An Analysis of Socioeconomic and Vital Statistics Using Nonparametric Smoothers

By

Kjell Doksum*, Jeffrey Gould, Robert Hammarstrand
Ashwini Mathur, Eric Vittinghoff

Technical Report No. 400
August 1993

*This research was supported in part by National Science Foundation Grant DMS-91106488.

Department of Statistics
University of California
Berkeley, California 94720

An Analysis of Socioeconomic and Vital Statistics Using Nonparametric Smoothers

Kjell Doksum, Jeffrey Gould, Robert Hammarstrand,*
Ashwini Mathur, Eric Vittinghoff

ABSTRACT

In this paper we consider issues involved in using nonparametric regression and correlation smoothing methods to analyze a moderately large data set consisting of socioeconomic and perinatal vital statistics for about 900 geographic regions in California. The issues considered range from theoretical to applied. Some of the theoretical questions considered are: What is the most efficient smoothing technique? Which technique best handles boundary effects? How do we choose the bandwidth? Some of the applied questions are: What is a good procedure for describing the relationship between socioeconomic and health variables? How can we study the nonlinear relationship between a covariate X_1 and a response variable Y in the presence of confounding covariates that are also nonlinearly associated with Y? After using Monte Carlo simulation to compare boundary corrected Nadaraya-Watson and Gasser-Müller kernel estimators with locally linear estimators based on fixed and variable bandwidths, we conclude that, for our type of data, a suitable smoother is the k-nearest-neighbor version of the locally linear estimator with k chosen by crossvalidating on the central 90% of the covariate values. In generalized nonparametric additive models, we use nonparametric partial regression and correlation methods to study the relationship between a response variable and a covariate after correcting for confounding variables.

KEY WORDS: Nonparametric regression, nonparametric correlation, generalized additive models.

*This research was supported in part by National Science Foundation Grant DMS-91106488.
The authors thank Mithra Moezzi for data management and for helpful comments.

1.INTRODUCTION

There are a growing number of interesting computer data bases consisting of moderately large to large data sets. For these data sets it is possible to use nonparametric regression techniques to analyze local relationships between variables without assuming a particular parametric structure on the joint distribution of the variables.

In this paper we consider the properties of several statistical regression smoothing techniques with the aim of choosing one of them to analyze a real data set. The smoothing procedures we consider are of three varieties: Nadaraya-Watson kernel estimators with and without boundary corrections, Gasser-Müller kernel estimators with and without boundary corrections, and locally linear smoothers with fixed bandwidth as well as locally linear smoothers based on k-nearest-neighbors.

For the case of one covariate X, we obtain information on the properties of these techniques by doing a Monte Carlo study over several models. The models have two different structures for the regression $E(Y|X=x)$ of Y on X : One is quadratic as in Hall and Wehrly (1991) and the other is the "bump" model considered by Härdle (1990) and Gasser, Kneip and Köhler (1991). In addition we consider symmetric and skew marginal distributions for the covariate X. We find that the locally linear techniques perform very well. In particular for regions of X where there are no abrupt changes in $E(Y|X=x)$, a very good choice is the k-nearest-neighbor locally linear technique which consists of finding the values $\hat{\alpha}$ and $\hat{\beta}$ that minimize

$$\sum_{i \in I_k(x)} \left[y_i - (\alpha + \beta x_i)\right]^2 K\left[\frac{x_i - x}{\delta(k)}\right]$$

where $K(u) = 0.75\ (1-u^2)\ I(|u| \leq 1)$ is the Epanechnikov kernel, $I_k(x)$ is the set of indices on the k-nearest-neighbors to x and $\delta(k)$ is the distance from x to its kth nearest neighbor. The estimate of $E(Y|X=x)$ is now $\hat{\alpha} + \hat{\beta}x$.

Locally linear methods have been considered by Stone (1977), Cleveland (1979), Müller (1988), Cleveland and Devlin (1988), among others. Recently their advantages in terms of asymptotic efficiency and reduced boundary effects have been estabilished by Fan (1992, 1993), and Fan and Gijbels (1992), among others.

We consider two different methods for choosing the bandwidth: Cross-validation based on all the data and cross-validation based on the data corresponding to the central 90% of the covariate values. We find that for the locally linear fixed bandwidth smoother, using 100% is generally preferable, while for the k-nearest-neighbor linear smoother, the 90% cross-validation rule is preferable.

In addition to $\mu(x) = E(Y|X = x)$ we use the local regression coefficient $\beta(x) = \mu'(x)$ and the local correlation coefficient $\rho(x) = \dfrac{\sigma_1\beta(x)}{\{\sigma_1^2\beta^2(x)+\sigma^2(x)\}^{1/2}}$, where $\sigma(x) = \{Var(Y|X = x)\}^{1/2}$ and $\sigma_1 = \{Var(X)\}^{1/2}$, to explore the relationship between a response Y and a covariate X. We do a Monte Carlo study of the properties of simple confidence intervals for $\mu(x)$ and $\beta(x)$ and find that procedures based on the locally linear methods give reasonably accurate confidence intervals for $E(\hat{\mu}(x))$ and $E(\hat{\beta}(x))$.

For the analysis of the relationship between a response Y and several covariates $X_1, X_2, \ldots, X_J$, we use an additive model and estimators that involve repeated use of univariate smoothers. In this context we introduce nonparametric partial regression and correlation curves and use these concepts to analyze the Improved Perinatal Outcome Data Management System data set described below using the response variable infant mortality and the three covariates median family income, percentage households on public assistance and percentage families with no husbands.

The data set

The Improved Perinatal Outcome Data Management(IPODM) system (maintained at University of California, Berkeley) provided the data we used for these analyses. The data base consists of perinatal, socioeconomic, and demographic data for California, aggregated by zip code. The perinatal data are for the six year period 1982-1987 and are derived from data for individual births published in the State of California's annual Birth Cohort Files, which encode birth certificate data. The demographic and socioeconomic data are from the 1980 U.S Census.

2. A COMPARISON OF SMOOTHERS IN TERMS OF BOUNDARY EFFECTS

Let Y be a real-valued response variable whose distribution depends on a real-valued covariate X. Then the regression of Y on $X = x$ is defined as

$$\mu(x) = E(Y|X = x) .$$

The curve $\mu(x)$ gives the mean relationship between Y and X at $X = x$ and is the fundamental parametric curve in most parametric regression analysis. When the density $f(x)$ of the predictor variable X has bounded support, kernel smoothers such as the Nadaraya-Watson and Gasser-Müller estimators are known to suffer from "edge effects" at points near the boundaries of the distribution of X. Bias, in particular, can increase substantially, unless $\mu'(x) \approx 0$ near the boundary; variance is also likely to increase, since in general fewer data points are used in this region. These boundary effects persist even when $f(x)$ does not have bounded support, since for x close to the extreme order statistics $x_{(1)}$ and $x_{(n)}$, there will be very few data points on one side of x to average over. Various modifications have been proposed to reduce such edge effects, including special asymmetric boundary kernels, as well as linear combinations of kernels with differing bandwidths, based on the jackknife. We compare three classes of estimators that are insensitive to boundary effects, and find that a locally-linear estimator is particularly efficient.

All three belong to the more general class of linear smoothers. Suppose that $(X_1,Y_1),\ldots,$ (X_n,Y_n) is a random sample of pairs with each (X_i,Y_i) distributed as X,Y. The different estimators (smoothers) can all be written in the form

$$\hat{\mu}(x) = \Sigma w_i(x) Y_i \tag{2.1}$$

where $w_i(x)$ are weights that depend on the $\{X_i\}$ only. The weights are large for indices i such that X_i is "close" to x, and small otherwise.

2.1 Boundary modified and reflexive Nadaraya-Watson smoothers

The basic Nadaraya-Watson estimator is of the form (2.1) with

$$w_i(x) = K\left[\frac{x-X_i}{h}\right] / \sum_{i=1}^{n} K\left[\frac{x-X_i}{h}\right] \tag{2.2}$$

where $h = h_n \to 0$ as $n \to \infty$ and K is a kernel function with $K(t) \geq 0$, $K(-t) = K(t)$, and $\int K^2(t)dt < \infty$. For instance, the Epanechnikov kernel is given by

$$K(u) = 0.75(1 - u^2) I(|u| \leq 1). \tag{2.3}$$

We will consider the same boundary modified Nadaraya-Watson estimates as in Hall and Wehrly (1991). Suppose the $\{X_i\}$ have support [a,b]. For $x \in [a,a+h] \cup [b-h,b]$, the estimate based on (2.1) and (2.2) with kernel (2.3) is typically severely biased. The bias is reduced for $x \in [a,a+h]$ if the kernel (2.3) is replaced on this interval by the boundary kernel

$$K_q(u) = (c_1 u + c_2)K(u)I(-1 \le u \le q) \tag{2.4}$$

where $q = (x-a)/h$, and c_1 and c_2 are the solutions to

$$\int_{-1}^{q} K_q(u)du = 1 \quad \text{and} \quad \int_{-1}^{q} uK_q(u)du = 0.$$

See Rice (1984). A similar boundary kernel K_q is defined for $x \in [b-h,b]$.

Finally a boundary bandwidth is used. That is, in (2.2), with K replaced by K_q near the boundaries, h is replaced by

$$h(x) = \begin{cases} 2h-(x-a) & \text{if } a < x < a+h \\ h & \text{if } a+h \le x \le b-h \\ 2h-(b-x) & \text{if } b-h < x < b. \end{cases} \tag{2.5}$$

Leave-one-out cross-validation is used to select h. See Hall and Wehrly (1991).

A second Nadaraya-Watson type estimator considered by Hall and Wehrly consists of using a reflection method to generate a set of pseudo-observations beyond the boundaries a and b. The Nadaraya-Watson estimator is computed using the original data and the pseudo-observations. See Hall and Wehrly (1991).

2.2 Boundary modified and reflexive Gasser–Müller smoothers

We now consider the fixed covariate case where we condition on the ordered $\{x_i\}$ and write

$$Y_i = \mu(x_i) + \sigma(x_i)\varepsilon_i,\ i = 1,\ldots,n,\ x_1 < x_2 < \cdots < x_n \tag{2.6}$$

where $\sigma(x) = Var(Y|X = x)$ and $\varepsilon_1,\ldots,\varepsilon_n$ are independently distributed with mean 0 and variance 1. Since we can define $\varepsilon_i = [Y_i - \mu(x_i)]/\sigma(x_i)$, the formulation (2.6) does not involve a reduction in generality. See Bhattacharya (1974). The Gasser-Müller estimator (1979, 1984) is of the form (2.1) with

$$w_i(x) = -\left[W\left(\frac{x-s_i}{h}\right) - W\left(\frac{x-s_{i-1}}{h}\right)\right] \tag{2.7}$$

where $W(x) = \int_{-\infty}^{x} K(t)dt$ is an integrated kernel and $s_i = (x_i + x_{i+1})/2$, $i = 1,....,n-1$, $s_0 = x_1$, and $s_n = x_n$. The Gasser-Müller estimator was originally proposed for the fixed design case, but is also appropriate in the random design case, Mack and Müller (1988).

