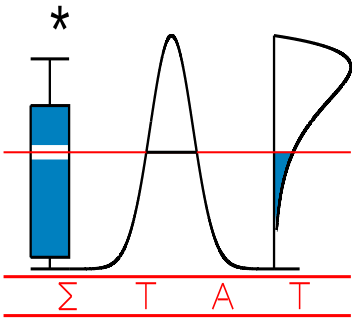
T E C H N I C A L

R E P O R T

11027

Quality of fit measures in the framework of quantile

NOH, H., EL GHOUCH, A. and I. VAN KEILEGOM



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Quality of Fit Measures in the Framework of Quantile Regression

Hohsuk Noh
Université catholique de Louvain *

Anouar El Ghouch
Université catholique de Louvain †

Ingrid Van Keilegom
Université catholique de Louvain ‡

May 6, 2011

Abstract

In regression experiments, to learn about the strength of the relationship between a covariate vector and a dependent variable, we propose a "coefficient of determination" based on the quantiles. Such a coefficient is a "local" measure in the sense that the strength is measured at a pre-specified quantile level. Once estimated, it can be used, for example, to measure the relative importance of a subset of covariates in the quantile regression context. Related to this coefficient, we also propose a new "local" lack-of-fit measure of a given parametric model. We provide some asymptotic results of the proposed measures and carry out a Monte Carlo simulation study to illustrate their use and performance in practice.

Key Words: Coefficient of determination; Inadequacy index; Lack-of-fit measure; Local polynomial estimation; Quantile regression; Nonparametric regression.

*H. Noh acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy).

†A. El Ghouch acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy).

‡I. Van Keilegom acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

1 Introduction

Given a random variable Y, the response, and a random vector $\underline{X} = (X_1, \ldots, X_d)^\top$ with $d \geq 1$, the covariates, the amount of variability in Y that can be explained by X, via an (unknown) link function, is a fundamental question in statistics. Typically, the link is given as the conditional mean function $m(\underline{X}) = \mathrm{E}(Y|\underline{X})$ but it may be any function of interest like the conditional median or, more generally, a conditional quantile function. When the mean regression function m is of interest, Doksum and Samarov (1995) proposed to use

$$\eta^2 = \frac{\mathrm{Var}(m(\underline{X}))}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{E}|Y - m(\underline{X})|^2}{\mathrm{E}|Y - \mathrm{E}(Y)|^2}$$

as a measure of the explanatory power of $\underline{X}$. This is nothing but a nonparametric version of the well known coefficient of determination R^2, which is largely used within the linear regression framework as an assess of the prediction power of a given linear model. While η^2 can address the question "is $\underline{X}$ (or a subset of it) important to understand the variation in Y?", it cannot answer some important questions like "does $\underline{X}$ exert any significant effect on the tail levels of Y?" or "is the effect of $\underline{X}$ as important on the tail levels of Y as on its central level?". To study such a question one should renounce the mean and adopt a function that gives a more comprehensive picture of the effect of $\underline{X}$ on Y like the conditional quantile. As we will show, by using the latter we are able to figure out the exact effect of $\underline{X}$ at a specific quantile level q of Y. To be clear, let us first consider the case of the median, i.e. $q = 1/2$. An L_1 natural analogue of η^2 is given by $1 - \mathrm{E}|Y - m_{0.5}(\underline{X})|/\mathrm{E}|Y - \xi_{0.5}|$, where $m_{0.5}$ is the conditional median of Y given $\underline{X}$ and $\xi_{0.5}$ is the marginal median of Y. An obvious generalization of such a measure is given by

$$R^1(q) = 1 - \frac{\mathrm{E}\left[\rho(Y - m(\underline{X}))\right]}{\mathrm{E}\left[\rho(Y - \xi_q)\right]},$$

where, for a fixed $0 < q < 1$, $m(\underline{X}) \equiv m_q(\underline{X}) = \arg\min_\theta \mathrm{E}[\rho_q(Y - \theta)|\underline{X}]$ is the (unique) conditional qth quantile of Y given $\underline{X}$, $\rho \equiv \rho_q(y) = (2q - 1)y + |y|$ is the well known check function and $\xi_q = \arg\min_\theta \mathrm{E}[\rho_q(Y - \theta)]$ is the (unique) qth marginal quantile of Y. Unlike η^2 which only assesses the mean power of $\underline{X}$, $R^1(q)$ enables us to get a more complete picture of the strength of $\underline{X}$ by varying q. A nice practical illustration of the importance of considering different values of q when studying the association between $\underline{X}$ and Y can be found in Koenker and Machado (1999), who cited an empirical study from Chamberlain (1994) on the U.S. labor union wage premium. Surprisingly, the $R^1(q)$ has never been studied in the literature neither from a theoretical nor from a practical viewpoint. The only exception that we have found is the work of Mckean and Sievers (1987), who proposed $R^1(1/2)$

