confidence interval for nonparametric regression is asymptotically correct. Then we assumed that the limiting distribution of the vector $\widehat{\boldsymbol{M}} = [\widehat{m}(95,0),\ \ \widehat{m}(95,10),\ \ldots\ ,\ \widehat{m}(95,100)]'$ was multinormal with mean vector $\boldsymbol{M}$ and covariance matrix Σ. With this we built an ellipsoidal confidence region by approximating the distribution of $n(\widehat{\boldsymbol{M}} - \boldsymbol{M})'\Sigma^{-1}(\widehat{\boldsymbol{M}} - \boldsymbol{M})$ by a chi-square distribution with 11 degrees of freedom. Hence, the $(1-\alpha)$ percent simultaneous confidence bands for $\boldsymbol{M}$ are given by $\widehat{\boldsymbol{M}} \pm (\chi^2_{1-\alpha;11})^{1/2} SE(\widehat{\boldsymbol{M}})$, where $\chi^2_{1-\alpha;11}$ is the $1-\alpha^{th}$ quantile from the chi-square distribution with 11 degrees of freedom. $SE(\widehat{\boldsymbol{M}})$ was approximated using a bootstrapped distribution, as described above. Figure 17 shows the estimate from Figure 15 with simultaneous 90% confidence bars.

7.2 Results:

As mentioned in Chapter I, these results are not intended to address the *efficacy* of cholestyramine, which has already been established. In other words, we did not attempt to construct a dose response curve with respect to cholestyramine for the reason that patients were not randomized to different levels of compliance. Rather, these results address the *usefulness* of the drug therapy, given the compliance behavior of the subjects taking the drug.

Table 9 shows the estimated effect of compliance on mean cholesterol at specific levels of compliance. Table 10 displays the estimated mean cholesterol as a function of the baseline measure and compliance. Baseline has been fixed at 300 milligrams of cholesterol. We observe at zero compliance a reduction in cholesterol of 15 milligrams from baseline, which corresponds to a 5% reduction in cholesterol. This reduction could be due to the cholesterol lowering diet on which the subjects received continuous counseling [LRC-CPPT (1984)]. At 70% compliance, a reduction of 37 milligrams corresponds to a 12.3% reduction in cholesterol. However, at 100% compliance we observe a reduction of 65.3 milligrams, or a 21.8% reduction in cholesterol.

In Chapter I, we initially explored the cholesterol means from the LRC study at nine visits. Although it would not be considered a true multivariate analysis, we can gain insight by applying model (1.2) to each of the nine visits individually. When looking at the cholesterol trend across time as a function of compliance, an important question was, "How much does the compliance history at visits 2 through v affect the cholesterol level at v?" Figure 18 is a collection of kernel estimates of the bivariate density of individual mean compliance through visit $v-1$ and the compliance at visit v. Figure 19 is attached for the purpose of giving the reader the 3-dimensional feel of the bivariate density estimates. It can be seen from the first bivariate density estimate (visit 2 to visit 3) there are two hills of mass at 0% and 100% compliance as well as some mass in-between. This supports the observation that between visit 2 and 3 a shakeout occurred where subjects were finding their respective level of compliance. In the remainder of the bivariate density estimates it can be seen that the subjects remained within their respective level of compliance, near 0% or 100%, for the rest of the observation period. The two hills of mass at 0% and 100% at each visit confirm that those averaging high or low compliance before visit v remained at that high or low level at visit v. These observations led us to believe that except at visit 3, most of the information about compliance is contained in the most recent visit. However, to address some of the clinical questions posed at the beginning, for visit v both compliance at v, X_{i2}, and mean compliance from visits 2 through $v-1$, X_{i1}, were included.

Figures 20(a) and 20(b) show the result of fitting model (1.2) for visits 3 through 9. The value of baseline, B, was fixed at 300 mg of total cholesterol. The values of mean compliance history, X_{i2}, and current compliance, X_{i1}, were both fixed at either 0%, 80% or 100%. The error bars delimit a 90% simultaneous confidence bar for each curve. The idea here was to get a picture of the total cholesterol trend of a typical patient with a baseline level of cholesterol equal to 300 mg while maintaining a consistent level of compliance. The solid curve in figure 20(a) representing 0% compliance is associated with an approximate 3% reduction in cholesterol. The dashed curve representing 80% compliance is associated with an approximate 8.2% reduction in cholesterol, or 5.2% reduction over diet alone. Based on the results of fitting a Cox Proportional Hazards model relating total cholesterol to incidence of coronary heart disease (CHD) the LRC-CPPT study (1984) reported that the benefits that could be expected for a 5.2% reduction in total cholesterol over diet alone was a 13.9% reduction in the relative risk of coronary heart disease, where relative risk is the estimated relative risk of an event in the cholestyramine group compared to the placebo group. The dotted curve in 20(b) representing 100% compliance is associated with an approximate 23% reduction in cholesterol, which would associate with a possible 52.6% reduction in CHD risk over the effect of diet alone.

Figure 20(a) shows a difference between the 0% and 80% compliance trends, but not as much separation as is seen in figure 20(b) between the 80% and the 100% compliance trends. The separation between the 0% and 80% compliance trends occurs at visits 3 and 4. Wheareas the separation between the 80% and 100% trends occurs for all visits and gives strong evidence of a greater reduction in total cholesterol between 80% and 100% compliance than between 0% and 80%. This confirms a feature of the data that was noted previously that the conditional mean trends of cholesterol given compliance were relatively flat until approximately 65%-70% compliance, where the trend began heading downward. This feature was particularly clear at visit 3. More analysis is needed to clarify the implications of this feature.

Figure 21(a) displays two scenarios. The solid line represents the trend in mean cholesterol for a typical patient with a baseline cholesterol of 300 who started at 100% of the dose, then reduced to 75% after one year. The dashed line represents the trend in mean cholesterol of a theoretical patient who maintained 100% compliance throughout the study. The error bars delimit a 90% simultaneous confidence band for each trend. This figure addresses the issue of what happened to mean cholesterol when the patient reduced the level of compliance. It is evident by the confidence bars that the two trends are different, and they separate at visit 4. The reduction in cholesterol for the solid trend over diet alone was approximately 5.2%, which associates with a possible 13.9% reduction in CHD. This provides evidence that within two years of reducing the dose to 75% the mean cholesterol increases and ultimately levels off at approximately the same reduction as that of patients who maintained 80% compliance throughout the entire study.

Figure 21(b) displays the trend in mean cholesterol when the patient started at 75% of the dose and increased to 100% after one year, and the trend in mean cholesterol of a theoretical patient who maintained 100% compliance throughout the study. The solid line could represent the scenario where the patient initially could not tolerate the full dosage because of negative side effects, then increased after one year after a tolerance was built up. The 90% confidence bars around each trend give evidence that the trends are indistinguishable. The reduction in cholesterol for the solid trend over diet alone was approximately 15.9%, which associates with a possible 42.6% reduction in CHD. Figure 21(b) provides evidence that patients who begin at 75% of the dose and increase to 100% after one year achieve very similar results as those who begin at 100% and maintain that level. These results have to be interpreted within the constraints of the summary measure of mean compliance for visits 1 to $v-1$ for the history of compliance at visit v.

Some have argued [discussion of Efron and Feldman (1984)] that comparing means at different levels of compliance is confounded by other factors than the effect of the drug because of the the non-randomized allocation of dose. In other words, the factors contributing to the self-selection of dose may have been the cause of the difference in cholesterol means instead of the effect of the level of drug ingested. Being aware of this possible criticism, the investigators in the LRC-CPPT (1984) study included a compliance type variable in the Cox Proportional Hazards model and concluded that there was no evidence for an effect of compliance on CHD incidence in the cholestyramine group other than that from the cholesterol-lowering effect of the drug. Additionally, this criticism is better directed at attempts to define the effectiveness of the drug. Since when investigating the usefulness of the drug, the non-randomized nature of dose and the factors affecting self-selection do not confound the interpretation of comparing cholesterol means at different levels of compliance.

CHAPTER VIII: SUMMARY AND FUTURE RESEARCH

8.1 Summary:

In this dissertation we attempted to motivate the use of the partially linear model (1.2) for the analysis of the LRC compliance data; and to justify the locally linear estimation of model (1.2) through results from analytic calculations and a simulation experiment.

In Chapter I we motivated model (1.2) by graphically exploring the relationships between the cholesterol variables and the compliance variables. We noted that we wanted to control the model bias as much as possible and to achieve this through nonparametric regression. The relative benefits and costs of controlling for model bias were discussed in terms of rates of convergence and variation of the estimates, where we saw that controlling for the model bias causes a slower rate of convergence and larger variance of the estimates; hence, larger confidence intervals. It was noted that for a fixed sample size, every variable added to the nonparametric function causes larger variation in the estimate. We developed an approach to this problem by justifying theoretically and empirically the parametric estimation of the part of the regression function associated with the baseline covariate. Thus, we were able to reduce the dimension of the function to estimate nonparametrically and still control the model bias, which resulted in the formulation of the partially linear model (1.2).

This research is not the first to examine the partially linear model. Thus, in Chapter II we explored the existing literature on kernel type nonparametric regression and the partially linear model. We reviewed kernel estimation literature and found that although all of the estimators have the same rate of convergence, some have better MSE properties than others. That is, the constants in the MSE expressions of some estimators have better interpretation and/or are smaller than others. It was clear from review in Chapter II that locally linear estimation had the most desirable MSE properties.

The issue for estimation of the partially linear model is how to maintain the parametric rate of convergence of the linear estimates and the usual MSE properties of the nonparametric estimate. An important result by Rice (1986) indicated that when the covariates associated with the linear part of the model have a dependence relationship with the covariates associated with the nonparametric function, the parametric rate of convergence may not be obtained by existing methods of estimation. This negative result was addressed by Denby (1986) and Speckman (1988) who independently developed a two stage estimation procedure that achieved the desired outcome for kernel type estimation: root n convergence for the estimates of the linear coefficients and the usual MSE properties for the nonparametric estimate. These results motivated us to propose using Speckman's (1988) two stage estimation approach using locally linear smoothing. This approach had not been examined in the literature. Specifically, what had been done before required the assumption of a fixed design for the nonparametric function, where results only addressed MSE properties of estimating a univariate function, and no examination of locally linear estimation. Therefore, in order to proceed in this direction we had to examine the large sample and finite sample properties of the estimates.

In Chapter III, we detailed exactly how we would estimate model (1.2). This involved showing how Speckman's (1988) two stage estimation method would combine with locally linear estimation as described by Fan (1993). After describing the estimation procedure, the next logical step was to justify the approach with asymptotic results. Thus, in Chapter IV, we summarize our analytic findings in the form of four theorems. We found that using our proposed estimation method preserved the desirable results of Speckman's two stage approach as well as achieved the desirable MSE properties associated with locally linear estimation. Immediately following the asymptotic results in Chapter IV are supporting proofs in Chapter V.

It is well known within the field of nonparametric estimation that asymptotic results may not be relevant in finite sample estimation. This is highly dependent upon the context. For instance, in the field of Physics, particle experiments may yield millions of observations, in which case the asymptotic results may be a good approximation of the properties of the estimates. However, in some biological experiments, each observation may require the sacrifice of a primate, which results in very small sample sizes where asymptotic results would provide little guidance. In such cases, it is instructive to examine the finite sample behavior of a proposed method. Since the sample size of our motivating example was 350 observations, we thought it necessary to examine the finite sample behavior of the proposed estimates of model (1.2). Our approach in Chapter VI was to design and implement simulation experiment such that we could maximize the amount of information gained about the estimators within the constraints of our computing resources and research schedule. The results indicated that the finite sample performance of the estimators was better than the asymptotic approximation would indicate for each sample size examined.

With the asymptotic and finite sample behavior of our proposed method established, we returned to the LRC data in Chapter VII. We addressed some important issues involved in estimating the partially linear model, such as smoothing parameter selection and confidence bars. We presented the results and discussed the information we gained about the relationship between cholesterol and compliance in light of the clinical questions posed in Chapter I. We also used model (1.2) to examine the estimates at each visit separately to see if we could gain information about trends over time.

8.2 Future Research:

From the discussion in Section 2.4, it was apparent that currently there does not exist a method for representing the variability of the regression function that is satisfactory to all. The analytical approach for calculating simultaneous confidence bands has received criticism that the bands are too wide. The bootstrap approach has open questions such as how to choose the number of points and interval spacing. For a satisfactory analytical approach, it seems that there needs to be some way of representing the variability of the curve within the infinite dimensional space of smooth curves. The current analytical approach detailed in Section 2.4, expresses the variability of the curve over infinite dimension space; however, fails to exclude the large space of non-smooth curves. Thus, the method produces confidence bands that are too wide. Consequently, this line of research remains an open question and we invite input and ideas from all sources.