A boundary version of the Gasser-Müller estimator is obtained by replacing K with K_q, given by (2.4), near the boundaries, and h with $h(x)$, given by (2.5). Finally, a reflexive version of the Gasser-Müller estimator is obtained by applying the Gasser-Müller smoother to pseudo-data generated by a reflection method. See Hall and Wehrly (1991).

2.3 Locally linear smoothers

Fan (1992) shows that locally-linear smoothers are less subject to boundary effects than non-modified Nadaraya-Watson and Gasser-Müller estimates. The locally-linear smoothers are defined as follows. Consider p grid points along the x-axis. Let x_0 denote any of the grid points, and let $y = a(x_0) + b(x_0)x$ be the weighted least squares line computed from the data (x_1,y_1),...., (x_n,y_n) with weights $w_1,....,w_n$, where $w_i = K((x_0-x_i)/h)$. The estimate $\hat{\mu}(x_0)$ is $a(x_0) + b(x_0)x_0$. Values of a and b are found at each of the grid points, and the curve $\hat{\mu}(\cdot)$ is completed using standard software to "connect the dots".

In addition, we consider locally-linear k-nearest-neighbor (*knn*) estimators. In this variant of the locally-linear smoother, the global bandwidth h is replaced by the variable local bandwidth $h(x_0)$, given by the distance between x_0 and its kth nearest neighbor.

Finally, for each variant, we consider two cross-validation procedures for selecting the bandwidth, one using all the data, and the other using only the subset corresponding to the central 90% of the values of X.

2.4 Results

To facilitate comparison between the three classes of smoothers, we checked the performance of the two locally-linear methods using the same simulation set-up employed by Hall and Wehrly with the boundary-modified and reflexive Nadaraya-Watson and Gasser-Müller methods. The simulation model is given by $y_i = (x_i - 1/2)^2 + \varepsilon_i$, $i = 1,....,100$, where the $\{\varepsilon_i\}$ are *iid* normal with mean 0 and standard deviation 0.1. For the random design case 400 data sets were generated with the $\{X_i\}$ *iid* U[0,1]; for each of these data sets $\hat{\mu}(x)$ was calculated at 101 evenly spaced grid points on [0,1]. For the fixed design case another 400 data sets were generated, each with $x_i = (i-1/2)/100$, $i=1,....,100$; for these simulations $\hat{\mu}(x)$ was calculated at the design points. Thus we ran 8 sets of calculations, one for each combination of locally-linear smoother and cross-validation method, using the same set of random normal errors $\{\varepsilon_i\}$ and, where applicable, the same random $\{X_i\}$, in each of the 8 combinations. We calculated estimates of mean integrated squared error (MISE) according to formulas given by Hall and Wehrly, along with the summary statistics given in their Table 2. Here we append our results to theirs:

Table 1

Summary Statistics for Estimated MISE, by Smoothing Method
400 Simulations, Uniform X-distribution

	Mean	Median	Standard deviation	Interquartile range
Random design				
N-W	.097	.085	.063	.059
N-WB	.285	.064	3.816	.067
N-WR	.087	.064	.118	.062
G-M	.087	.067	.072	.068
G-MB	.098	.077	.074	.073
G-MR	.083	.065	.066	.066
LL100	.074	.058	.064	.061
LL90	.071	.059	.056	.060
LLknn100	.087	.073	.062	.067
LLknn90	.076	.059	.059	.067
Fixed design				
N-W	.079	.070	.046	.051
N-WB	.066	.056	.049	.047
N-WR	.062	.053	.046	.043
G-M	.059	.052	.039	.042
G-MB	.062	.052	.042	.046
G-MR	.058	.049	.040	.041
LL100	.062	.049	.048	.053
LL90	.064	.050	.048	.055
LLknn100	.080	.065	.055	.064
LLknn90	.068	.054	.052	.053

Note: N-W is the unmodified Nadaraya-Watson estimator, N-WB is Nadaraya-Watson with boundary kernel, and N-WR is Nadaraya-Watson with reflected data; and similarly for the Gasser-Muller (G-M) estimator. LL100 is the locally-linear estimator with cross-validation using 100% of the data, while

LL90 is the same estimator with cross-validation using 90% of the data; LLknn100 and LLknn90 refer to corresponding implementations of the locally-linear variant with *knn* bandwidth. All locally-linear methods use the Epanechnikov kernel (2.3)

The simulation results suggest that the locally-linear methods are quite competitive in both the random and the fixed design cases. In addition, we see that while the performance of the fixed-bandwidth locally-linear smoother is essentially the same with either cross-validation method, the *knn* version does considerably better if the data-fraction is found using cross-validation on only 90% of the data points.

Overall, our conclusion from this table is that fixed bandwidth locally-linear method with bandwidth chosen using all the data is highly efficient. It is very simple, easy to compute and handles the boundary problem as well as the boundary adjusted and reflexive N-W and G-M procedures. With cross-validation using 90% of the data, the *knn* variant is also competitive; simulations reported below show that in some situations it may outperform the fixed-bandwidth version.

Envelope Curves

We also summarize our results using percentile curves as in Hall and Wehrly (1991). At each grid point t_i, we compute the median, 5th and 95th percentiles of the 400 simulated values of $\hat{\mu}(t_i)$. The results are given in Figures 1 and 2. Comparing the median fits with the true regression functions suggests that the differences in bias are trivial; likewise variance in the border region is similar whether 90% or 100% of the data is used to select bandwidth or data fraction. The plots are consistent with Fan and Gijbel's(1992) theoretical demonstration that the bias of locally-linear methods does not increase substantially in the boundary region, and that the variance of the locally-linear estimates in this region is comparable to that of other methods with similarly limited bias. In terms of envelope curves, the fixed-bandwidth locally-linear method with cross validation using all the data performs as well as or better than the competitors, given this simulation set-up. Note that the *knn* variant, which has the virtue of adapting bandwidth to local differences in the density of the predictor, would not be expected to out-perform the fixed-bandwidth variant in either the equispaced fixed design or the uniform random case.

However, the plots suggest that the *knn* version is slightly less biased in the interior of the data, with either cross-validation method.

Skew X-distribution

We also consider the case where the $\{X_i\}$ have a skew distribution, rather than being symmetrically distributed on [0,1]. In the random design case, the $\{X_i\}$ are iid with density

$$f(x) = \begin{cases} 3(1-x)^2 & , x \in [0,1] \\ 0 & , \text{otherwise.} \end{cases}$$

In other respects the model is as before. In the fixed design case, $x_i = F^{-1}[(i-1/2)/100]$, $i = 1,\ldots,100$, where $F(x) = \int_0^x 3(1-t^2)dt = 1-(1-x)^3$, $0 \le x \le 1$. With both designs, $\mu(x)$ is estimated at the grid points $i/100 \le \max\{x_i\}$; note that in the fixed design, $\max(x_i) \equiv 1-(1-99.5/100)^{1/3} = 0.829$. MISE is estimated by the average squared error at these grid points. As with the uniform design case, we calculated estimates of MISE on each of 400 Monte Carlo runs for 8 combinations of locally-linear smoother and cross-validation method. Summary statistics for these values are given in Table 2.

Table 2

Summary Statistics for Estimated MISE, by Smoothing Method
400 Simulations, Skew X-distribution

	Mean	Median	Standard deviation	Interquartile range
Random design				
LL100	.092	.061	.091	.081
LL90	.100	.065	.094	.095
LLknn100	.124	.098	.097	.098
LLknn90	.121	.091	.104	.100
Fixed design				
LL100	.109	.078	.100	.107
LL90	.110	.079	.103	.107
LLknn100	.131	.103	.096	.099
LLknn90	.144	.121	.102	.120

These results strongly suggest that given Hall and Wehrly's quadratic regression function, the locally-linear method with fixed bandwidth performs better than the *knn* variant when the predictor distribution is skew. There is not much difference between the two cross-validation methods, but cross-validation based on 100% of the data performs a little better. Surprisingly, estimated MISE is smaller with the random predictor distribution. However, since the grid points at which MISE is estimated are different for the two cases, they are not directly comparable.

Envelope Curves.

We also give, in Figures 3 and 4, the envelope curves for the skew design. They graphically show how the *knn* methods are biased in the region where the $\{x_i\}$ are sparse. In this region the variable bandwidth $h(x_0)$ is large, and the distribution of the $\{x_i\}$ included in the window $[x_0-h(x_0), x_0+h(x_0)]$ is highly asymmetric. In short the *knn* fits in this region are strongly influenced by points at the left edge of the window; at the same time, there are relatively few points in its center. In such situations the *knn* method is biased toward linearity. In contrast, the locally-linear methods with fixed width are remarkably unbiased in this region. The outer envelope curves show that *both* methods are highly variable where the $\{x_i\}$ are sparse.

Simulations Using Variants of Härdle's Bump Model

In addition to the variants of Hall and Wehrly's model with quadratic mean function, we also consider the "bump" model used in several contexts by Härdle (1990). For this model the regression function is given by $\mu(x) = 1-x+exp(-200(x-c)^2)$. We reset c from 1/2 to 1/3, moving the bump to the left, so that in the skew designs there is enough data in the region of the nonlinearity. The error distribution is *iid* Normal, as in Hardle as well as Hall and Wehrly. But in this case the value of $\sigma^2(x) = 0.1^2$ used in their simulations results in such noisy estimates that we reset this parameter as well, using values 0.05 and 0.025. We also restrict our attention to the fixed predictor distributions and to LL100 and LLknn90, since these estimators did consistently better with Hall and Wehrly's model. The results of these simulations, each using 400 Monte Carlo samples, are given in Table 3.

Table 3
Summary Statistics for Estimated MISE
Uniform and Skew Fixed Variants of Hardle's "Bump" Model
By Error Variance and Smoothing Method

	Mean	Median	Standard Deviation	Interquartile Range
$\sigma^2(x) = 0.05^2$				
Uniform Fixed Design				
LL100	.067	.063	.024	.023
LLknn90	.088	.088	.018	.023
Skew Fixed Design				
LL100	.454	.453	.044	.056
LLknn90	.068	.064	.028	.036
$\sigma^2(x) = 0.025^2$				
Uniform Fixed Design				
LL100	.020	.020	.005	.006
LLknn90	.022	.022	.004	.005
Skew Fixed Design				
LL100	.427	.428	.018	.024
LLknn90	.020	.019	.007	.008

These results show that the locally-linear method with fixed bandwidth performs somewhat better than the *knn* method in the uniform fixed design case, while in the skew fixed design case it performs considerably worse. As with Hall and Wehrly's simulation set-up, the *knn* adaptation would not be expected to help with a uniform random or equispaced fixed design. However, with a skew X-distribution, its potential value is demonstrated. Clearly this method narrows the bandwidth in the region of the bump, as required, and widens it where the data are sparse, on the right. At the same time, this result clearly depends on the shape of Härdle's regression function, which is linear where the data are sparse. This is precisely where the *knn* smoother is biased toward linearity. Envelope curves are given in Figures 5 and 6. The curves confirm that while both estimators perform essentially the same in the uniform fixed design, the *knn* method does much better in the skew fixed design. Taken together with the results using the skew variant of Hall and Wehrly's set-up, Table 3 suggests that the choice between the fixed-bandwidth and *knn* methods depends substantially on the unknown quantity we wish to estimate.