The LRC data used as an illustrative example in this dissertation is longitudinal in nature, i.e. the same variables were measured consistently over several years. We considered the relationship between cholesterol at the third visit with baseline cholesterol and compliance at visits one and two. However, in Chapter VI we also used model (1.2) to examine the trend of cholesterol as a function of the covariates over several visits. Model (1.2) was not constructed to fully analyze longitudinal data. Therefore, in order to consider (1.2) as a possible model for longitudinal analysis, refinements would need to be made. One suggestion would be to calculate the covariance of the curves across all visits, which would take into account the multivariate aspect of the data. A more promising direction taken by Moyeed and Diggle (1994) is to consider (1.2) in the context of data that consists of a collection of short time series where the measurements within series are correlated. A useful modeling framework for this is that the observed sequence of measurements on an experimental unit is sampled from a realization of a continuous-time stochastic process.

An important issue not covered in this dissertation is the choice of smoothing parameters. In the past few years, there have been important developments in data based smoothing parameter selection [Jones, Marron and Sheather (1992)]. Recently, Ruppert, Sheather and Wand (1994) have examined the analytical and finite sample properties of some plug in bandwidth selectors that have performed well in the arena of density estimation. They considered univariate locally linear regression. In Speckman's (1988) paper, he proposed a cross-validated estimate of the smoothing parameter in the partially linear model. However, it has been shown that the cross-validated estimate suffers from large variability [Jones, Marron and Sheather (1992)]. We suspect that the methodology developed in Ruppert, Sheather and Wand (1994) holds promise for estimation of the smoothing parameter for model (1.2). To verify this, we need to extend the results to the estimation of a bivariate function and the estimation of a partially linear model. This would provide a vital tool for practical data analysis.

IX. REFERENCES

Chambers, J. M. and Hastie, T. (eds) (1991) *Statistical Models in S,* Wadsworth, Pacific Grove, California.

Chen, H. and Shiau, J. -J. H. (1991). A Two-Stage Spline Smoothing Method for Partially Linear Models. *Journal of Statistical Planning and Inference.* 27: 187-202.

Chung, K. L. (1974) *A Course in Probability Theory.* Academic Press, N. Y.

Denby, L. (1986). Smooth Regression Function. *Statistical Resiarch Report #26.* AT&T Bell Laboratories, Princeton, New Jersey.

Efron, B. and Feldman, D. (1991) "Compliance as and Explanatory Variable in Clinical Trials," (with discussion), *Journal of the American Statistical Association*, 86: 9-26.

Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Nonparametric Estimates of the Relation Between Weather and Electricity Sales. *Journal of the American Statistical Association.* 81: 310-320.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Dekker.

Eubank, R. Hart, J. D. and Speckman, P. (1990). Trigonometric Series Regression Estimators with an Application to Partially Linear Models, *Journal of Multivariate Analysis,* 32:70-83.

Eubank, R. and Speckman, P. (1993). Confidence Bands in Nonparametric Regression, *Journal of the American Statistical Association,* 88:1287-1301.

Falk, M. (1993) Asymptotically Optimal Estimators of General Regression Functionals, *Journal of Multivariate Analysis,* 47:59-81.

Fan, J. (1992). Design Adaptive Nonparametric Function Estimation: A Unified Approach, *Journal of the American Statistical Association,* 87:998.

Fan, J. and Gijbels, I. (1993). Variable Bandwidth and Local Linear Regression Smoothers, *Annals of Statistics,* 20:2008-2036.

Fan, J. (1994). Local Linear Regression Smoothers and their Minimax Efficiency, *Annals of Statistics* 21:196-216.

Fan, J. and Marron, J. S. (1993) Fast Implementations of Nonparametric Curve Estimators. *University of North Carolina Department of Statistics Mimeo Series #2093.*

Gasser, T. and Müller, H. G. (1979). Kernel Estimation of Regression Functions. (In) *Smoothing Techniques for Curve Estimation.* Eds. Gasser and Rosenblatt. Heidelberg: Springer-Verlag.

Gasser, T., Müller, H. G., Köhler, W., Molinari, L. and Prader, A. (1984). Nonparametric Regression Analysis of Growth Curves, *Annals of Statistics,* 12:210-229.

Gasser, T. and Kneip, A. (1989) Discussion of "Linear Smoothers and Additive Models", by Hastie and Tibshirani, *Annals of Statistics,* 17:453-555.

Grady, J. J. (1992). Structural Covariance Matrices for Incomplete Longitudinal Data. *Institute of Statistics Mimeo Series #2102T, University of North Carolina Department of Biostatistics.*

Green, P., Jennison, C. and Seheult, A. (1985). Analysis of Field Experiments by Least Squares Smoothing. *Journal of the Royal Statistical Society, Series B.* 47: 299-315.

Härdle, W. (1989). *Applied Nonparametric Regression.* New York: Cambridge University Press.

Härdle, W. and Marron, J. S. (1985). Optimal Bandwidth Selection in Nonparametric Regression Function Estimation, Annals of Statistics, 13:1465-1481.

Hart, D. and Wehrly, T. E. (1986). Kernel Regression Estimation using Repeated Measurements Data. *Journal of the American Statistical Association.* 81: 1080-1088.

Hastie, T. J. and Tibshirani, R. J. (1991) *Generalized Additive Models,* Chapman and Hall, New York.

Heckman, N. E. (1986). Spline Smoothing in a Partly Linear Model. *Journal of the Royal Statistical Society, Series B.* 48: 244-248.

Lipids Research Clinic Program (1984). "The Lipid Research Clinics Coronary Primary Prevention Trial Results, Parts I and II," *Journal of the American Medical Association*, 251: 351-374.

Moyeed, R. A. and Diggle, P. J. (1994). Rates of Convergence in Semi-parametric Modelling of Longitudinal Data. *Unpublished Manuscript.*

Nadaraya, E. A. (1964). On Estimating Regression. *Theory Prob. Appl.* 10: 186-190.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode, *Annals of Mathematical Statistics.* 33: 1065-1076.

Rice, J. (1986). Convergence Rates for Partially Splined Models. *Probability Letters.* 4: 203-208.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics.* 27: 642-669.

Ruppert, D. and Wand, M.P. (1992). Multivariate Locally Weighted Least Squares

Regression. *Technical Report 92-4, January 1992, Cornell University.*

Ruppert, D., Sheather, S. J. and Wand, M. P. (1994). An Effective Bandwidth Selector for Local Least Squares Regression, *Unpublished Manuscript.*

Shiau, J., Wahba, G. and Johnson, D. R. (1986). Partial Spline Models for the Inclusion of Tropopause and Frontal Boundary Information in Otherwise Smooth Two and Three Dimensional Objective Analysis. *Journal of Atmospheric and Oceanic Technology.* 3: 714-725.

Sen, P. K. and Singer, S. M. (1993) *Large Sample Methods in Statistics: An Introduction with Applications.* Chapman and Hall.

Silverman, B. W. (1985). Spline Smoothing: The Equivalant Variable Kernel Method, *Annals of Statistics,* 12:898-916.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Speckman, P. (1988). Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society, Series B.* 50: 413-436.

Stone, C. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression, *Annals of Statistics,* 10:1040-1053.

Wahba, G. (1984). Partial Spline Models for the Semiparametric Estimation of Functions of Several Variables, *In: Statistical Analysis of Time Series,* 312-329. Institute of Statistical Mathematics. Tokyo.

Wand, W. P. and Jones, M. C. "Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation," *Unpublished Manuscript.*

Watson, G. S. (1964). Smooth Regression Analysis, *Sankhyā, Series A.* 26: 359-372.

Wehrly, T. E. and Hart, J. D. (1988). Bandwidth Selection for Kernel Estimators of Growth Curves with Correlated Errors (abstract), *Institute of Mathematical Statistics Bulletin,* 17:236.

	Min		*Inflection*		*Max*	
n	h_1	h_2	h_1	h_2	h_1	h_2
100	1.181	1.722	-0.012	0.331	-0.699	-1.598
500	0.801	1.454	0.055	0.057	-0.486	-0.701
2500	0.487	0.778	0.010	-0.003	-0.328	-0.300

Table 1. Bias of the local linear estimate, uniform design.

	Min		*Inflection*		*Max*	
n	h_1	h_2	h_1	h_2	h_1	h_2
100	1.081	1.816	-0.002	0.317	-0.820	-1.654
500	0.679	1.461	-0.035	0.085	-0.528	-0.705
2500	0.362	0.767	-0.053	0.032	-0.398	-0.283

Table 2. Bias of the local linear estimate, beta design.

	Min		*Inflection*		*Max*	
n	h_1	h_2	h_1	h_2	h_1	h_2
100	1.725	1.220	1.607	0.688	1.342	0.673
500	0.739	0.552	0.736	0.437	0.718	0.442
2500	0.422	0.296	0.406	0.289	0.412	0.309

Table 3. Standard deviation of the local linear estimate, uniform design.

	Min		*Inflection*		*Max*	
n	h_1	h_2	h_1	h_2	h_1	h_2
100	1.388	1.151	1.339	0.662	1.475	0.729
500	0.683	0.532	0.702	0.440	0.674	0.450
2500	0.381	0.286	0.397	0.303	0.445	0.311

Table 4. Standard deviation of the local linear estimate, beta design.

	Min		*Inflection*		*Max*	
n	h_1	h_2	h_1	h_2	h_1	h_2
100	4.370	4.454	2.583	0.583	2.290	3.007
500	1.188	2.419	0.545	0.194	0.752	0.687
2500	0.415	0.693	0.165	0.084	0.277	0.185

Table 5. MSE of the local linear estimate, uniform design.

	Min		*Inflection*		*Max*	
n	h_1	h_2	h_1	h_2	h_1	h_2
100	3.095	4.623	1.793	0.539	2.848	3.267
500	0.928	2.418	0.494	0.201	0.733	0.699
2500	0.276	0.670	0.160	0.093	0.356	0.177

Table 6. MSE of the local linear estimate, beta design.

	Uniform		*Beta*	
n	h_1	h_2	h_1	h_2
100	1.026	-0.066	1.369	0.451
500	0.787	-0.060	0.543	0.119
2500	0.442	0.045	0.316	0.223

Table 7. Bias of the linear coefficient estimate x 100

	Uniform		*Beta*	
n	h_1	h_2	h_1	h_2
100	11.98	11.01	11.73	10.91
500	4.751	4.551	4.588	4.755
2500	1.966	1.978	1.949	2.042

Table 8. Standard deviation of the linear coefficient estimate x 100.

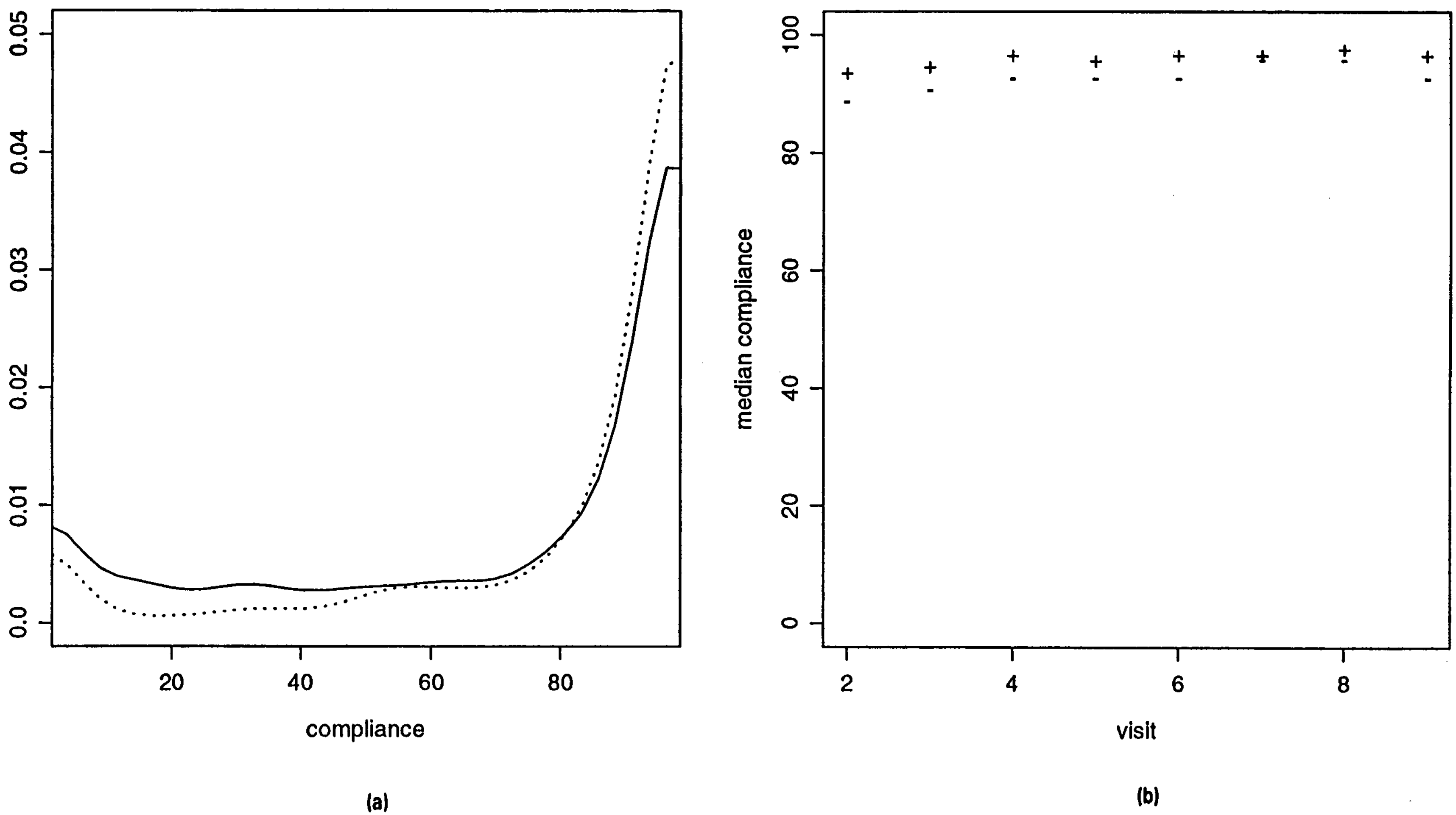
X_1, X_2	(0,0)	(10,10)	(20,20)	(30,30)	(40,40)	(50,50)	(60,60)	(70,70)	(80,80)	(90,90)	(100,100)
$\hat{m}(x)$	72.30	74.05	66.73	61.69	54.88	54.05	46.55	50.06	43.03	34.75	21.67

Table 9. Locally linear estimates at specific values of the compliance covariates.

X_1, X_2	(0,0)	(10,10)	(20,20)	(30,30)	(40,40)	(50,50)	(60,60)	(70,70)	(80,80)	(90,90)	(100,100)
$\hat{a}b + \hat{m}(x)$	285.3	287.1	279.7	274.7	267.9	267.1	259.6	263.1	256.0	247.8	234.7

Table 10. Estimates of total cholesterol at baseline of 300 and specific values of the compliance covariates.

Figure 1. Compliance to Treatment, LRC Data



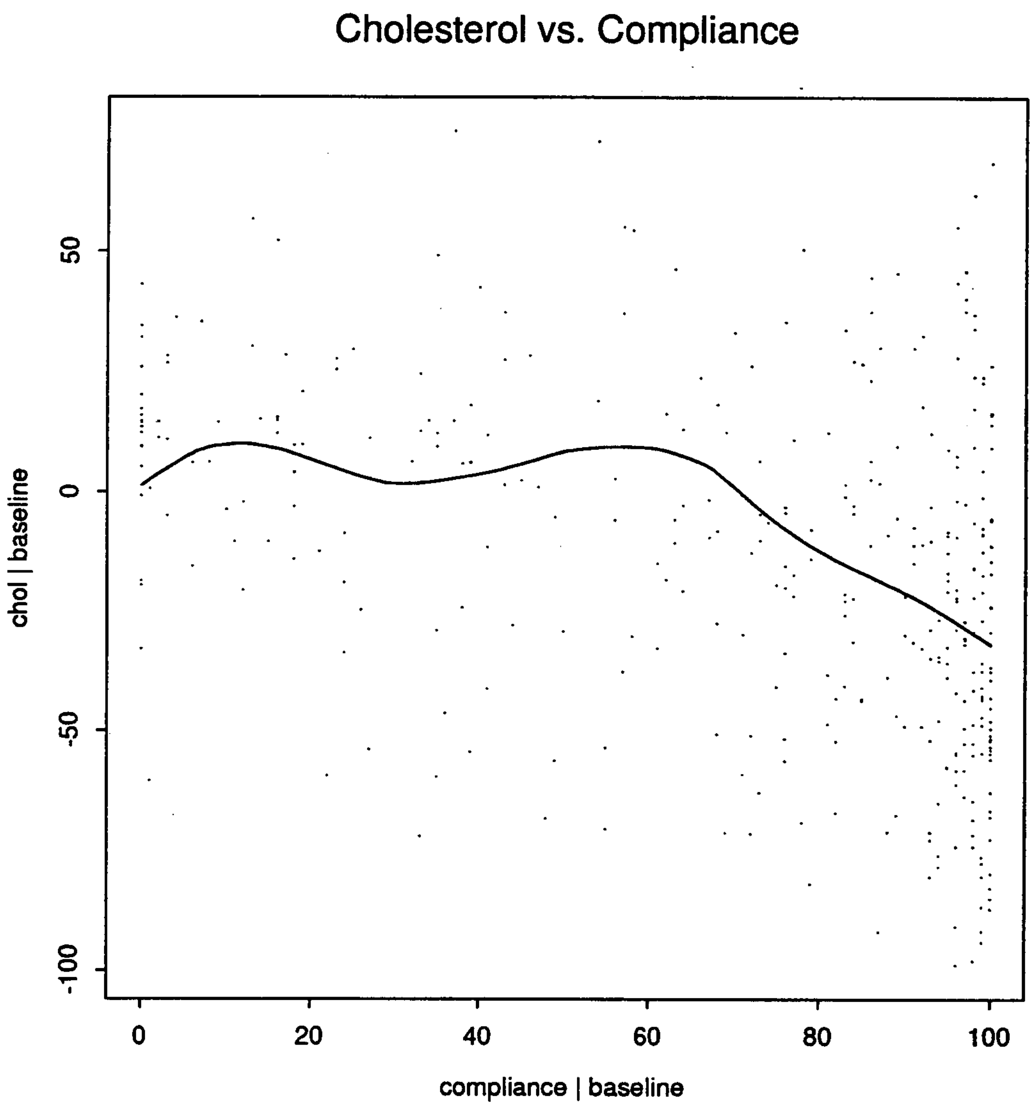

Cholesterol vs. Compliance
chol | baseline
compliance | baseline
50
0
-50
-100
0
20
40
60
80
100

Figure 2.

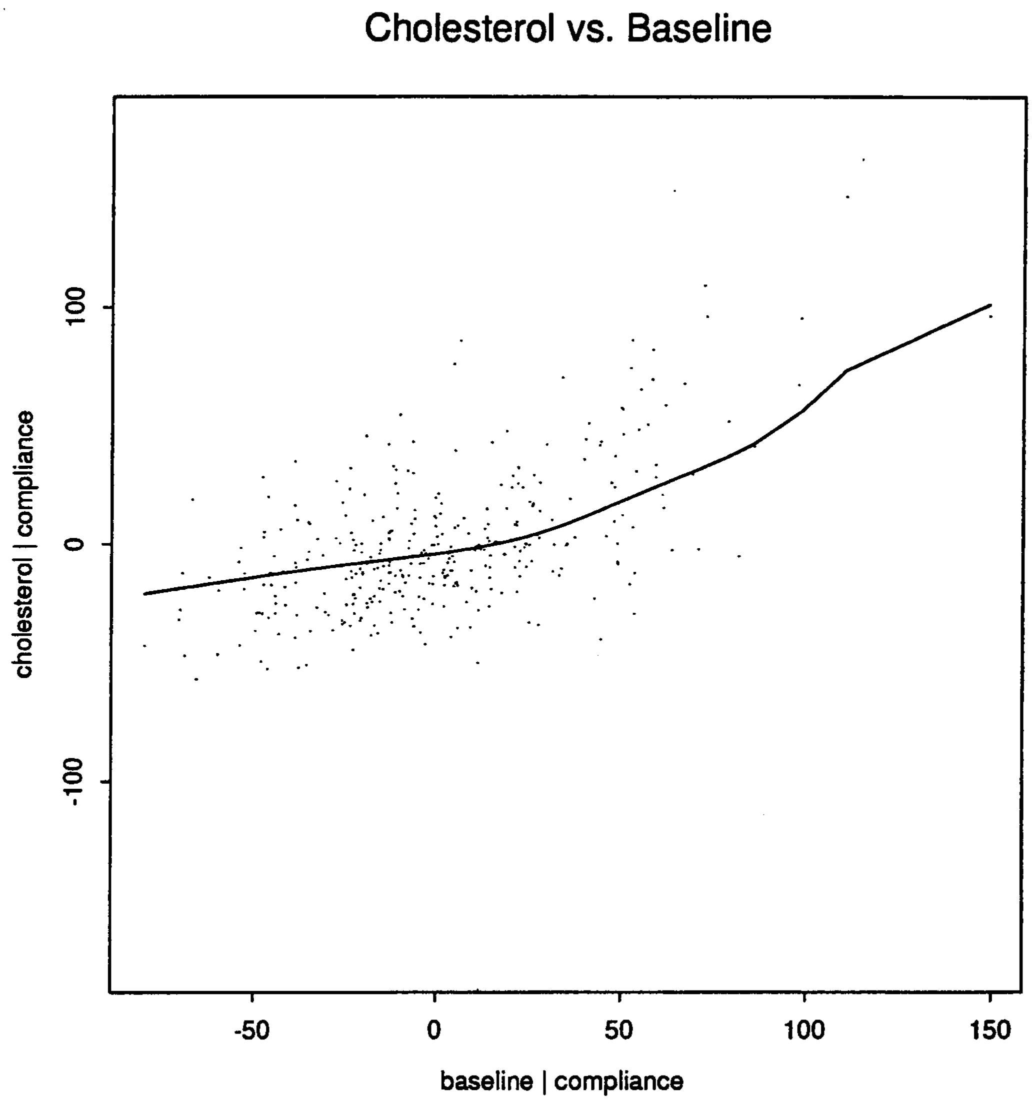

Cholesterol vs. Baseline
100
0
-100
cholesterol | compliance
-50
0
50
100
150
baseline | compliance

Figure 3.

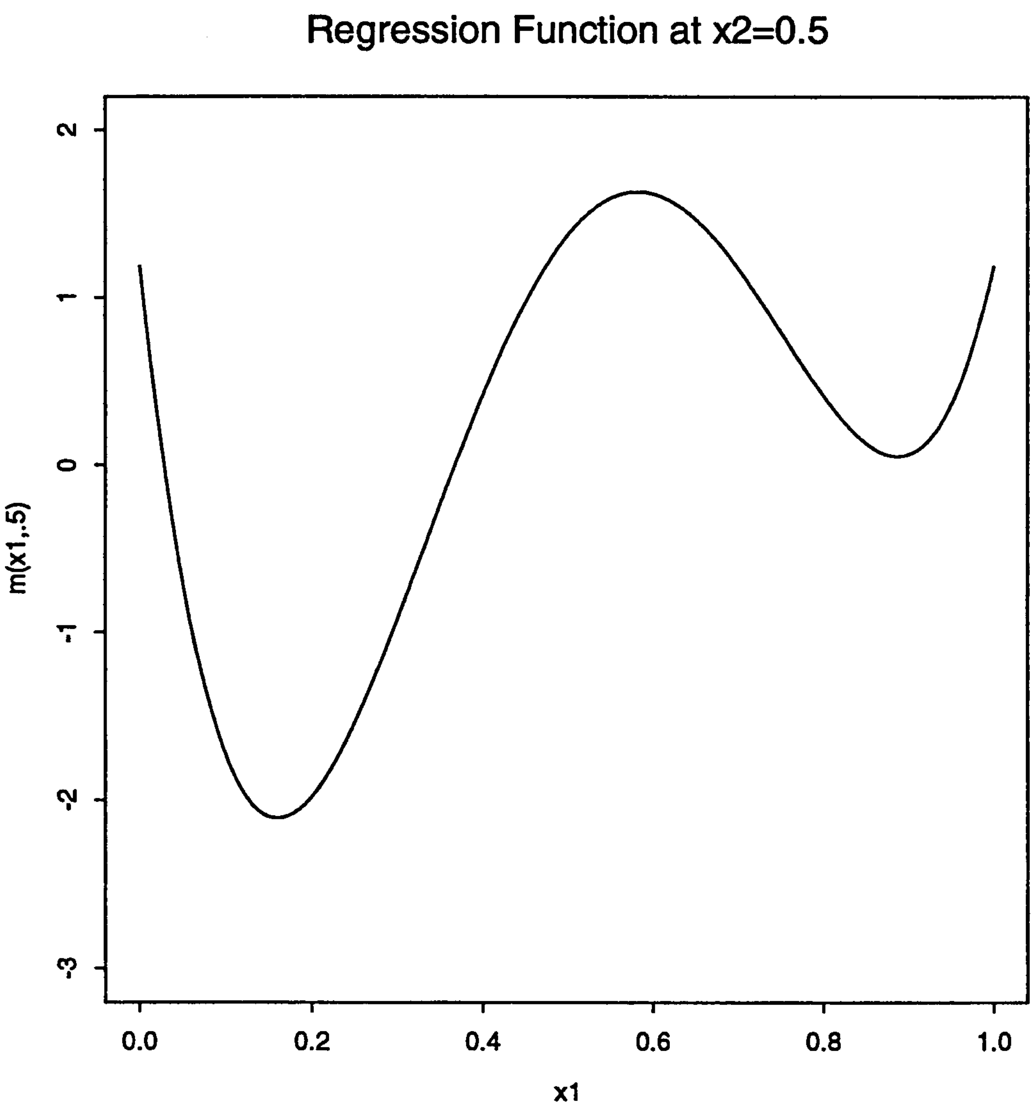


Figure 4.