3. REGRESSION, STANDARD DEVIATION AND CORRELATION CURVES

In linear statistical inference, regression coefficients, error standard deviations and correlation coefficients are the key parameters of interest. In nonlinear situations, the regression coefficient is replaced by the *regression coefficient curve*:

$$\beta(x) = d\mu(x)/dx,$$

which gives the rate of change in the conditional mean of Y as x changes. The error standard deviation is replaced by the *conditional standard deviation curve*:

$$\sigma(x) = \{Var(Y \mid X = x)\}^{1/2},$$

and the correlation coefficient is replaced by the *correlation curve*:

$$\rho(x) = \frac{\sigma_1\beta(x)}{\{\sigma_1^2\beta^2(x)+\sigma^2(x)\}^{1/2}},$$

where σ_1 is the standard deviation of X. $\rho(x)$ is a standardized version of the regression coefficient curve $\beta(x)$. It combines the local rate of change $\beta(x)$ with the local standard deviation $\sigma(x)$ to form an invariant local measure of the strength of the relationship between X and Y near $X = x$ which coincides with the Pearson correlation coefficient ρ in linear models. See Bjerve and Doksum (1993) and Doksum, Blyth, Bradlow, Meng and Zhao (1993).

3.1 Estimating local correlation

Estimating $\rho(x)$ is straightforward using locally-linear smoothers. In particular, we set $\hat{\beta}(x) = b(x)$, the slope of the local weighted least squares line at x, as defined in section 2.3; and estimate the local variance by the fitted values of a second locally-linear smooth of the squared residuals $\{(y_i - \hat{\mu}(x_i))^2\}$. Both slope and variance are estimated at the same grid points as $\hat{\mu}$, and by the same smoothing method, but using potentially different bandwidths. Then $\hat{\rho}(x)$ is computed by "plugging in" these component estimates.

We tried several cross-validation techniques for estimating $\rho(x)$ and found that none were satisfactory for the sample sizes and models considered in Section 2 (*cf.* Muler, Stadtmüller and Schmidt (1987)). In the real data example we look both to cross-validation and our sense of plausible smoothness to choose the smoothing parameter k.

4. POINTWISE CONFIDENCE INTERVALS USING LOCALLY LINEAR SMOOTHERS

With all 4 locally-linear smoothers, $\hat{\mu}(x) = \sum w_i(x) Y_i$, where the $\{w_i(x)\}$ are weights depending only on x and X_i (and on the smoothing method). Expressions for the locally-linear slope estimates follow that same pattern. Conditional on the $\{x_i\}$, pointwise standard errors for the locally-linear estimates of mean and slope are immediate. For instance, $se(\hat{\mu}(x)) = \sqrt{\sum w_i^2(x)\sigma^2(x_i)}$. These standard errors are estimated by "plugging in" the locally-linear estimates of the conditional variance.

We used Monte Carlo trials with the variants of Hall and Wehrly's quadratic simulation model to estimate the coverage probabilities for the resulting pointwise confidence intervals at 3 points over the range of x, both before and after correction for bias. In particular, we estimated the coverage probabilities of the biased pointwise 95% confidence intervals

$$[\hat{\mu}(x) - 1.96\, se(\hat{\mu}(x)),\quad \hat{\mu}(x) + 1.96\, se(\hat{\mu}(x))]$$

by the average of $I[|t_\mu(x)| \le 1.96]$ over Monte Carlo samples, where $t_\mu(x) = [\hat{\mu}(x) - \mu(x)]/se(\hat{\mu}(x))$. Likewise, the coverage probability of the unbiased interval

$$[\hat{\mu}(x) - (\mathrm{E}\hat{\mu}(x) - \mu(x)) - 1.96\, se(\hat{\mu}(x)),\quad \hat{\mu}(x) - (\mathrm{E}\hat{\mu}(x) - \mu(x)) + 1.96\, se(\hat{\mu}(x))]$$

was estimated by the average of $I[|t_\mu^*(x)| \le 1.96]$, where $t_\mu^*(x) = [\hat{\mu}(x) - \mathrm{E}\hat{\mu}(x)]/se(\hat{\mu}(x))$. With a simulation model, $\mathrm{E}\hat{\mu}(x) = \sum w_i(x)\mu(x_i)$ is straightforward to compute. Coverage probabilities of the biased and unbiased intervals for $\beta(x) = \mu'(x)$ were estimated by the same strategy. Similarly, we attempted to quantify the effect of estimating $\sigma^2(x)$ on our confidence intervals by recomputing estimated coverage probabilities using the same series of simulated data sets, but substituting the true variance for the estimates.

The results of 400 Monte Carlo runs with the 4 variants of Hall and Wehrly's model are summarized in Table 4. For each design, cross-validated smoother, and location, four estimates are given: those in the first row are not corrected for the bias in $\hat{\mu}(x)$, while those in the second have been corrected for this source of error; similarly, the figures in parentheses result from recomputing these biased and unbiased estimates using the known variance. The same cross validation rules (as indicated in the table) were used for $\mu(x)$ and $\beta(x)$.

Table 4
Estimated Coverage Probabilities (%) of 95% Confidence Intervals for Mean and Slope by Simulation Design, Smoothing Method, and Location, in 400 Simulations

Design	$\mu(x)$			$\beta(x)$		
		x			x	
Smoothing Method	0.25	0.50	0.75	0.25	0.50	0.75
Random Uniform						
LL100	86(87)	86(88)	86(88)	92(95)	93(95)	92(93)
	93(95)	94(93)	93(94)	93(95)	93(95)	92(94)
LL90	88(89)	88(89)	88(90)	92(95)	94(95)	91(92)
	93(95)	93(93)	92(94)	93(94)	95(96)	92(94)
LLknn100	89(92)	87(88)	89(90)	88(89)	93(95)	86(87)
	92(95)	92(92)	93(94)	93(94)	93(95)	92(93)
LLknn90	91(94)	91(92)	90(92)	91(93)	93(95)	90(91)
	92(95)	93(94)	93(94)	93(94)	93(95)	92(94)
Fixed Uniform						
LL100	87(89)	87(90)	87(89)	93(94)	91(94)	90(93)
	93(95)	95(96)	91(93)	93(95)	91(94)	91(94)
LL90	88(89)	88(91)	89(90)	92(95)	92(94)	92(94)
	93(96)	94(96)	92(94)	92(95)	92(94)	92(94)
LLknn100	92(94)	91(92)	89(92)	90(91)	93(95)	90(92)
	94(96)	94(96)	92(94)	93(95)	93(95)	94(95)
LLknn90	92(95)	93(94)	91(93)	93(96)	94(96)	91(93)
	93(95)	94(96)	93(95)	94(96)	94(96)	92(93)
Random Skew						
LL100	80(81)	88(91)	78(92)	92(92)	81(84)	71(90)
	95(95)	90(93)	77(93)	94(96)	92(94)	77(95)
LL90	85(85)	88(91)	77(93)	93(93)	85(87)	75(95)
	94(95)	89(93)	77(93)	95(96)	91(94)	77(97)
LLknn100	89(90)	84(84)	77(77)	92(93)	63(67)	35(36)
	94(96)	92(93)	90(94)	93(94)	93(93)	89(90)
LLknn90	85(88)	85(87)	73(74)	92(93)	67(69)	37(38)
	93(95)	91(93)	90(94)	92(94)	93(95)	90(92)
Fixed Skew						
LL100	81(82)	89(92)	74(93)	92(93)	81(85)	71(90)
	94(95)	91(93)	73(93)	94(95)	92(94)	74(96)
LL90	86(86)	90(92)	75(93)	92(94)	84(86)	78(93)
	93(95)	89(92)	75(93)	94(95)	89(93)	80(96)
LLknn100	90(91)	83(87)	70(76)	92(94)	67(68)	36(39)
	95(97)	92(94)	90(94)	93(94)	94(95)	90(94)
LLknn90	85(87)	86(89)	65(66)	92(93)	61(62)	31(33)
	94(95)	92(94)	89(94)	93(94)	92(94)	90(94)

For the uniform designs, the estimated coverage probabilities of the uncorrected intervals using estimated variance are 2-9 percentage points too low for $\mu(x)$, and 1-9 points low for $\beta(x)$. On average the coverage probabilities for the slope are closer to 95%. Correction for bias in $\hat{\mu}(x)$ reduces the shortfall by at least half, with smaller improvements resulting from substitution of known for estimated variance. These improvements are comparable for all four smoothers.

For the skew designs, however, the picture is more complex. Most obviously, the estimated coverage probabilities of the totally uncorrected intervals, both for mean and slope, are much too low at $x = 0.75$, and still far from satisfactory at $x = 0.50$; and in the case of the slope confidence intervals, the shortfall is worse with the *knn* methods than with their fixed bandwidth counterparts. At these points, where the data are sparse, most of the deficit with the *knn* smoothers results from the bias in $\hat{\mu}(x)$, whereas with the fixed bandwidth methods most comes from estimating $\sigma^2(x)$. Envelope plots shown in Figures 3 and 4 show the first kind of bias clearly; likewise envelope plots of fixed bandwidth estimates of $\sigma^2(x)$ (not shown) demonstrate the second. The *knn* estimators of variance are also biased low, but less severely. However, these results may depend on the fact that with this simulation $\hat{\mu}(x)$ is curved, whereas $\sigma^2(x)$ is linear (in fact, constant): the point being that where the data are sparse the *knn* estimates are more strongly biased towards linearity than the fixed bandwidth results.

These results suggest that our simply-computed confidence intervals for mean and slope have reasonable coverage probabilities in regions where the data are sufficiently dense, but perform poorly when this condition does not hold. In addition, they suggest that the *knn* smoother may be better suited to the estimation of variance, which is commonly modeled more simply than the mean, slope, or local correlation: but for these functions, on which we are reluctant to impose simple models, the fixed bandwidth method may often be better suited.

Note that the bias correction used here is only computable in Monte Carlo trials where $\mu(x)$ is known. However, our bias corrected results apply to confidence intervals for $E(\hat{\mu}(x))$ and $E(\hat{\beta}(x))$ rather than $\mu(x)$ and $\beta(x)$. This is consistent with the "descriptive" approach to statistics where estimators estimate what they estimate rather than a predetermined population parameter (c.f. Bickel and Lehmann (1975)). For example, in this approach the fact that the sample median may be a biased estimate of the population mean is not of concern. In our case this descriptive approach is appropriate when $E(\hat{\mu}(x))$ and $E(\hat{\beta}(x))$ reflect interesting aspects of the relationship between a response Y and a covariate X.

5. ADDITIVE MODELS AND PARTIAL CORRELATION CURVES

Let Y be a real valued response variable whose distribution depends on a real valued vector $\mathbf{X} = (X_1, X_2, \ldots, X_J)$. Then the regression of Y on $\mathbf{X} = \mathbf{x}$ is defined as

$$\mu(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x}).$$

Here we assume an additive model for $\mu(\mathbf{x})$: that is, for i.i.d vectors $(X_{i1}, \ldots, X_{iJ}, Y_i)$,

$$Y = \alpha + \sum_{j=1}^{J} m_j(x_j) \tag{5.1}$$

where the $\{m_j(x)\}$ are smooth functions such that that expectation of $E(m_j(X_{ij})) = 0$, and $\varepsilon_1, \ldots, \varepsilon_n$ are independently distributed with mean 0 and variance σ^2. The functions $\{m_j(X_j)\}$ are identifiable up to an arbitrary constant.

The constant α and the functions $\{m_j(\cdot)\}$ can be estimated using the backfitting algorithm of Friedman and Stuetzle (1981) (c.f. Hastie and Tibshirani (1989)). This algorithm has the following form. In the initialization, set $\hat{\alpha} = \bar{y} = n^{-1}\sum_{i=1}^{n} y_i$, and $\hat{m}_j(\cdot) \equiv 0$, $j = 1, \ldots, J$. Then each $m_j(\cdot)$ is re-estimated in turn, with the new value given by a smooth of the adjusted response

$$\tilde{y}_{ij} = y_i - \hat{\alpha} - \sum_{k \neq j} \hat{m}_k(x_{ik})$$

on the corresponding predictors $\{x_{ij}\}$. Iteration continues over this cycle of updates from $j=1$ to J until changes in $RSS = \sum_{i=1}^{n} [y_i - \hat{\alpha} - \sum_{j=1}^{J} \hat{m}_j(x_{ij})]^2$ are sufficiently small. Note that while the $\{\hat{m}_j(\cdot)\}$ are not necessarily unique, the estimated regression surface is. Nonetheless, we found in our application that the estimated curves corresponding to each predictor were not affected by the order in which the additive functions were updated in the backfitting algorithm.

The algorithm depends on the assumption of additivity, since if

$$\tilde{Y}_{ij} = Y_i - \alpha - \sum_{k \neq j} m_k(x_{ik}) = m_j(x_{ij}) + \varepsilon_i$$

then all the information about $m_j(\cdot)$ is contained in the $\{(\tilde{Y}_{ij}, X_{ij})\}$. The additive model also allows us to define a partial regression coefficient curve and a partial local correlation curve. We define the partial regression coefficient curve as

$$\beta_j(x_j) = \partial\mu(\mathbf{x})/\partial x_j = m'(x_j)$$

and the partial correlation curve as

$$\rho_{YX_j}(x_j) = \frac{\sigma_j \beta_j(x_j)}{\{\sigma_j{}^2\beta_j{}^2(x_j) + \sigma^2\}^{1/2}}$$

where σ_j is the standard deviation of X_j. As in the case of a single covariate, $\rho_{YX_j}(x_j)$ is a standardized version of partial regression coefficient curve $\beta_j(x_j)$. Also note that $Var(\tilde{Y} | X_j) = Var(Y | \mathbf{X}) \equiv \sigma^2$. This allows us to estimate the partial curves using the techniques for a single covariate. The partial correlation curve allows us to study the strength of the local relationship between a response variable and a particular covariate adjusting for other covariates, in the case of a homoscedastic additive model (5.1). Unlike the $\{\hat{m}_j(\cdot)\}$, which can only be estimated up to an additive constant, the partial correlation curves do not depend on the location of the dependent variable.

6. ANALYSIS OF THE IPODM DATA SET USING ADDITIVE MODELS

Description

The data set consists of aggregate vital statistics and measures of infant health for zipcode areas in California. In particular, we analyze the relationship of the local infant mortality rate to median family income, the proportion of families with no husband, and the proportion of households on public assistance; in addition we examine how this relationship differs in urban and rural areas. We define infant mortality rate as the number of deaths before age 1 per 1000 live births; to ensure the stability of our results, we limit the analysis to zipcode areas which have at least 150 births. In addition, we classify a zipcode area as rural if the percentage of rural population is at least 75%; otherwise we call it urban. The data used in our analysis consist of 921 observations, 667 of which are for urban areas.

We give summary measures of the 4 variables of interest in Table 5.

Table 5
Mean and Standard Error of Demographic Measures, by Area

Measure	Overall	Urban	Rural
Infant deaths per 1000 live births	6.75 (0.09)	6.5 (0.09)	7.32 (0.29)
Median family income	$21719 (226.85)	$22841 (275.80)	$17981 (253.96)
Percent households on public assistance	9.77 (0.22)	9.28 (0.27)	11.41 (0.34)
Percent families without husbands	6.36 (0.11)	6.81 (0.14)	5.01 (0.12)

The differences between rural and urban areas with respect to infant mortality median family income, percent of households on public assistance and percent of families without husbands were statistically significant using a univariate t-test(p-value < 0.01).

Analysis

We used a locally-linear smoother in our analysis of the IPODM data set. Relying on the simulation results we selected the *knn* variant to estimate the mean, slope and variance functions because the distributions of our predictors are skewed, and because we did not expect to find sharp nonlinearities in the tails. For these estimates we set k equal to the maximum of the 90% cross-validated value and the nearest integer to $0.7n$, since smaller values give implausibly rough estimates. In the additive model, these initial values were used until convergence. Then new values of k were chosen by the same rule for each adjusted variate at convergence, and final estimates of the $\{\hat{m}_j(\cdot)\}$ computed using the new values of the smoothing parameter.

In exploratory analyses of the combined rural and urban data we smoothed infant mortality rates on the three predictors one at a time. The resulting univariate estimates of mean and local correlation are shown in Figure 7. We found that on average the infant mortality rate declines rapidly, from 10.5 to 6.5 per 1000, as median family income increases from less than $10,000 to $20,000, and somewhat less rapidly thereafter. This nonlinear relationship is reflected in the univariate local correlation curve, with somewhat stronger (negative) correlation in the region of lower median family income. The univariate relationship of infant mortality to the proportion of families with no husband is also nonlinear: there is virtually no relationship among zipcode areas with fewer than 6% of such families, but a steep rise in the average infant mortality rate, from 6 to 14 per 1000, as the proportion of these families rises to 30%. In contrast, the increase in average infant mortality rate with the proportion of households on public assistance is almost linear, as both the mean and local correlation functions suggest.

In turn we used an additive model to adjust our estimate of the association of each of these predictors with the infant mortality rate for the effect of the other two. These results are shown in Figure 8. In the additive model, the relationship of infant mortality with all three predictors is attenuated, but qualitatively similar to the result of the univariate smooth. That is, the nonlinearities of both the mean and local correlation estimates remain, but $\hat{m}(\cdot)$ is "flattened" and $\hat{\rho}(\cdot)$ is closer to 0. Clearly the association of each predictor with the infant mortality rate is to some extent confounded by the other two. The apparent differences in the width of confidence bands for the additive results are the artifact of scale.

We also fit univariate and additive models to the subsets of the data corresponding to rural and urban zipcode areas. Univariate and additive results for the rural areas are shown in Figures 9 and 10 respectively. This subset of the data differs in some respects from the overall picture. In particular, the local correlation of the infant mortality rate with the proportion of households on public assistance is weaker for values of the predictor below 20% before adjustment, and everywhere virtually nil after adjustment. In contrast, the local correlation of the infant mortality rate with the proportion of families with no husband was somewhat stronger for relatively small and large values of the predictor, both before and after adjustment, although the range of the predictor is considerably shorter among the rural zipcode areas. However, the association of infant mortality with median family income is essentially the same as in the combined data.

Since the majority of zipcode areas are urban, it is not surprising that results for this subset strongly resemble the estimates obtained using all the data. These univariate and additive estimates are shown in Figures 11 and 12 respectively. Only a few slight contrasts are in evidence. The correlation of the infant mortality rate with median family income is more strongly attenuated in the additive model among urban zipcode areas than in the combined data. The local correlation of the infant mortality rate with the proportion of households on public assistance is relatively linear, and suffers less attenuation in the additive model than with the combined data. Adjustment has the effect on our estimate of the association of the infant mortality rate with the proportion of families with no husband among urban zipcode areas of fairly sharply attenuating the univariate result for values of the predictor smaller than 10%. But these are not large differences. In general, stratification of the data into rural and urban subsets does not suggest any remarkable interactions of this predictor with the other three.

Clearly, statistical analysis based on ecological data is difficult to interpret. Aggregate zipcode area statistics are at best indirect measures, both of the conditions of the particular families where infant mortality does or does not occur, and of social conditions that might affect this outcome, including the accessibility and quality of medical care. Nonparametric smoothers are useful tools for exploring the relationship of these aggregate measures, since parametric forms are not imposed *a priori*; similarly, additive models, together with the local correlation measure, enable us to apply nonparametric smoothers in the important context of multiple regression, and thus control for the effect of covariates (including nonlinear effects). At the same time statistical inference using these tools is not yet well developed, so that our statistical methods, like our data, are better suited for generating hypotheses than to testing them.

References

Bhattacharya, P.K. (1974). "Convergence of sample paths of normalized sums of induced order statistics." *The Annals of Statistics*, 2, 1034-1039.

Bickel, P.J. and Lehmann, E.L. (1975). "Descriptive statistics for nonparametric models. I." *The Annals of Statistics*, 3, 1038-1044.

Bjerve, S. and Doksum, K.A. (1993). "Correlation curves: measures of association as functions of covariate values." *The Annals of Statistics*, 21. To appear, June 1993.

Cleveland, W.S. (1979). "Robust locally weighted regression and smoothing scatterplots." *Journal of the American Statistical Association*, 74, 829-836.

Cleveland, W.S. and Devlin, S.J. (1988). "Locally weighted regression: An approach to regression analysis by local fitting." *Journal of the American Statistical Association*, 83, *596-610.*

Doksum, K.A., Blyth,S., Bradlow, E., Meng, X.L. and Zhao, H. (1993). "Correlation curves as local measures of variance explained by regression." *Journal of the American Statistical Association*, 88. *To appear.*

Fan, J. (1992). "Design adaptive nonparametric regression." *Journal of the American Statistical Association*, 87, *998-1004.*

Fan, J. and Gijbels, I. (1992). "Variable bandwidth and local linear regression smoothers." *The Annals of Statistics*, 20, 2008-2036.

Fan,J. (1993). "Local linear regression smoothers and their minimax efficiencies." *The Annals of Statistics*, 21. To appear.

Friedman, J.H., and Stuetzle, W. (1981). "Projection pursuit regression." *Journal of the American Statistical Association*, 76, 817-823.

Gasser, T. and Müller, H.G. (1979). "Kernel estimation of regression functions." *Smoothing techniques for curve estimation : proceedings of workshop held in Heidelberg, April 2-4, 1979.* edited by Gasser, T. and Rosenblatt, M.

Gasser, T. and Müller, H.G. (1984). "Estimating regression functions and their derivatives by the kernel method." *Scandinavian Journal of Statistics*, 11, *171-185.*

Gasser, T., Kneip, A. and Köhler, W. (1991). "A flexible and fast method for automatic smoothing." *Journal of the American Statistical Association*, 86, 643-652.

Hall, P. and Wehrly, T.E (1991). "A geometrical method for removing edge effects from kernel-type nonparametric regression estimators." *Journal of the American Statistical Association*, 86, *665-673.*

Hastie, T.S. and Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman and Hall, London, New York.