Bandwidth Grid

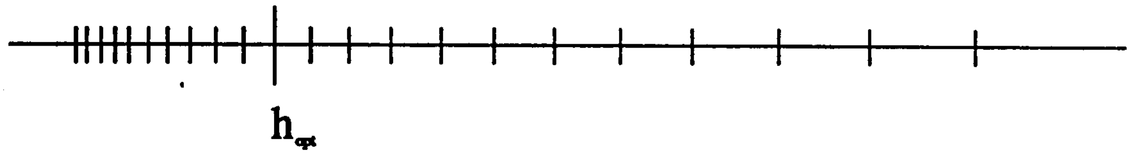


Figure 5.

Local Linear Estimate, Maxpoint, Asymptotic Optimal Bandwidth

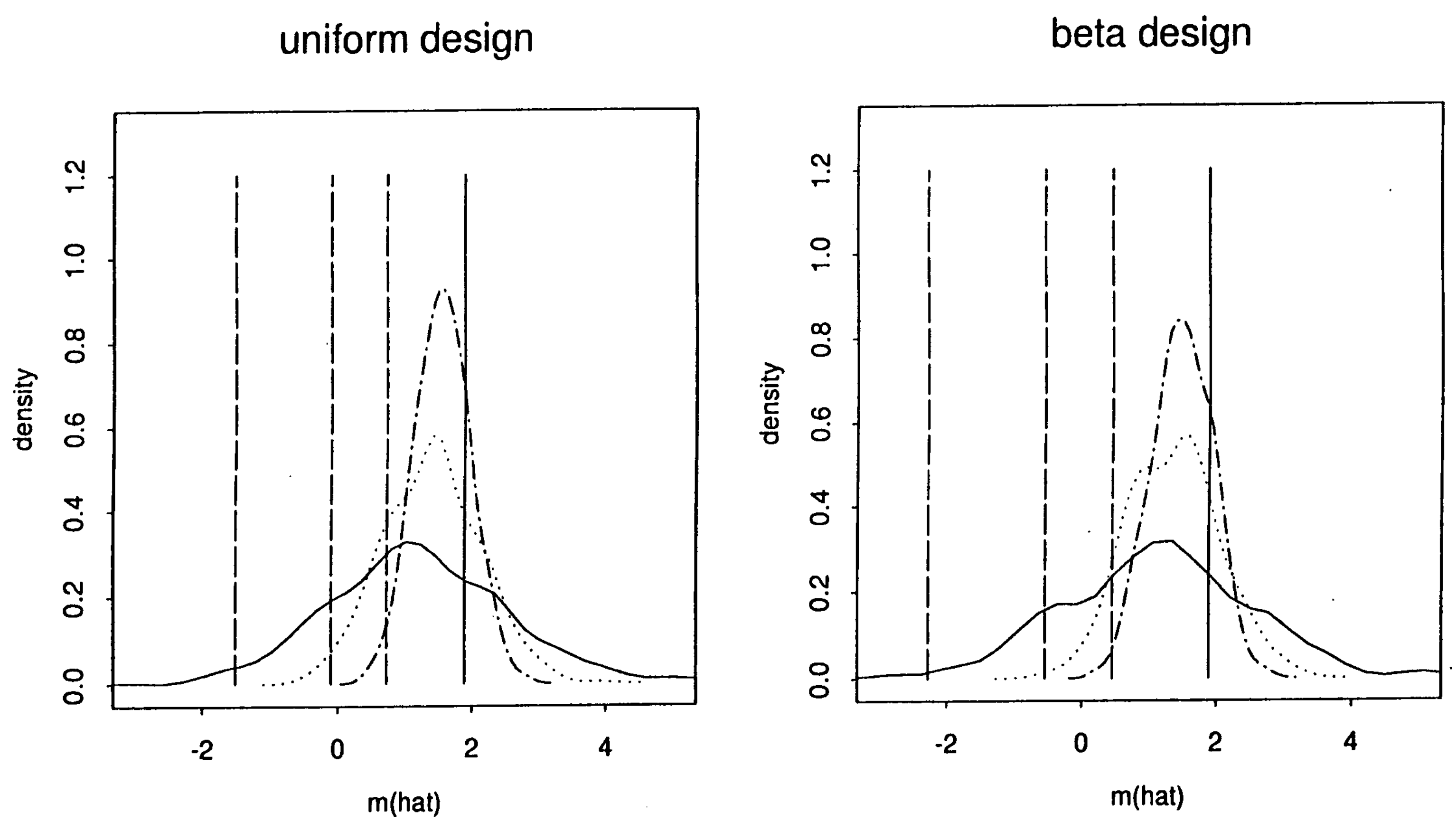


Figure 6.

n=100: solid, n=500: dotted, n=2500: dot-dashed

Local Linear Estimate, Inflection Point, Asymptotic Optimal Bandwidth

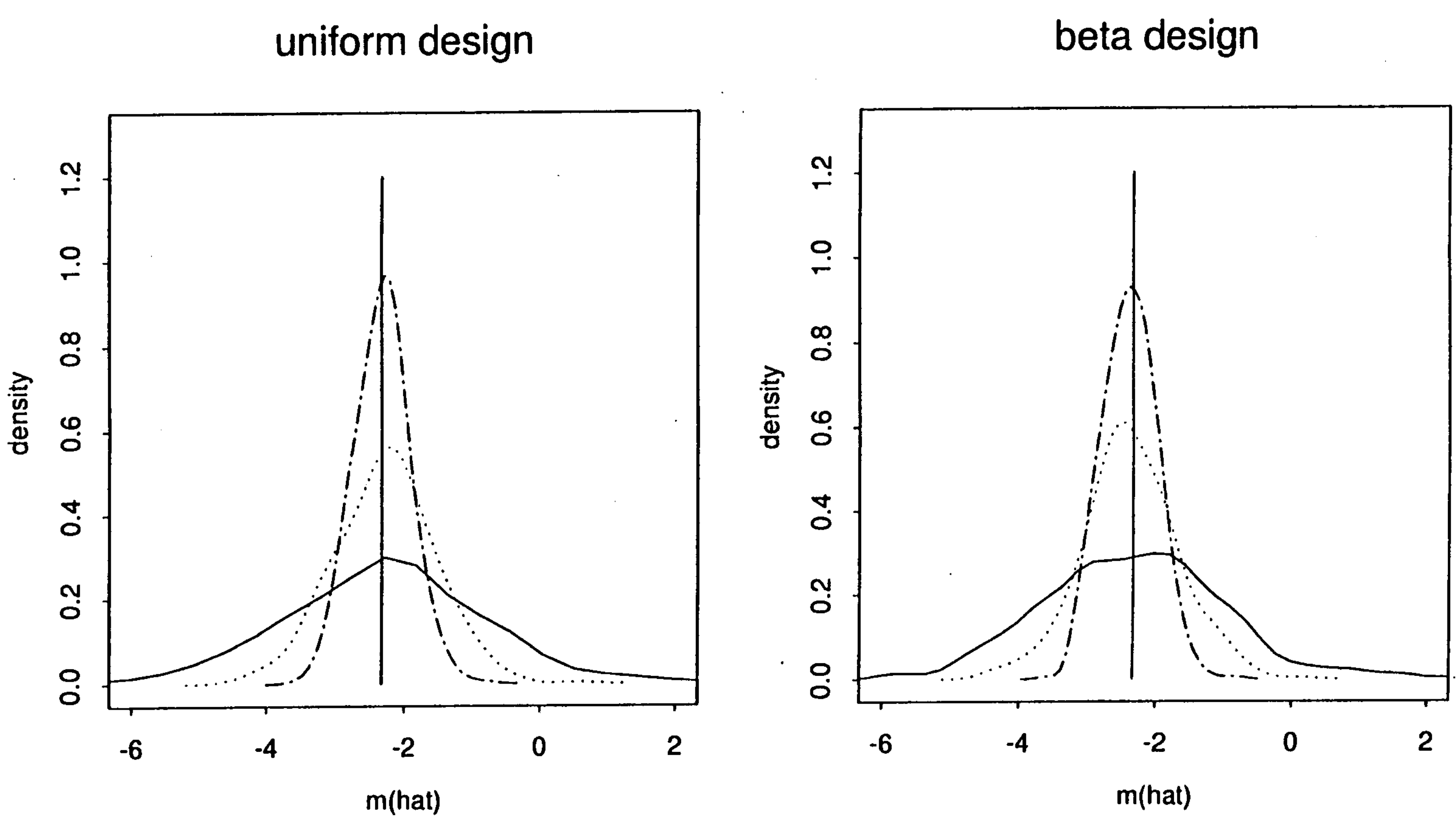


Figure 7.

n=100: solid, n=500: dotted, n=2500: dot-dashed

Local Linear Estimate, Minpoint, Asymptotic Optimal Bandwidth

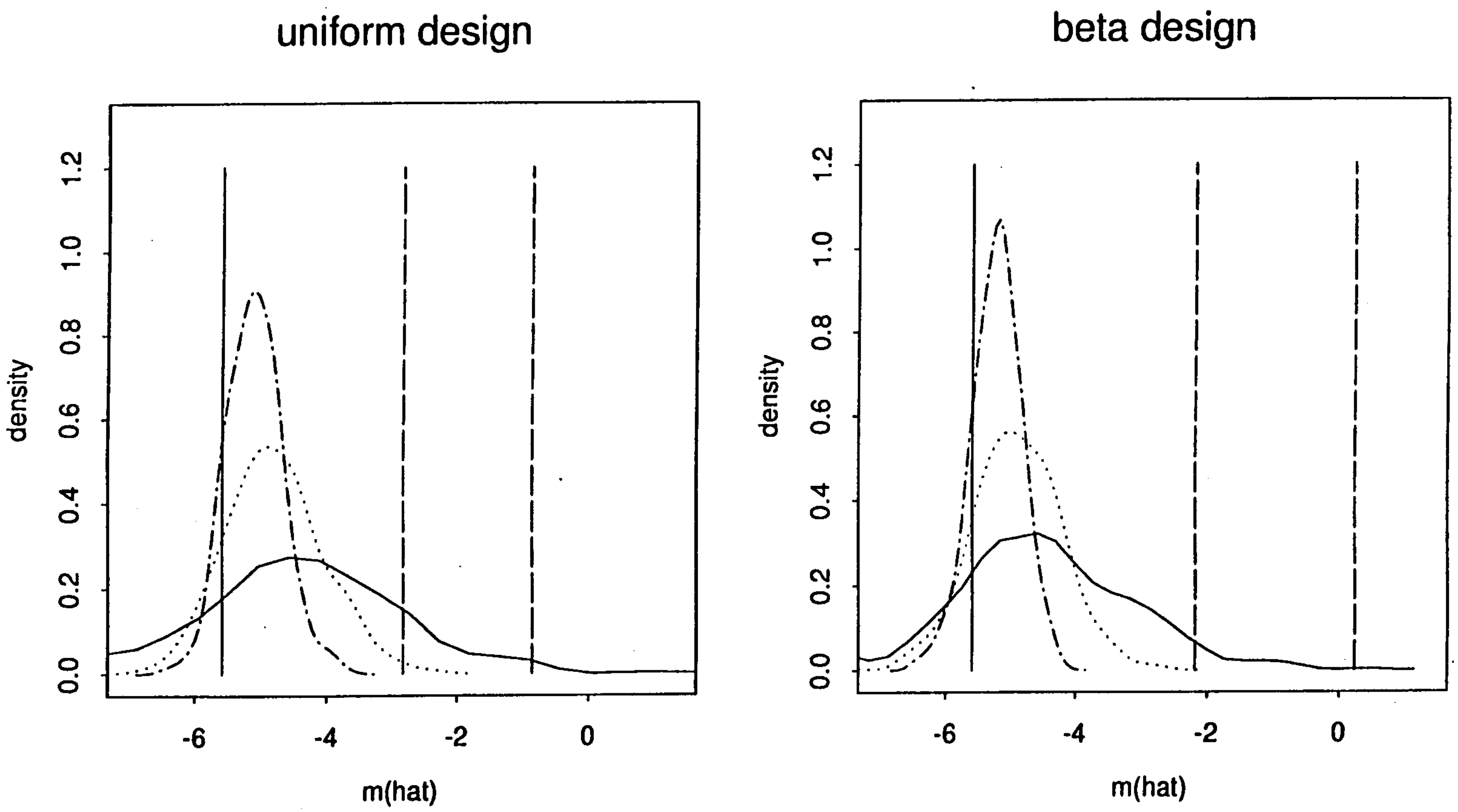


Figure 8.