Härdle, W. (1990). *Applied Nonparametric Regression.* Econometric Society Monographs. Cambridge University Press. Cambridge; New York.

Mack, Y.P. and Müller, H.G. (1988). "Convolution type estimators for nonparametric regression." *Statistics and Probability Letters*, 7, 229-239.

Müller, H.G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics. Springer. Berlin; New York.

Müller, H.G., Stadtmüller, U. and Schmidt, T. (1987). "Bandwidth choice and confidence intervals for derivatives of noisy data." *Biometrika*, 74, 743-750.

Rice, J.A. (1984). "Boundary modification for kernel regression." *Communications in Statistics*, Series A, 13, 893-900.

Stone, C.J. (1977). "Consistent nonparametric regression." *The Annals of Statistics*, 5, 595-620.

Figure 1: The 5th, 50th and 95th percentile of
400 estimates of m(x). The true curve is represented by a solid line.
Random Design Case

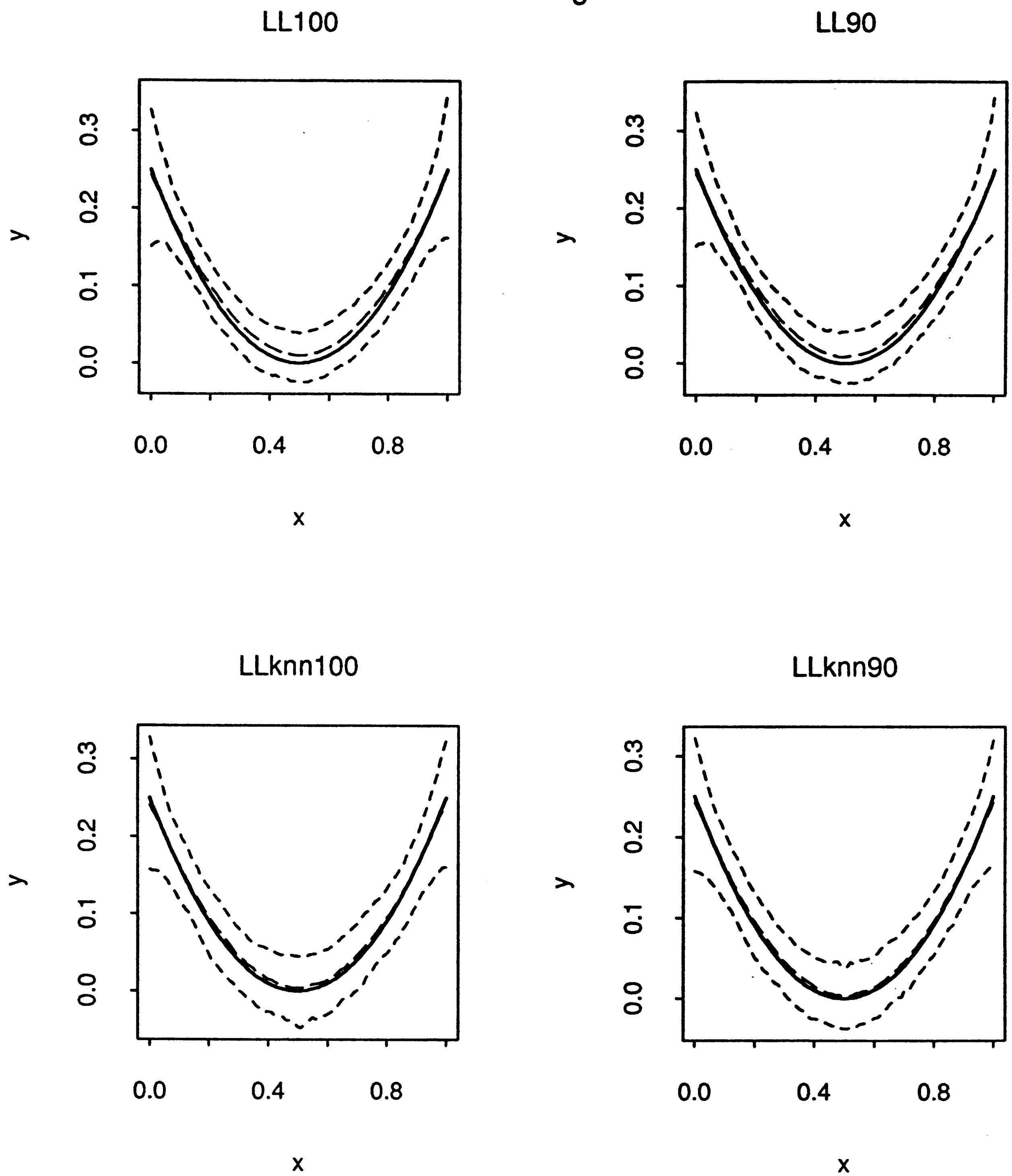

Figure 2: The 5th, 50th and 95th percentile of
400 estimates of m(x). The true curve is represented by a solid line.
Fixed Design Case

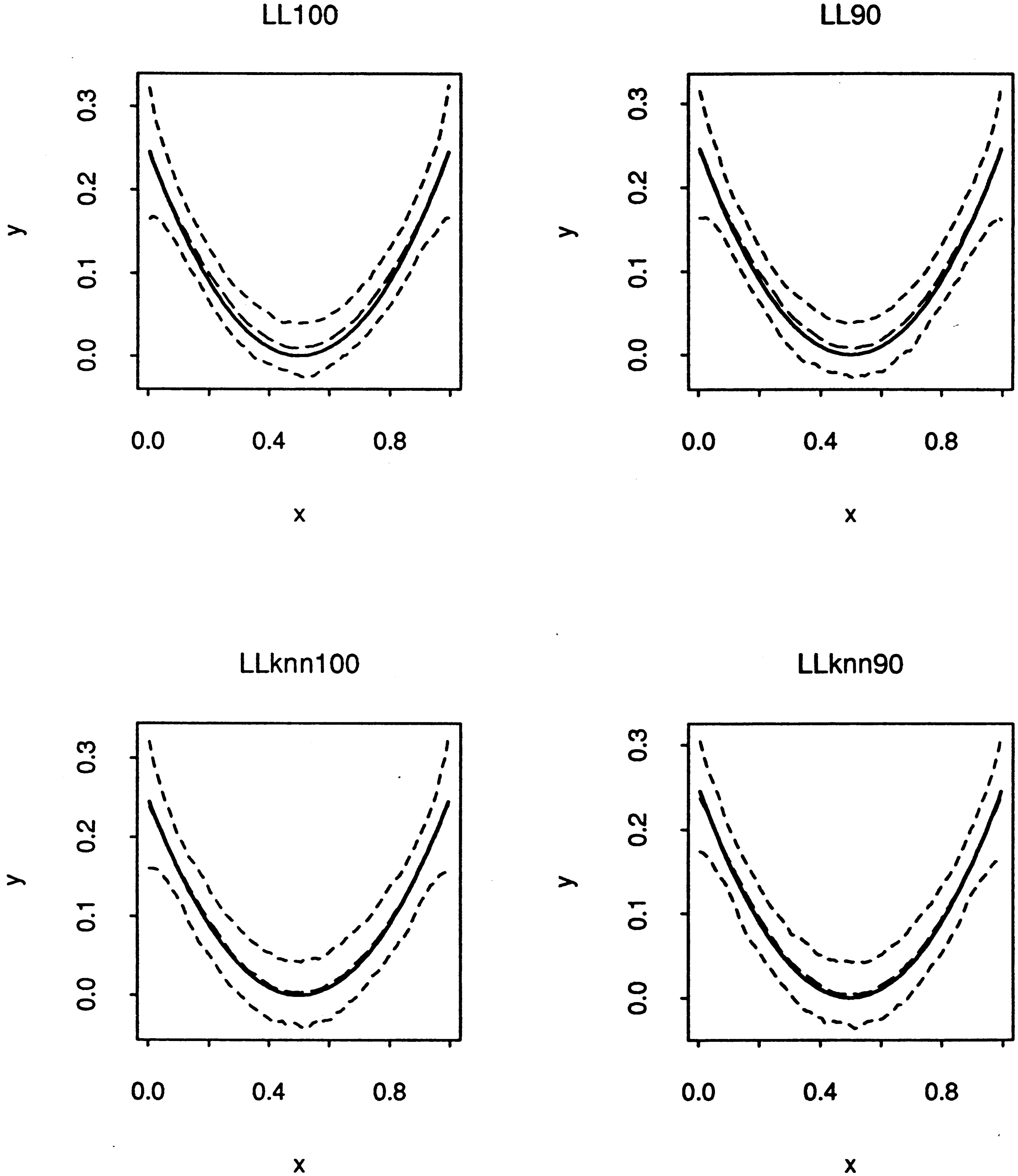

Figure 3: The 5th, 50th and 95th percentile of
400 estimates of m(x). The true curve is represented by a solid line.
Skew Random Design Case