n=100: solid, n=500: dotted, n=2500: dot-dashed

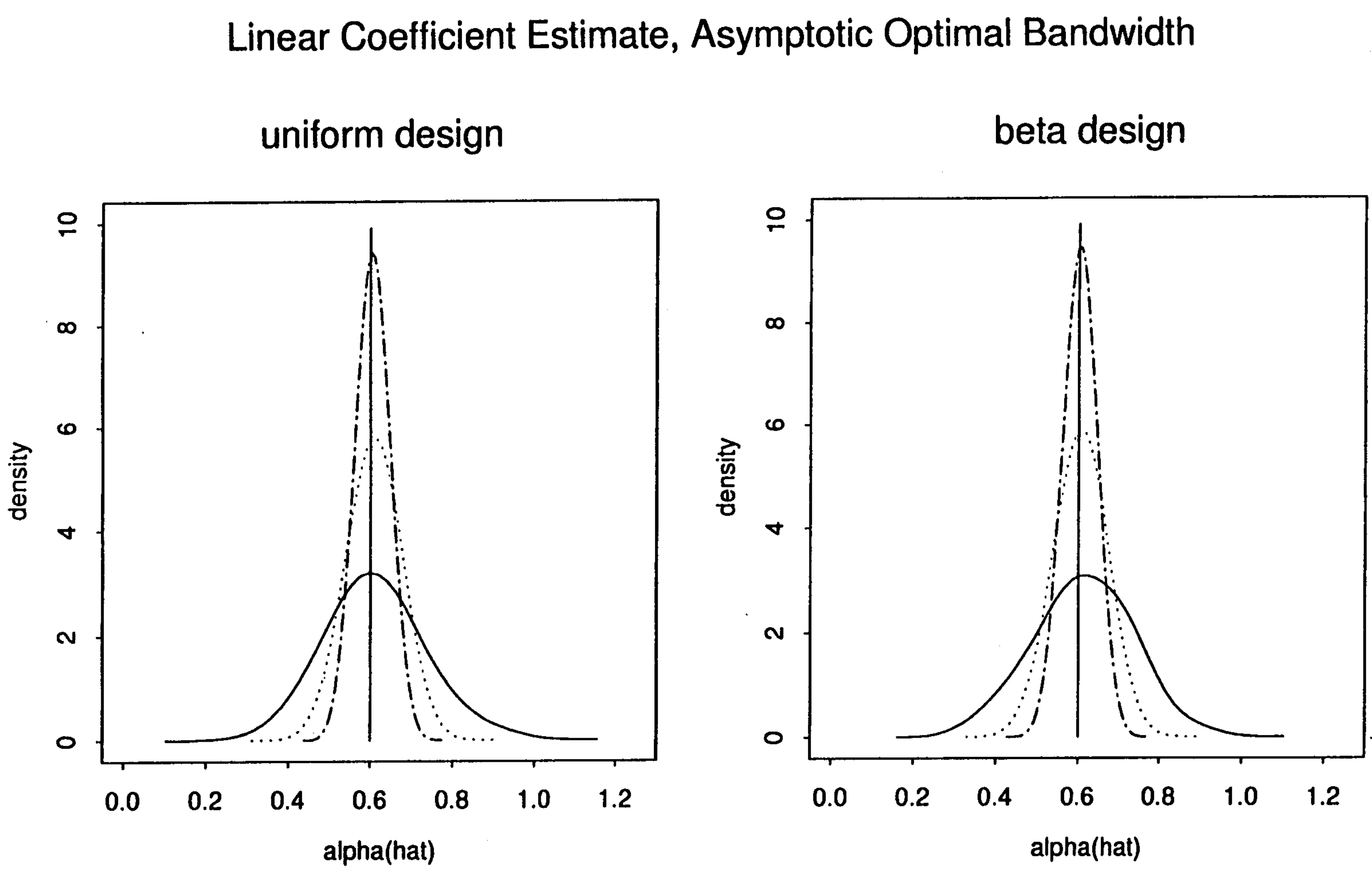


Figure 9.

n=100: solid, n=500: dotted, n=2500: dot-dashed

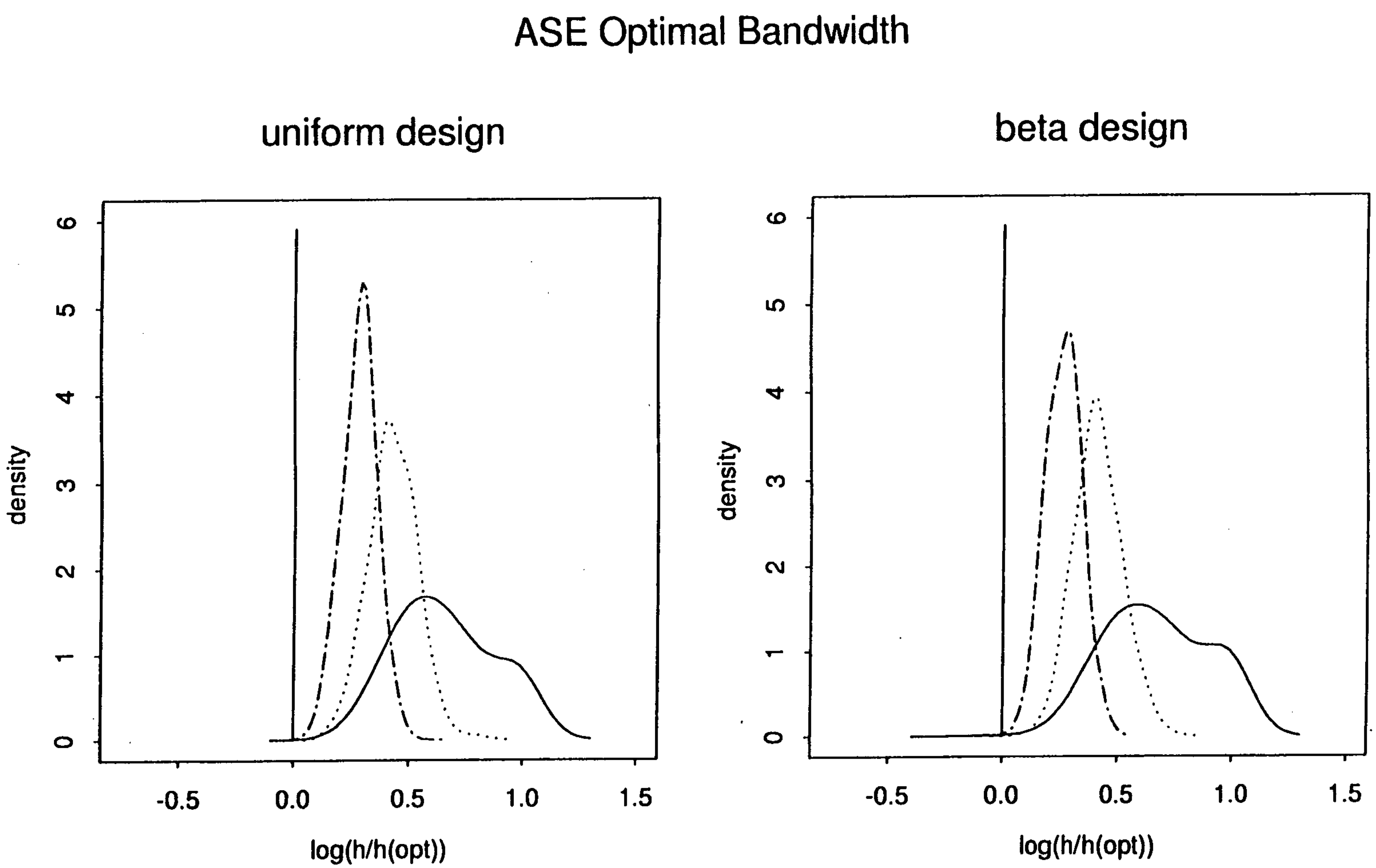


Figure 10.

n=100: solid, n=500: dotted, n=2500: dot-dashed

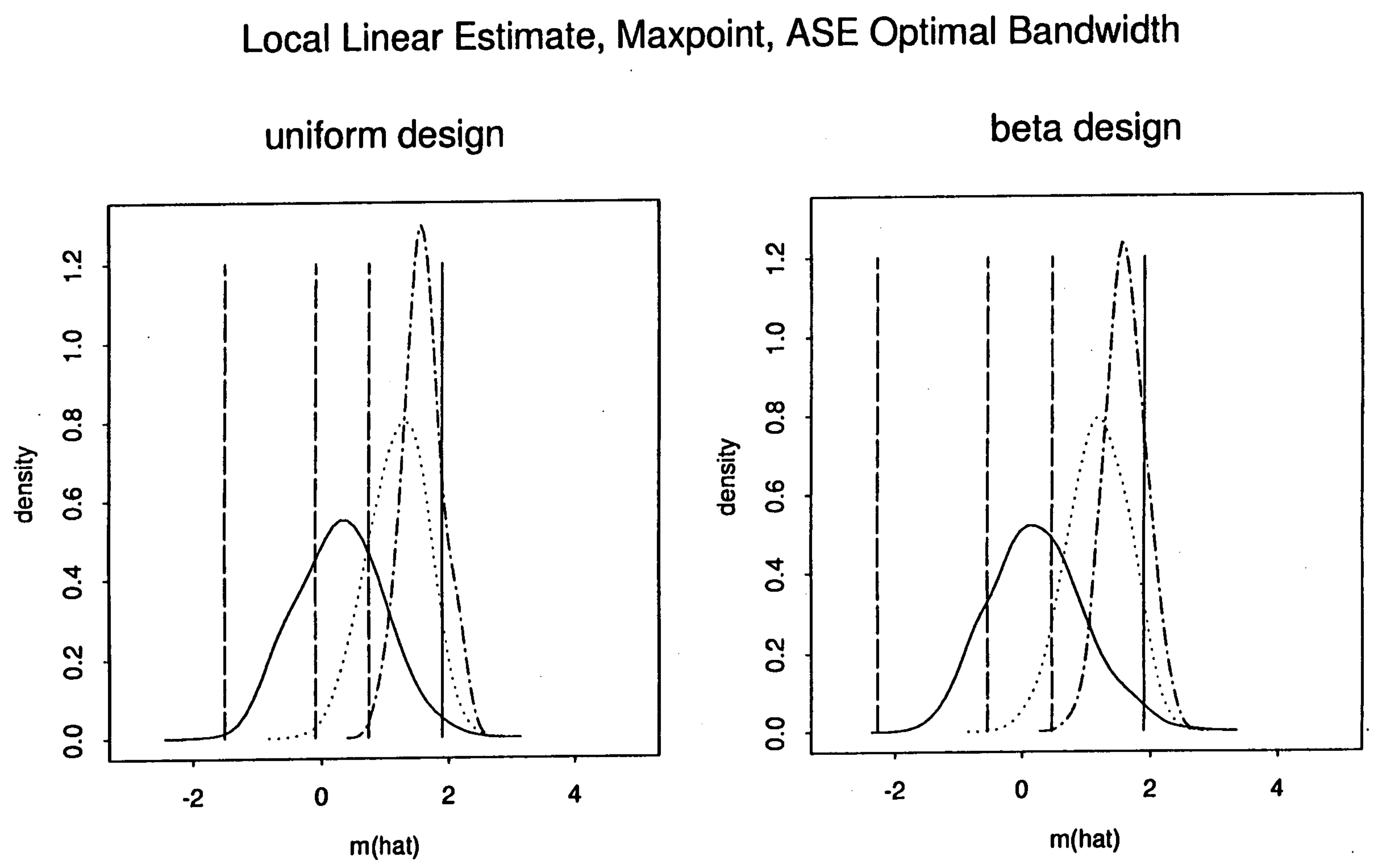


Figure 11.