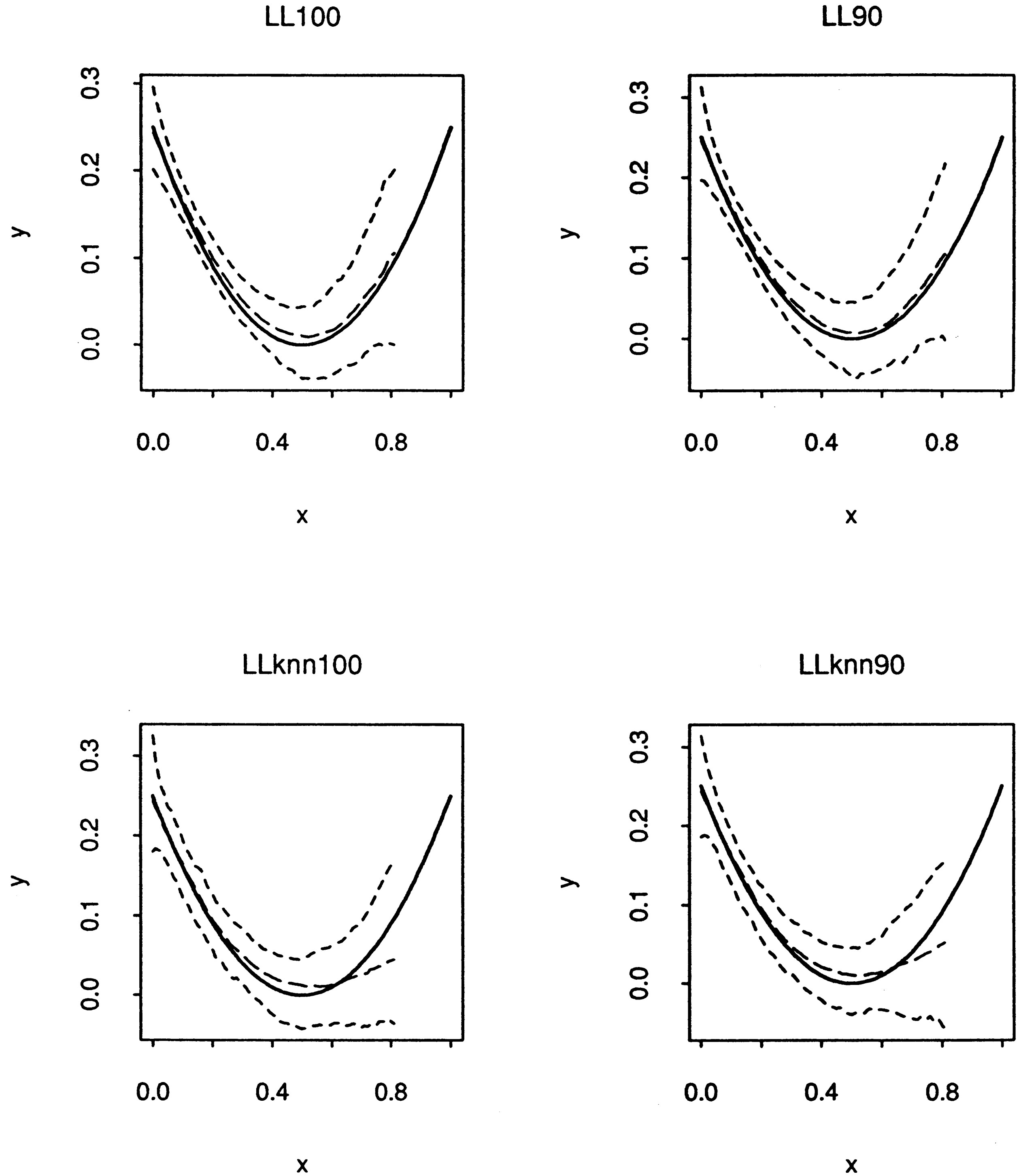

Figure 4: The 5th, 50th and 95th percentile of
400 estimates of m(x). The true curve is represented by a solid line.
Skew Fixed Design Case

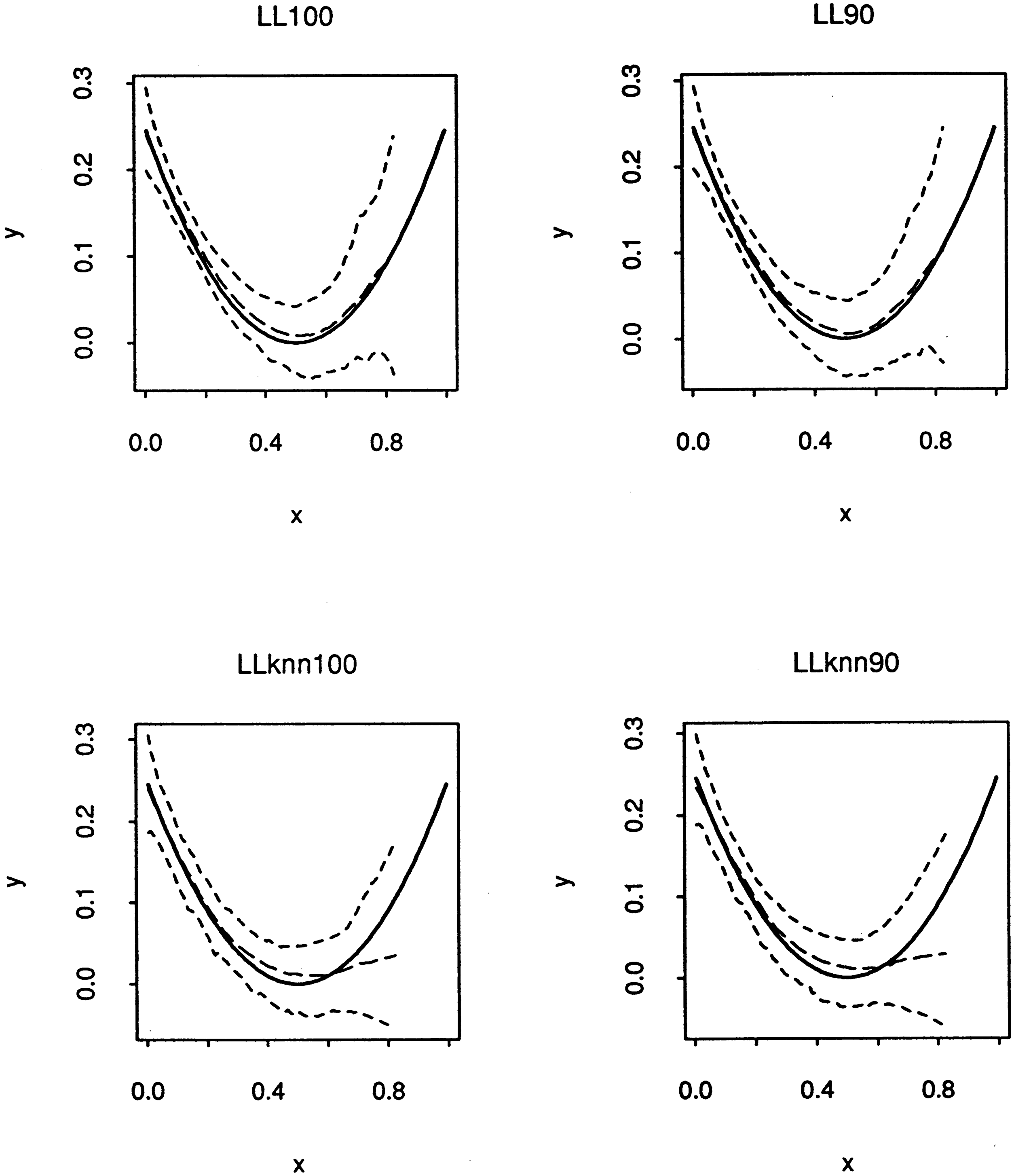

Figure 5: The 5th, 50th and 95th percentile of
400 estimates of m(x). The true curve is represented by a solid line.
Bump Model with Standard Deviation 0.05

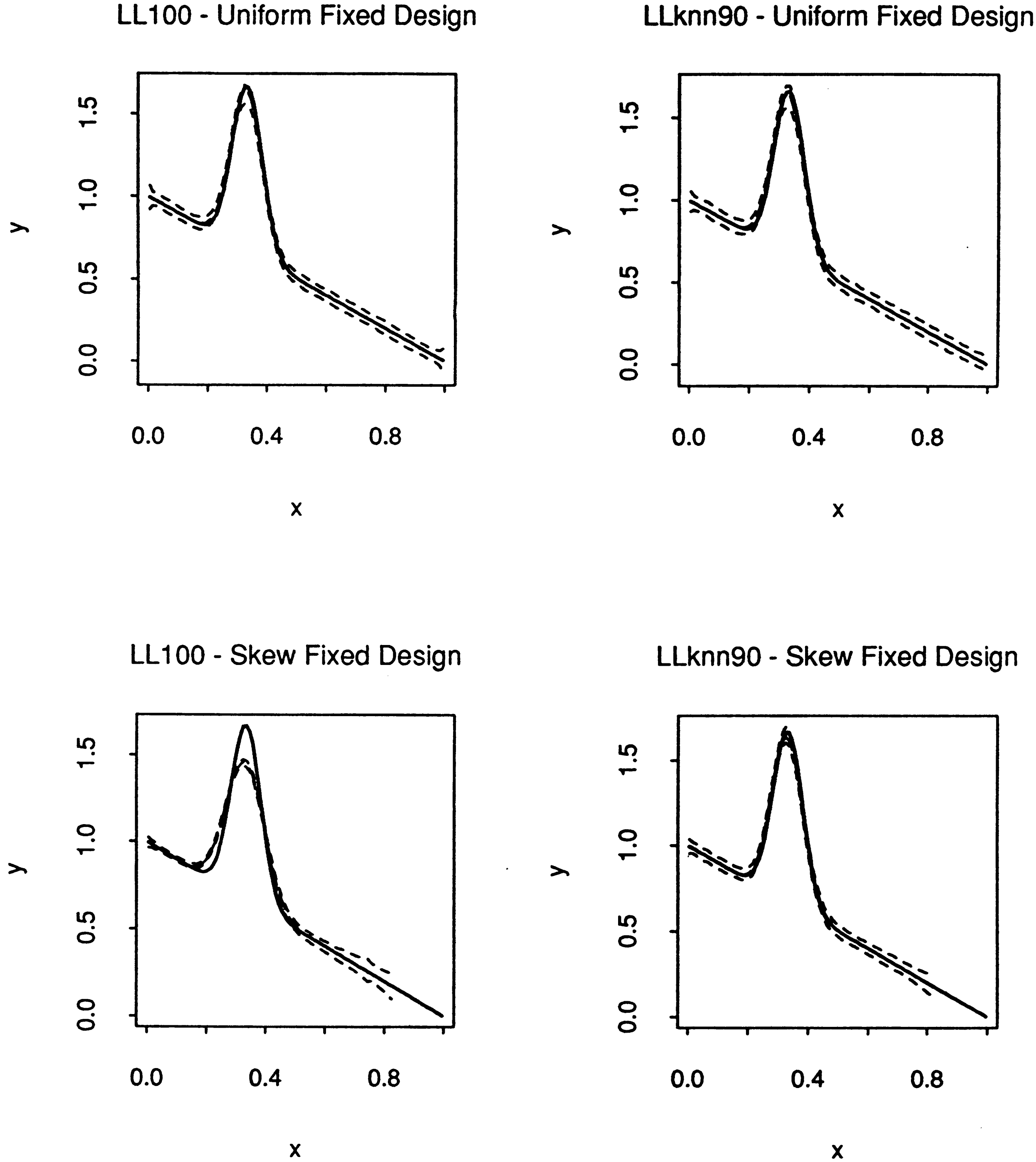

Figure 6: The 5th, 50th and 95th percentile of
400 estimates of m(x). The true curve is represented by a solid line.
Bump Model with Standard Deviation 0.025