n=100: solid, n=500: dotted, n=2500: dot-dashed

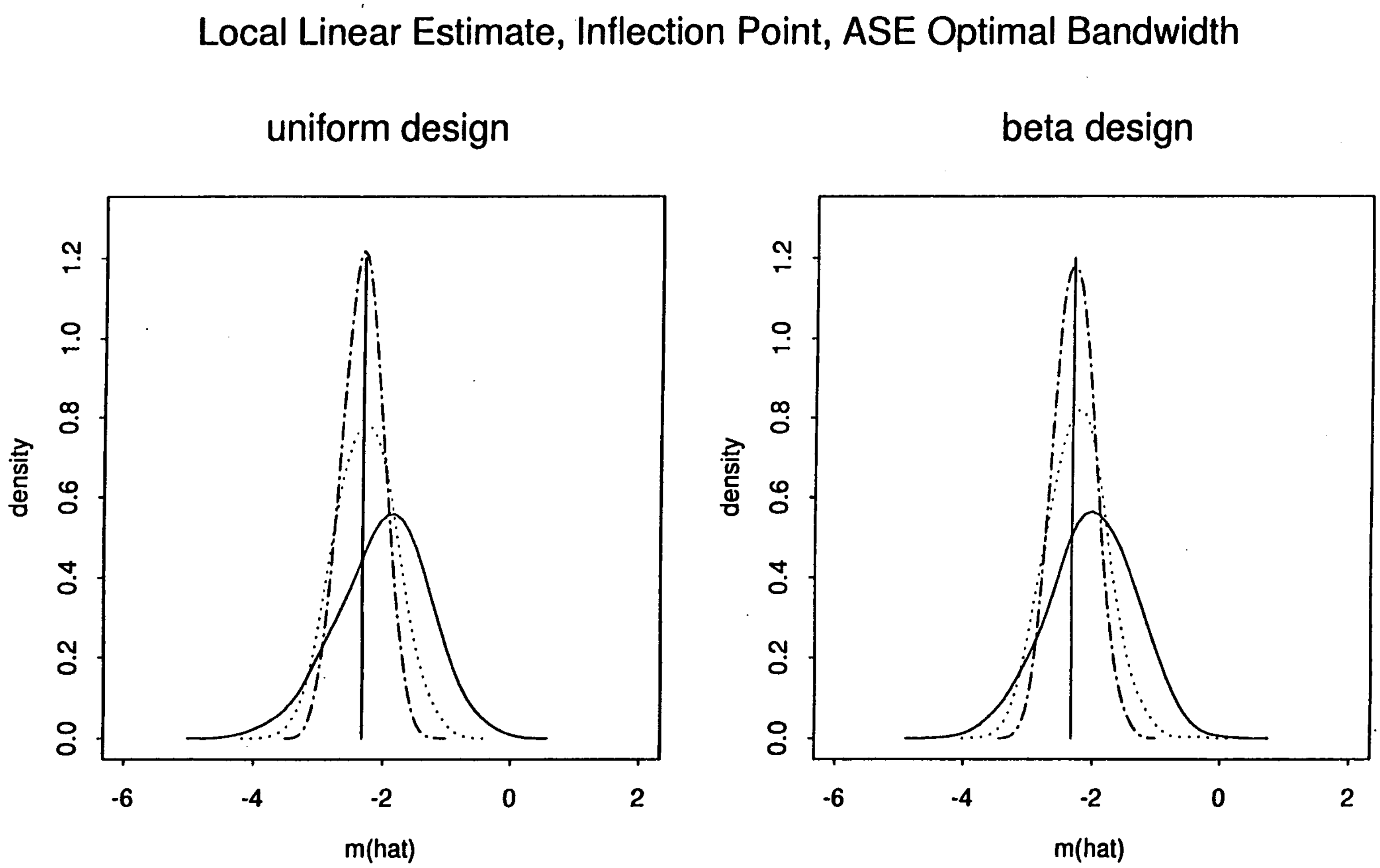


Figure 12.

n=100: solid, n=500: dotted, n=2500: dot-dashed

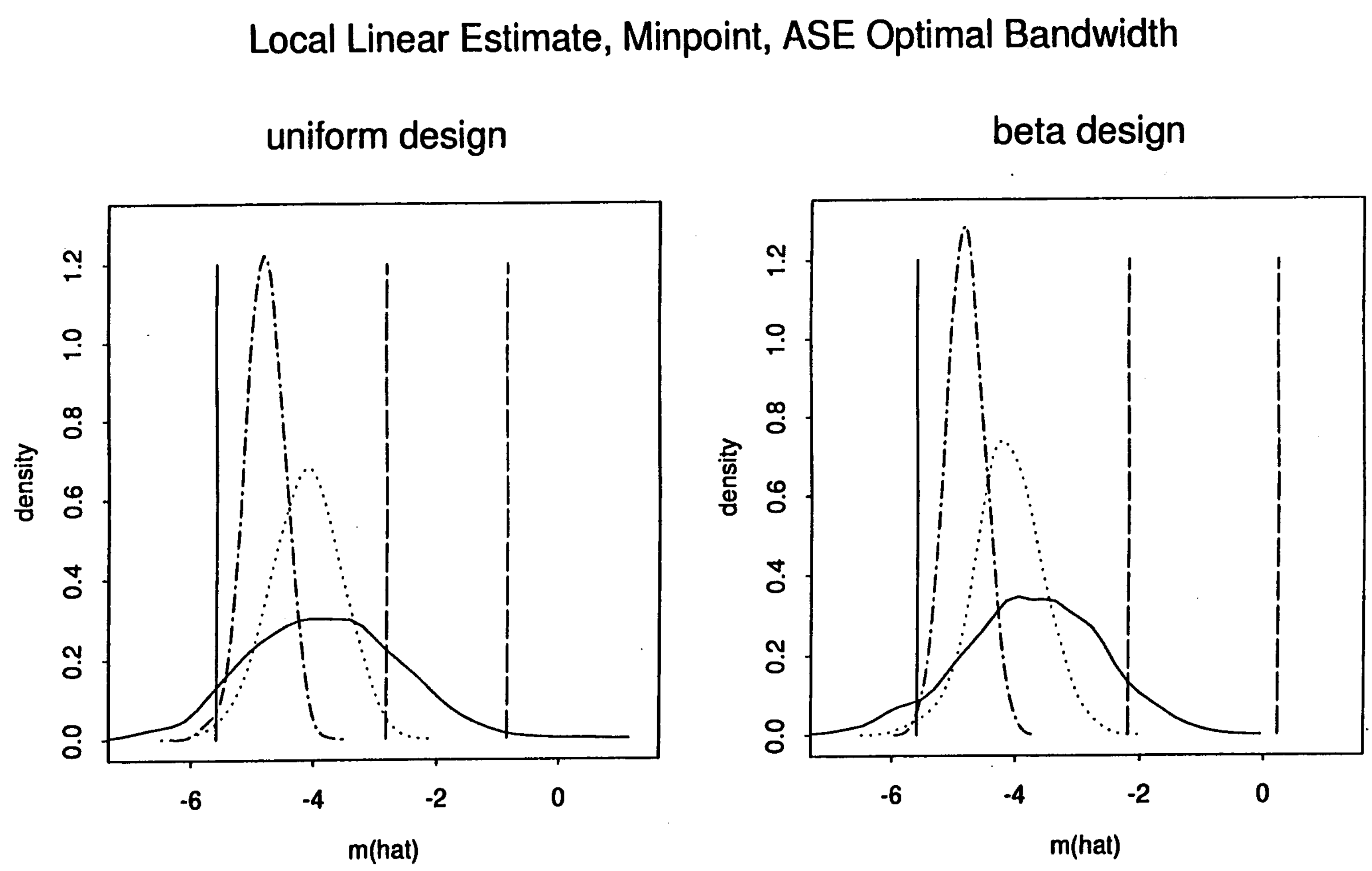


Figure 13.

n=100: solid, n=500: dotted, n=2500: dot-dashed

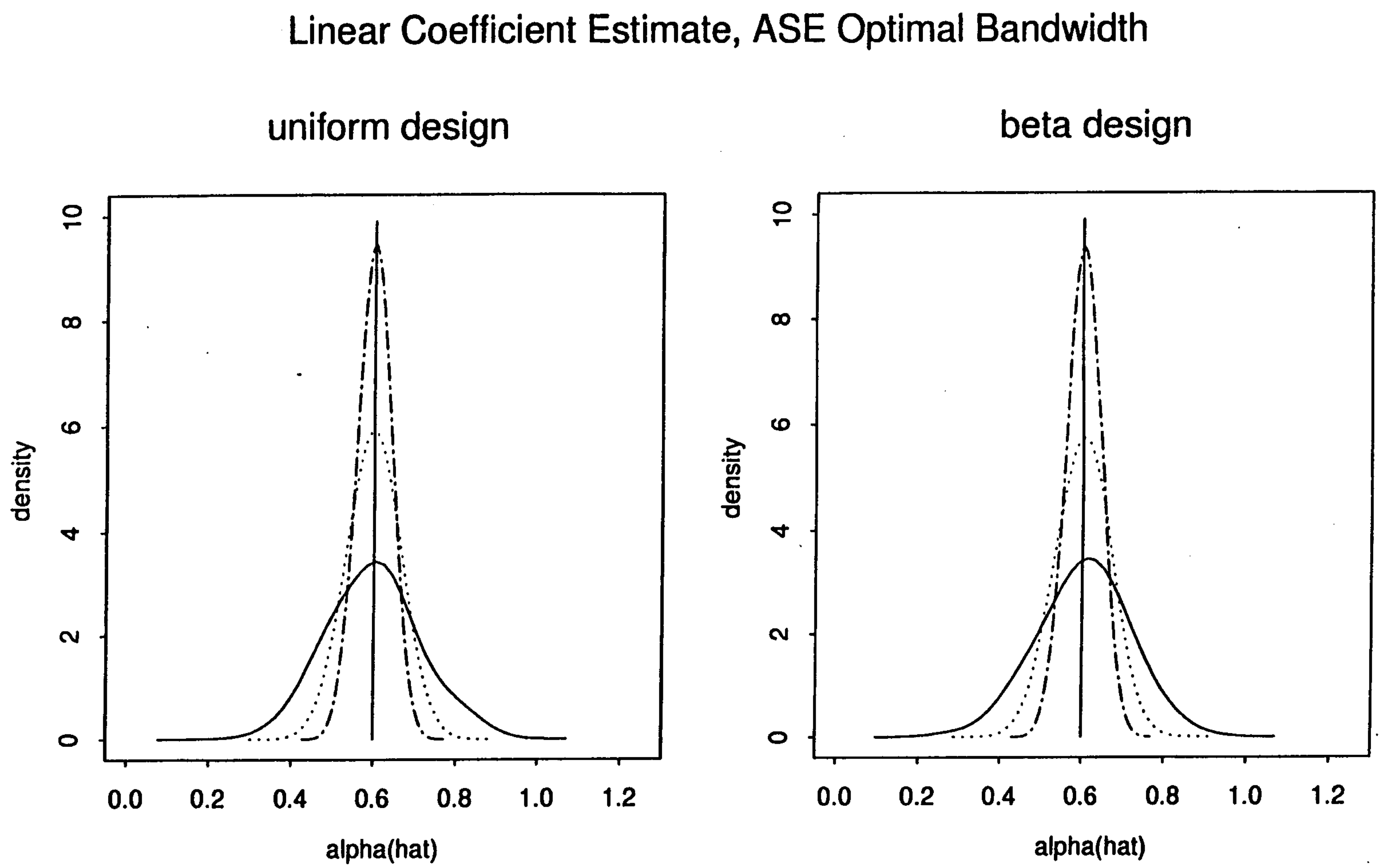


Figure 14.

n=100: solid, n=500: dotted, n=2500: dot-dashed

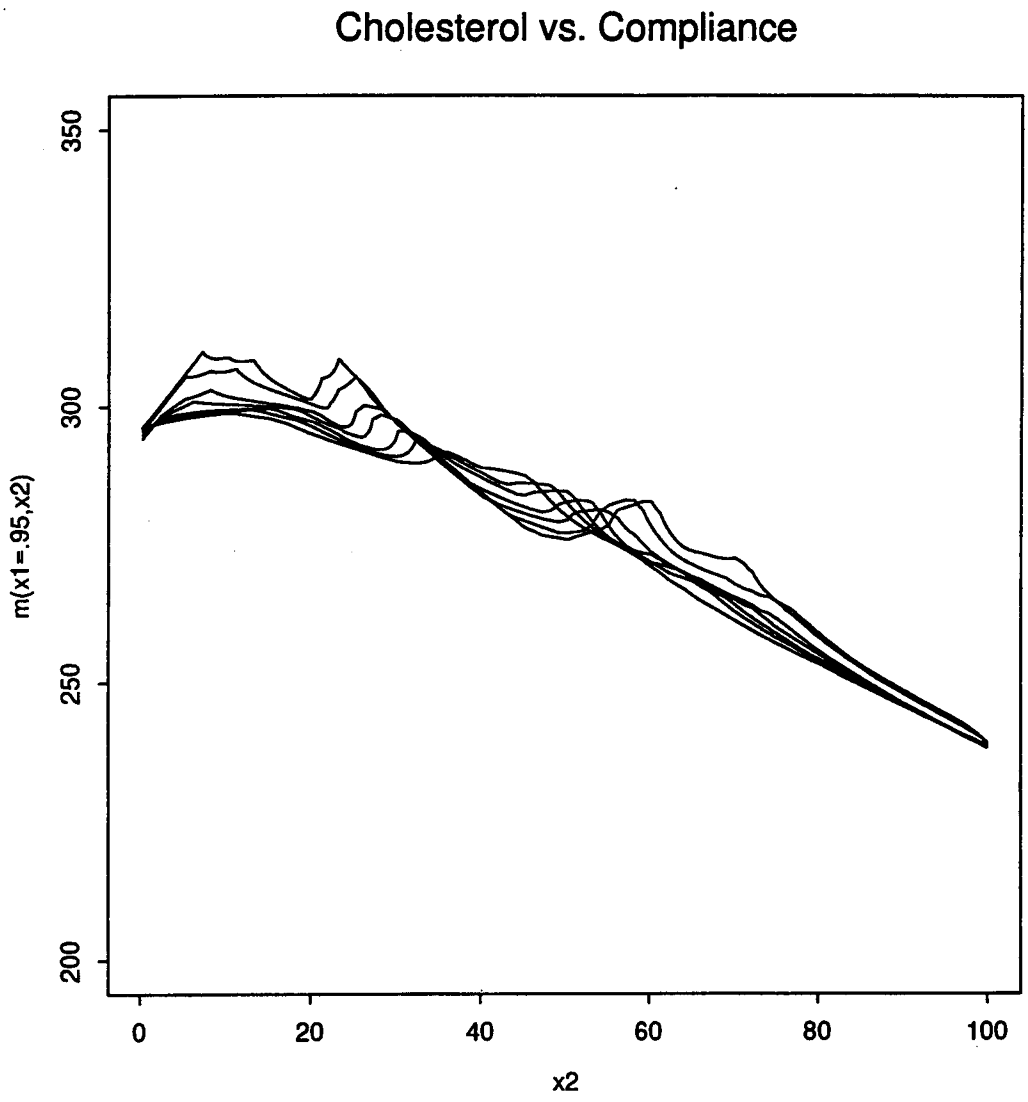

Cholesterol vs. Compliance
m(x1=.95,x2)
350
300
250
200
0
20
40
60
80
100
x2

Figure 15.

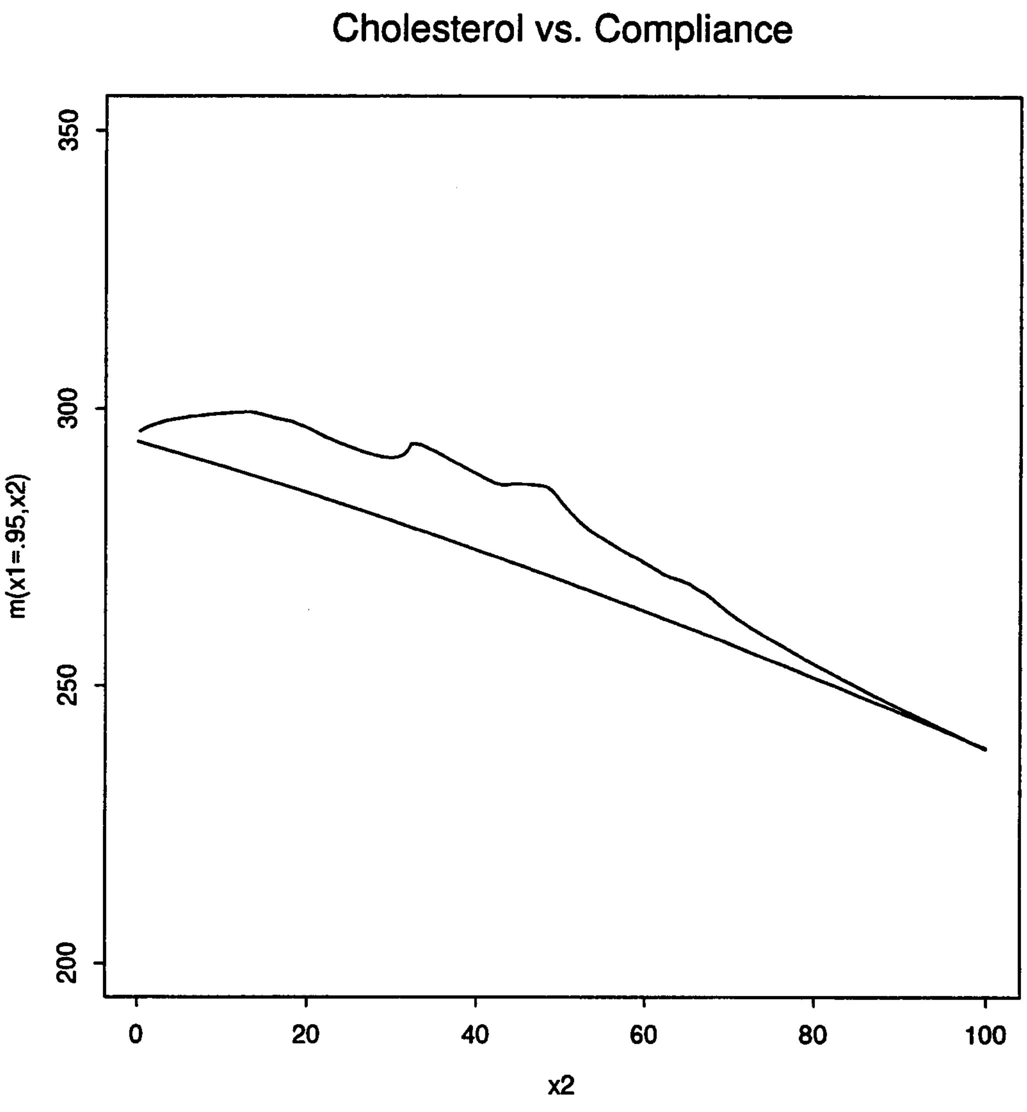

Cholesterol vs. Compliance
350
300
250
200
m(x1=.95,x2)
0
20
40
60
80
100
x2

Figure 16.

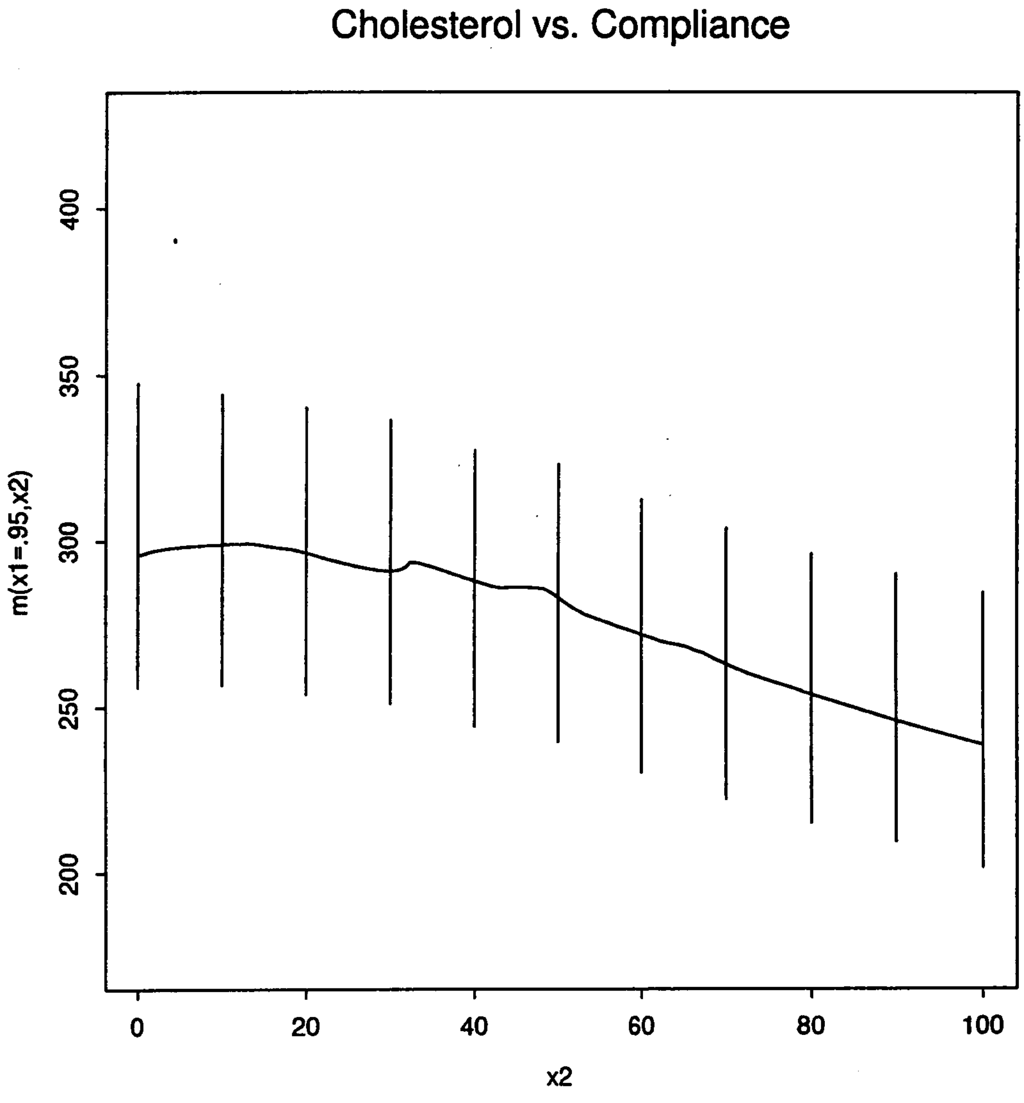

Cholesterol vs. Compliance
m(x1=.95,x2)
200
250
300
350
400
0
20
40
60
80
100
x2

Figure 17.