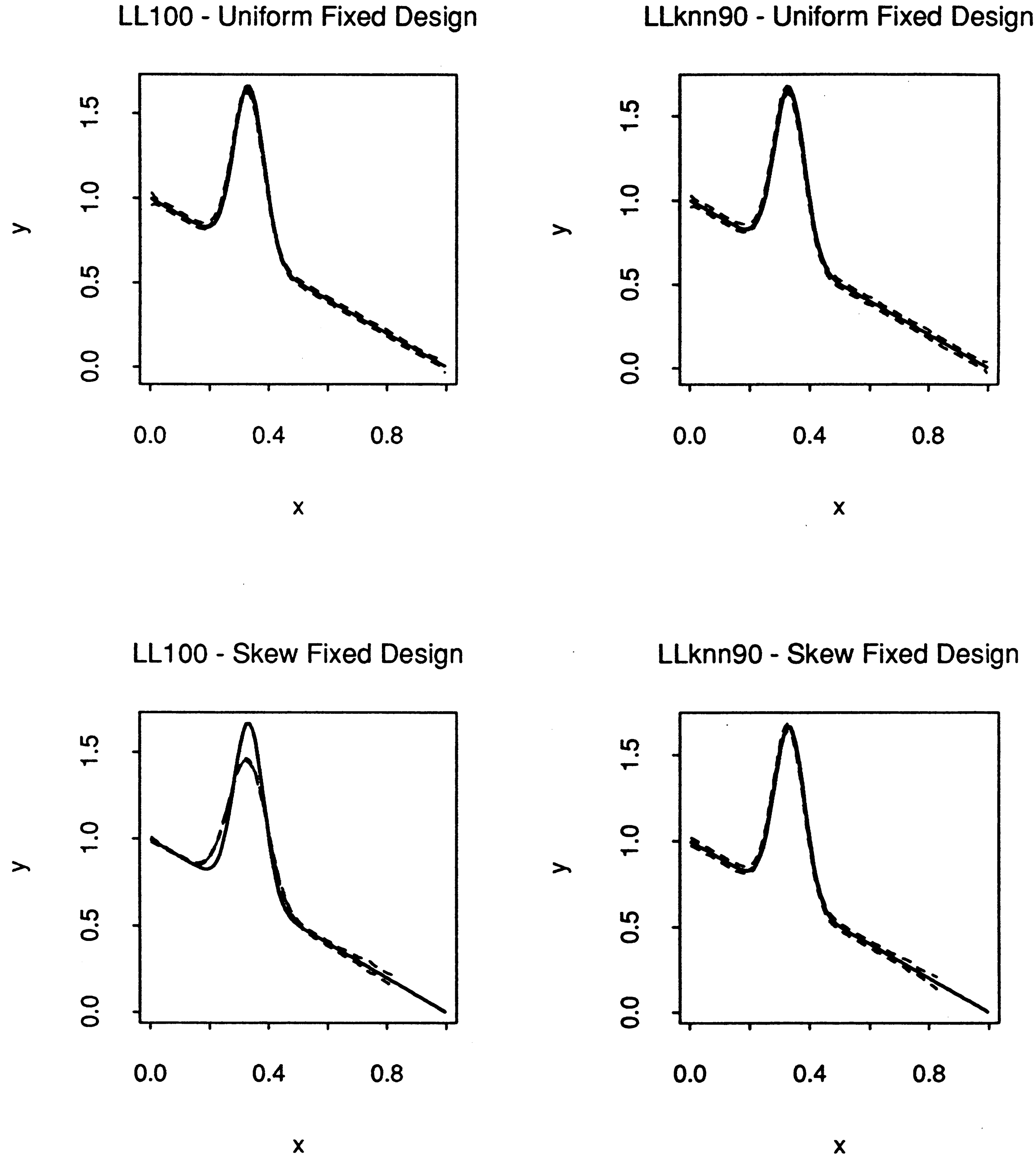

Figure 7: Local regression and correlation
Urban and & Rural areas combined(Univariate model)

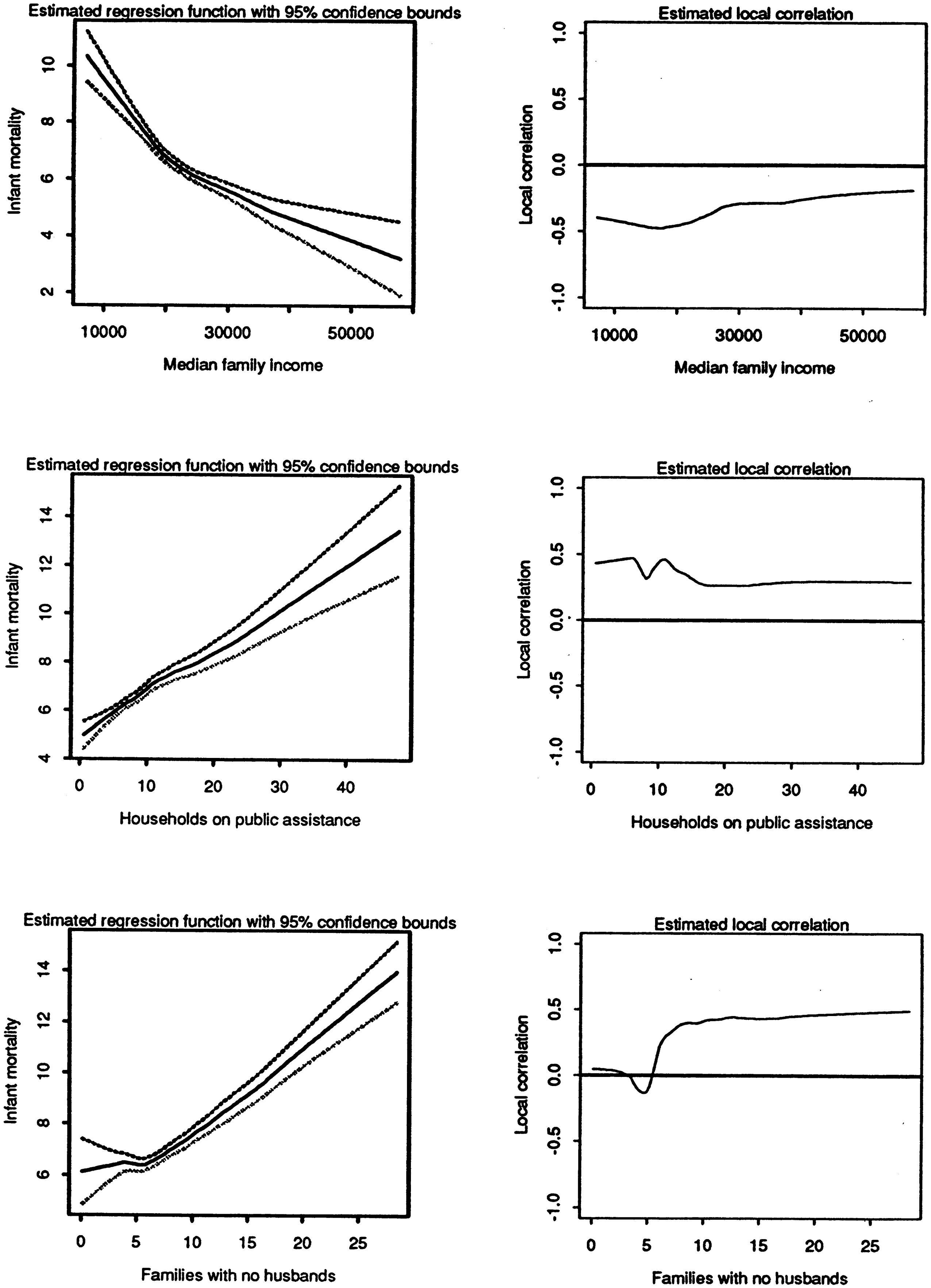

Figure 8: Partial local regression and correlation
Urban and & Rural areas combined(Additive model)

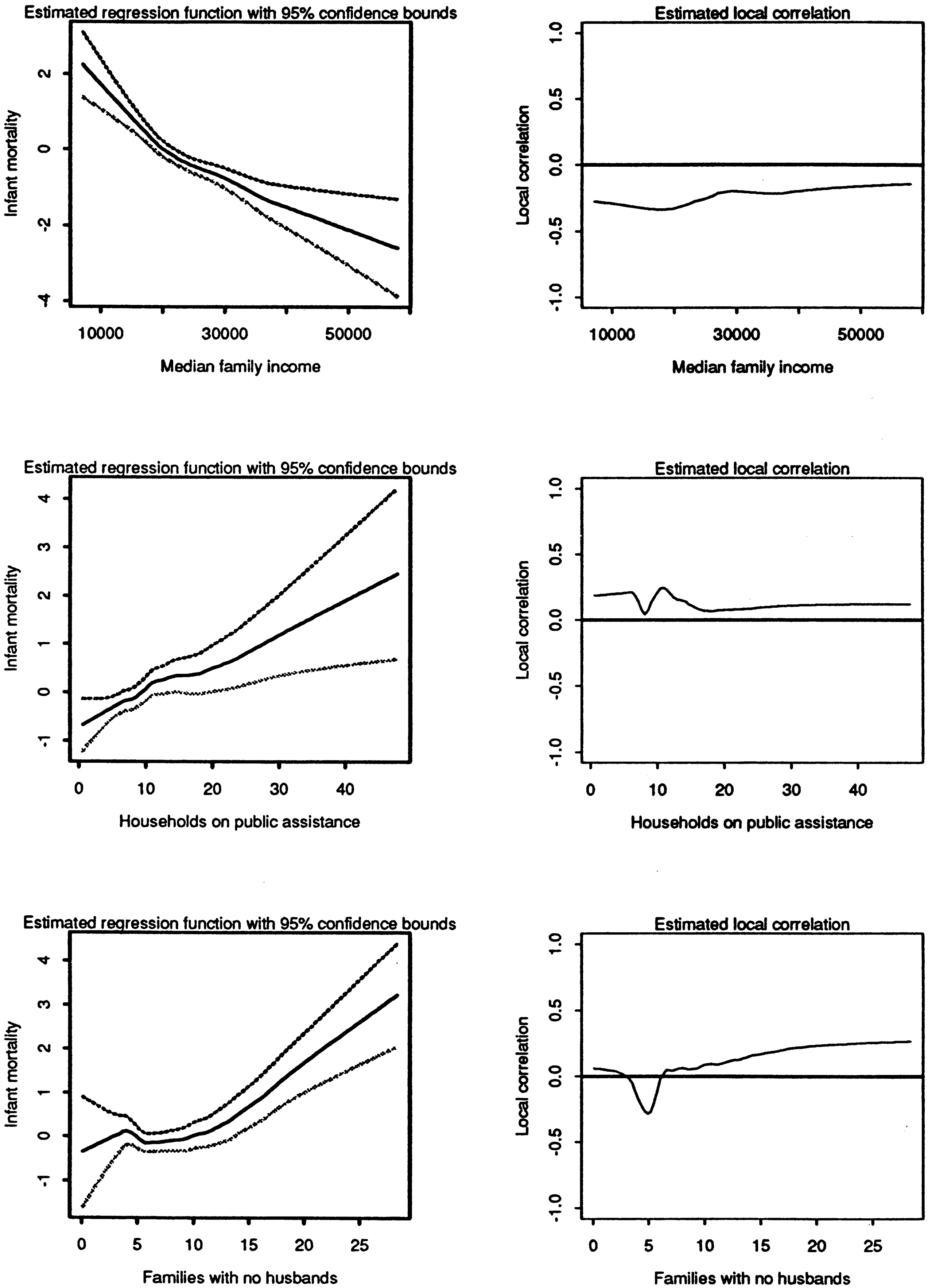

Figure 9: Local regression and correlation
Rural areas(Univariate model)