Figure 18. Contour Plots for Compliance at Present Visit vs. Mean of Past Visits

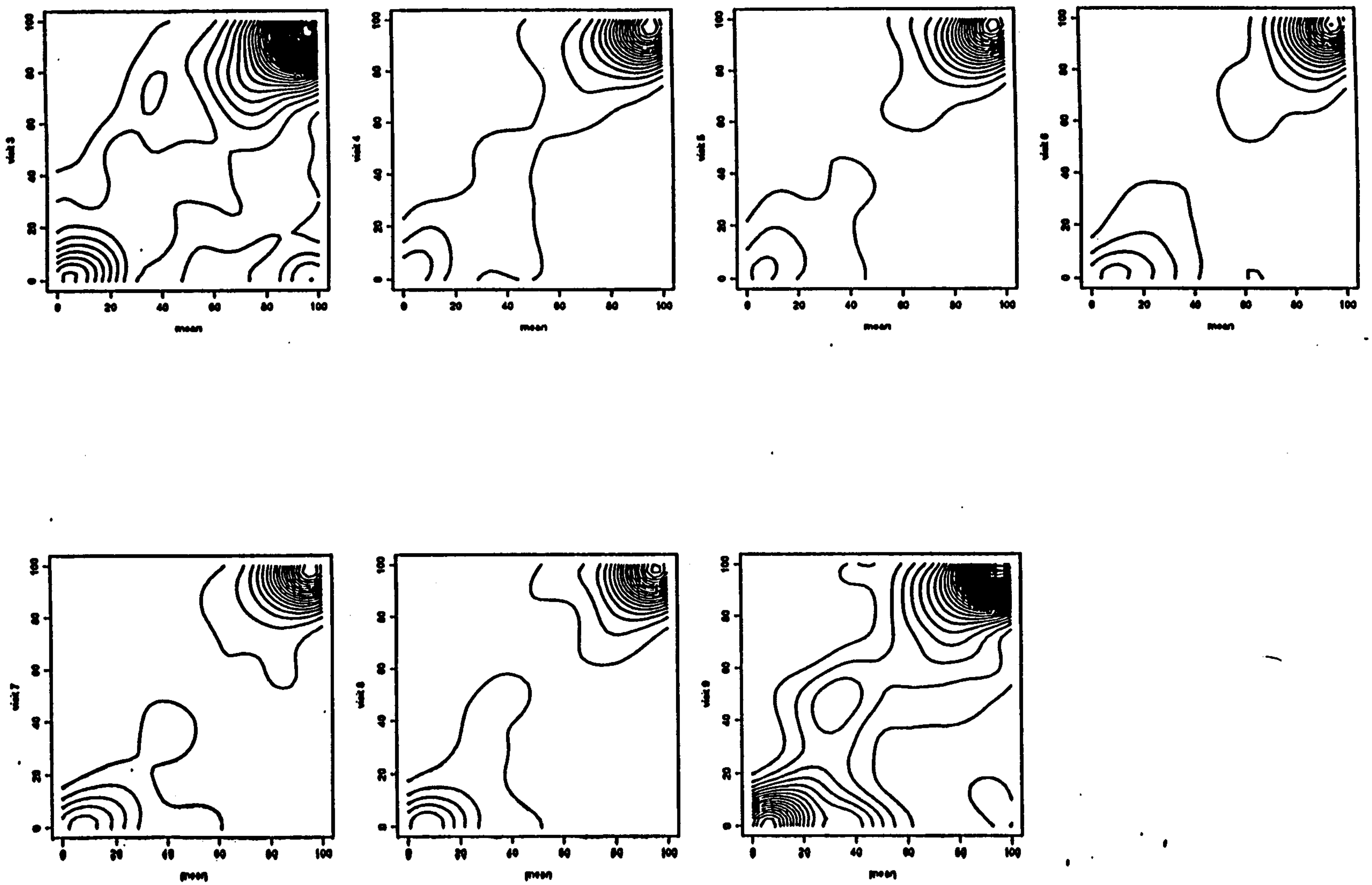


Figure 19. Perspectives of Compliance at Present Visit vs. Mean of Past Visits

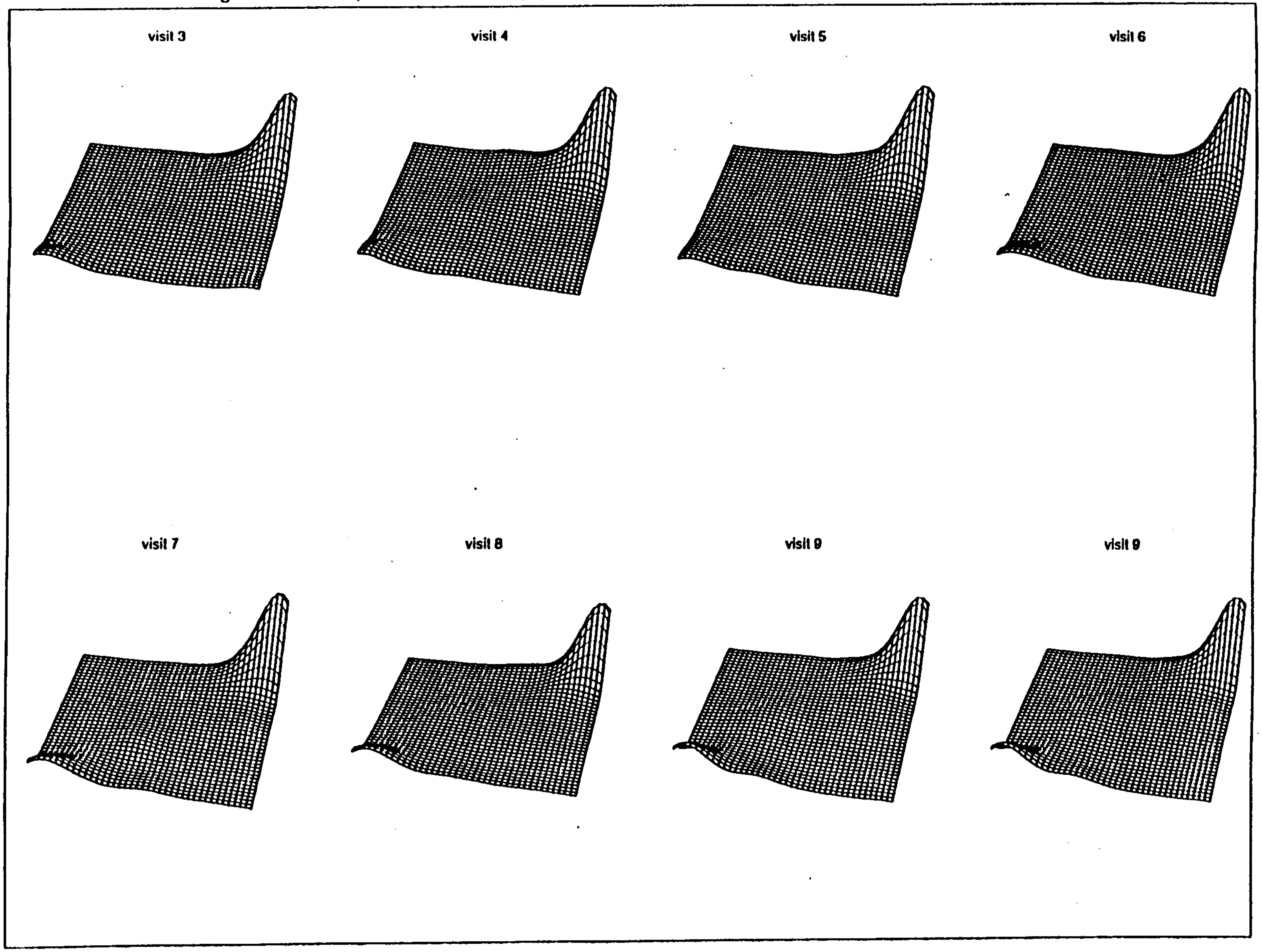
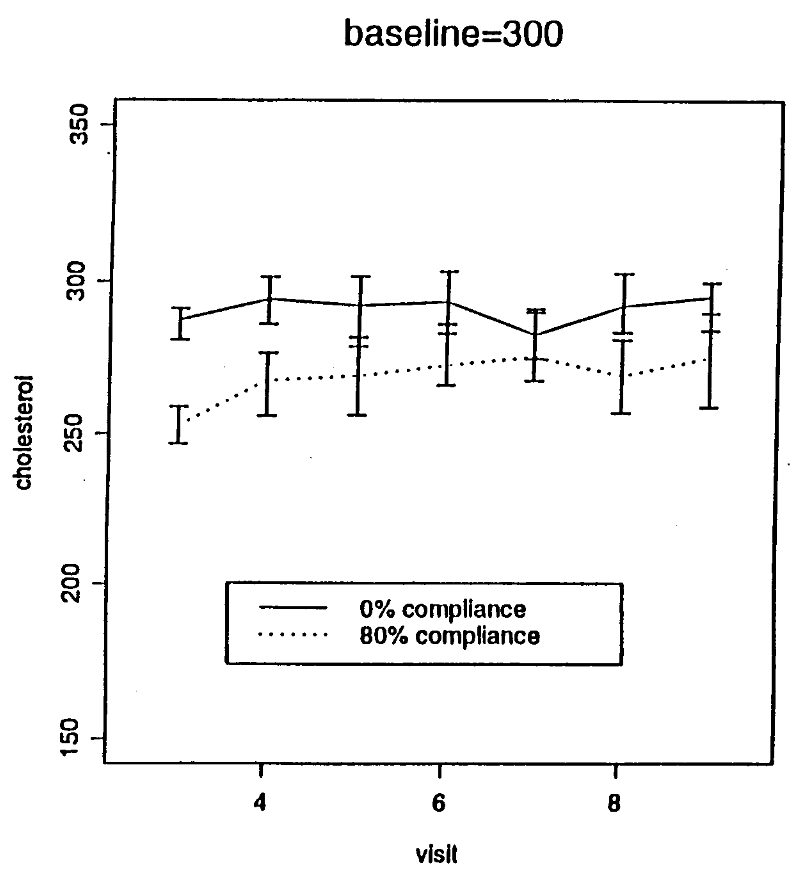
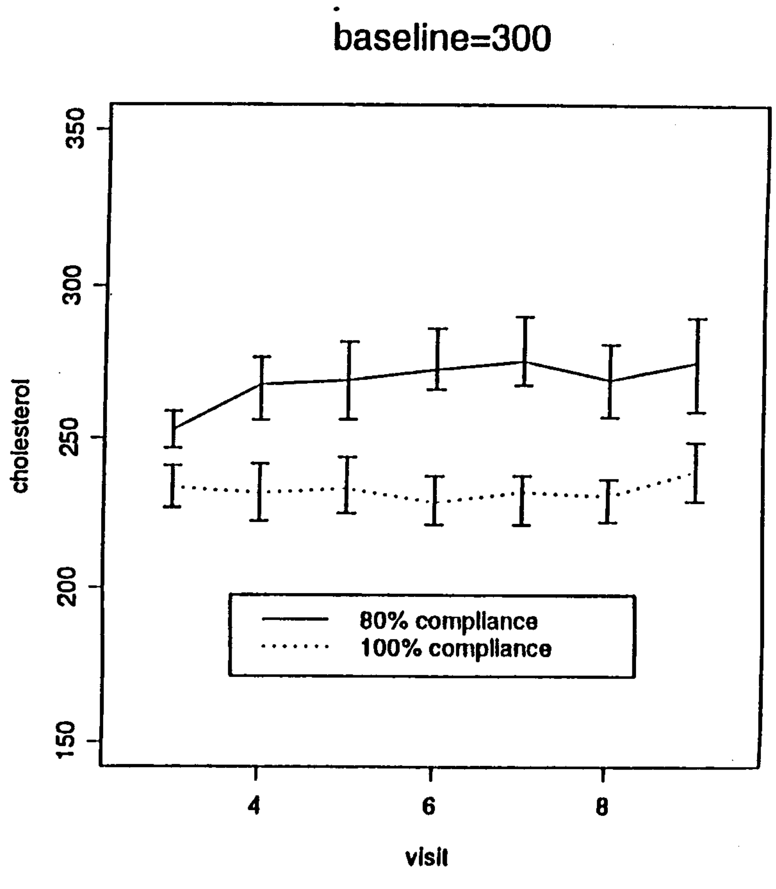


Figure 20. Trends of Mean Cholesterol at Fixed Levels of Compliance

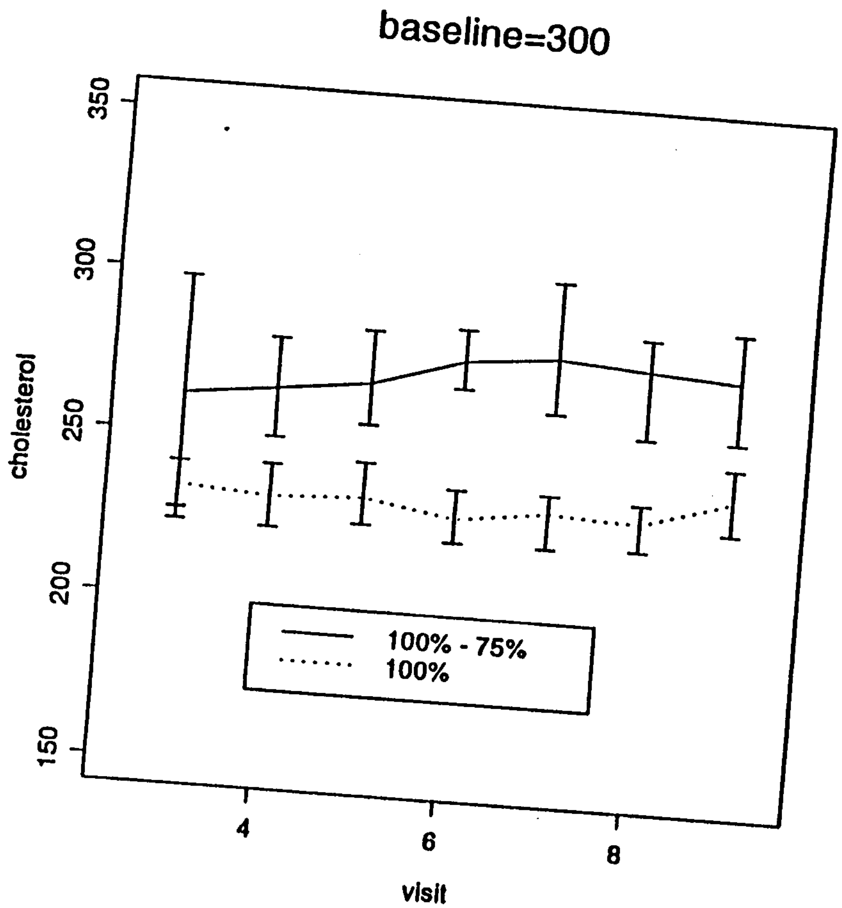


(a)

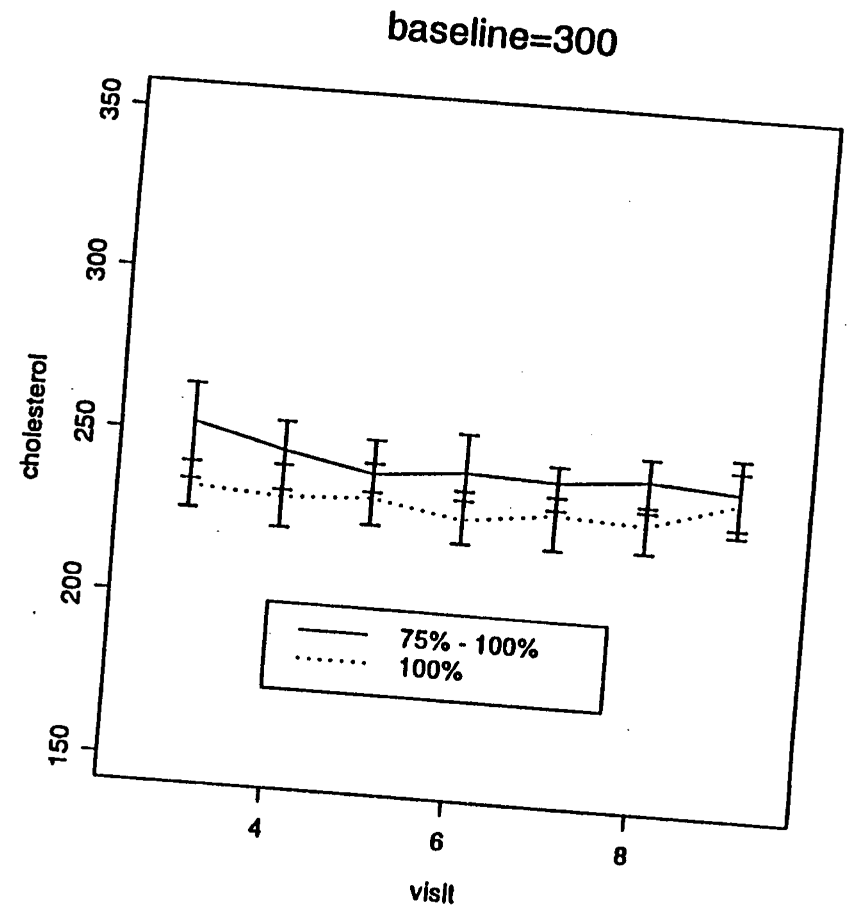


(b)

Figure 21. Trends of Mean Cholesterol at Fixed Levels of Compliance



(a)



(b)

Youth, divine treasure
Already you are gone, never to return
At times I want to cry, but cannot
And times I cry, but did not want to