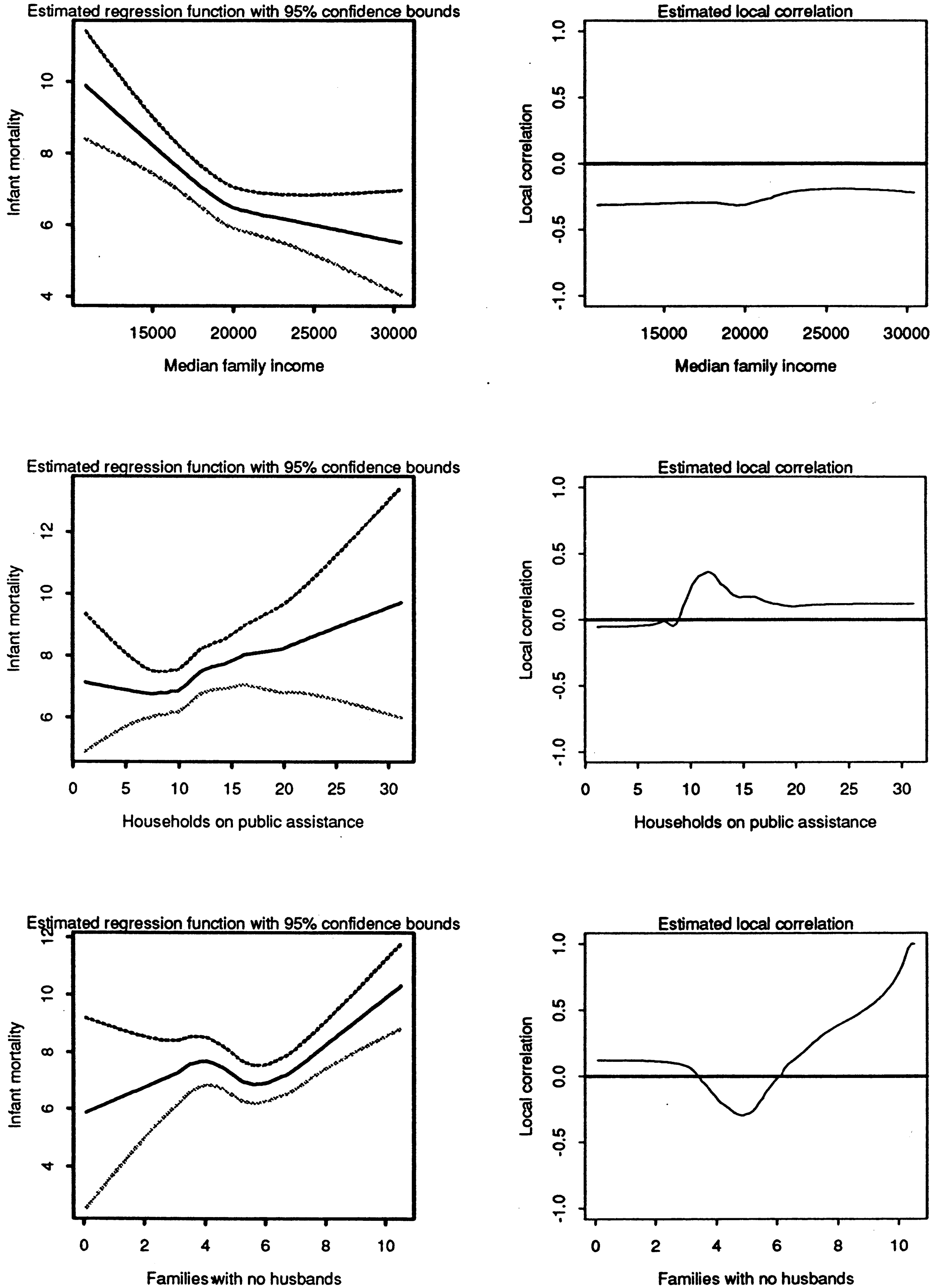

Figure 10: Partial local regression and correlation
Rural areas(Additive model)

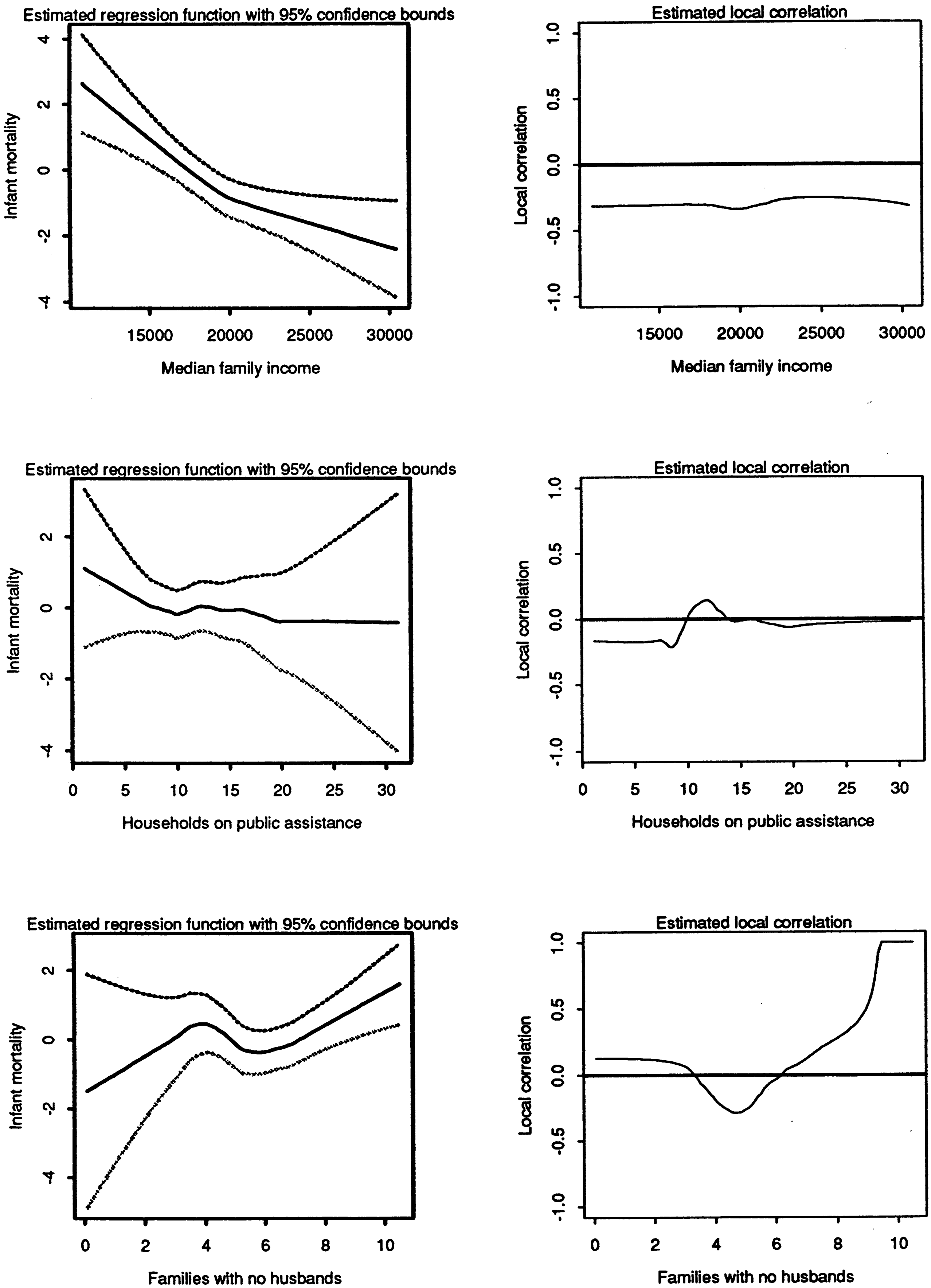

Figure 11: Local regression and correlation
Urban areas(Univariate model)

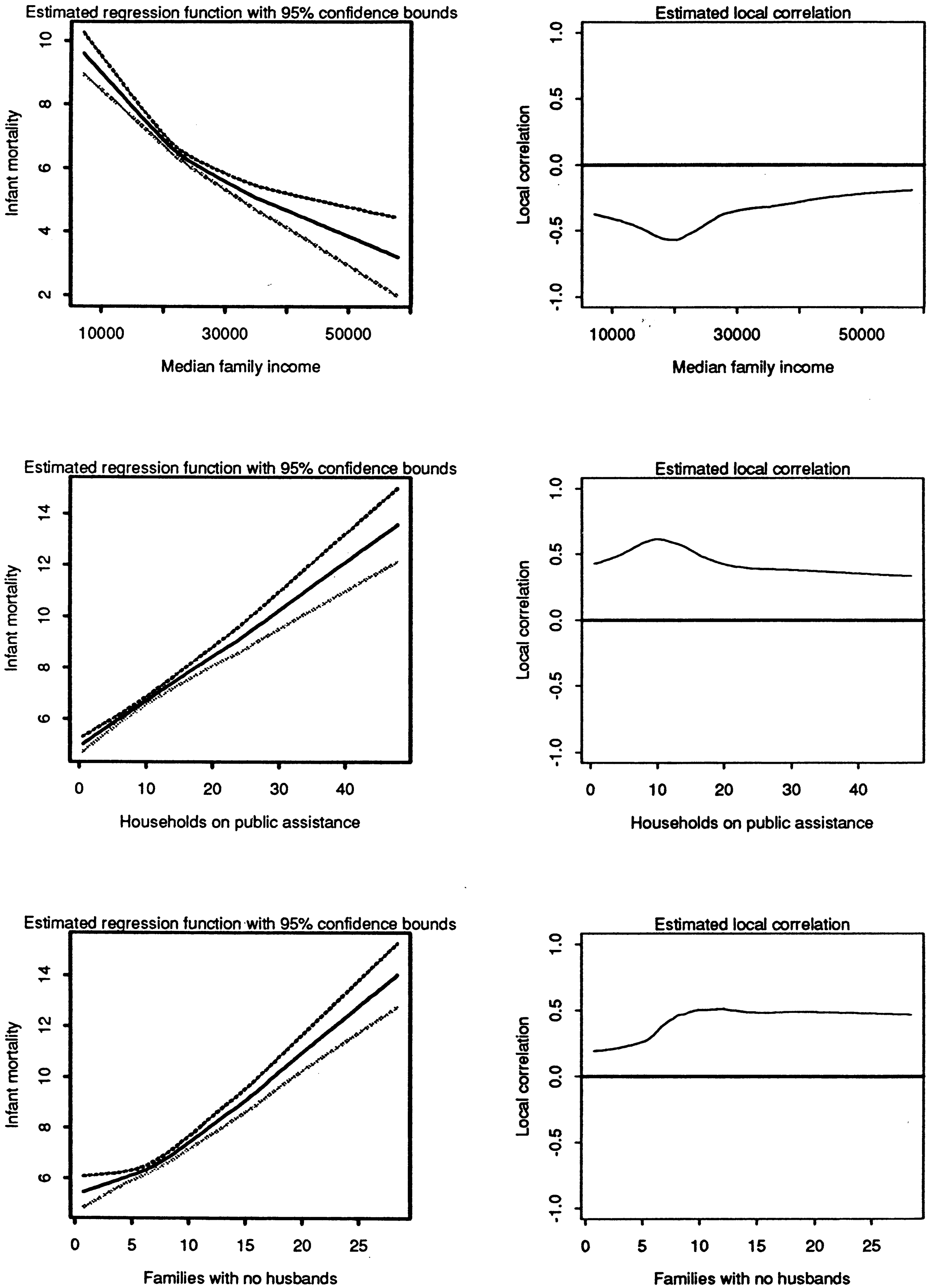

Figure 12: Partial local regression and correlation
Urban areas(Additive model)