as a robust alternative to R^2 in the case of a parametric linear model. Assuming a fixed design, they suggested a parametric version of the estimator of $R^1(1/2)$ and studied its consistency. This motivates us to consider $R^1(q)$ in a more general context, where (1) no parametric restriction is made on the form of m, (2) $\underline{X}$ is allowed to be a random vector of any given dimension $d \geq 1$ and (3) the quantile level q can be any value within $(0, 1)$. Under relatively weak assumptions, we propose a consistent estimator of $R^1(q)$, for which we obtain a Bahadur representation that allows us to prove its asymptotic normality. Note that by allowing m to be a constant it can be seen that $0 \leq R^1(q) \leq 1$. $R^1(q) = 0$ corresponds to the case when $m(\underline{X}) = \xi_q$ with probability one, i.e. no variability is captured by $\underline{X}$, while $R^1(q) = 1$ corresponds to the case when $m(\underline{X}) = Y$ with probability one, i.e. all the variability in Y at its qth quantile level is captured by $\underline{X}$ via the quantile function m. In other words, $R^1(q)$ exactly quantifies the relative gain of introducing the covariate $\underline{X}$ in estimating the qth quantile of Y. It can be used to robustly reduce the dimensionality of $\underline{X}$ by keeping only the significant components of it or to discriminate between different combinations of fixed number of covariates with the objective of selecting the best combination with the highest explanatory power.

Another related and important problem of common interest is to assess the discrepancy of a given parametric model, say $m(\theta, \underline{X})$, for estimating a certain conditional quantile of Y. The objective is to measure the inevitable loss of information that can be attributed to (and only to) the parametric restriction. From similar motivations as for $R^1(q)$, we can define

$$\Phi(q) = 1 - \frac{\mathrm{E}[\rho(Y - m(\underline{X}))]}{\mathrm{E}[\rho(Y - m(\theta^*, \underline{X}))]},$$

where $\theta^* \equiv \theta^*(q)$ is a "pseudo-true" parameter, i.e. an argument that minimizes $\mathrm{E}[\rho(Y - m(\theta, \underline{X}))]$ with respect to all θ in a set Θ, or equivalently $m(\theta^*, \cdot)$ is the best approximation to the true regression function m that can be found within the parametric family $\{m(\theta, \cdot)\}_{\theta \in \Theta}$. The numerator of $\Phi(q)$, i.e. $\mathrm{E}[\rho(Y - m(\theta^*, \underline{X}))] - \mathrm{E}[\rho(Y - m(\underline{X}))]$, is a nonnegative scalar that represents the amount of unexplained "variation", as measured by the "distance" ρ, due to the fact that one uses the parametric model $m(\theta^*, \cdot)$ instead of the true quantile regression curve m. Thus, it follows that $\Phi(q)$ is the fraction of the parametric residual variation that can be completely attributed to the lack of fit of the parametric quantile function $m(\theta, \cdot)$. In the following, $\Phi(q)$ will be shortly called the q-inadequacy index of $m(\theta, \cdot)$. Like $R^1(q)$, the q-inadequacy index is a local measure. By local we mean that it quantifies the quality (or to be more correct the poorness) of $m(\theta, \cdot)$ for a given fixed quantile level q. Allowing $m(\theta, \cdot)$ to be constant, one can easily check that $0 \leq \Phi(q) \leq R^1(q) \leq 1$. Unlike $R^1(q)$, the case $\Phi(q) = 0$ corresponds now to the best case in which $m(\theta^*, \underline{X})$ coincides almost surely with the true q-quantile

curve $m(\underline{X})$. The case $\Phi(q) = R^1(q)$ represents the worst case when the parametric model fails to capture any variation in the data that could have been captured if $m(\underline{X})$ was used. In such case $m(\theta^*, \underline{X}) = \xi_q$, with probability one. In the mean regression framework, this type of inadequacy index was already proposed in El Ghouch et al. (2010). They derived the Bahadur-type representation of their inadequacy index, denoted ζ^2, under some weak assumptions. Unlike $\Phi(q)$, ζ^2 is a global inadequacy measure over the entire conditional distribution of Y. Our q-inadequacy index $\Phi(q)$ may be used as a robust decision rule to find the best approximation among several candidate parametric models when a fixed set of covariates is given. Furthermore, our index has the advantage of finding an appropriate model corresponding to a given quantile level q (different values of q might correspond to different models). We select the model with the smallest value of $\Phi(q)$ as an appropriate model.

The paper is organized as follows. First, we propose estimators for $R^1(q)$ and $\Phi(q)$ in Section 2. The asymptotic properties of the proposed estimators are established in Section 3. Regarding $\Phi(q)$, due to some technical limitations, the asymptotic properties are only proved when the parametric model $m(\theta, \cdot)$ is linear. In Section 4, we will illustrate how our measures can be applied using some examples and simulations. Some general conclusions are given in Section 5, while the proofs and assumptions of the asymptotic results are deferred to Section 6.

2 Estimation

We assume that $\{(Y_i, \underline{X}_i)\},\ 1 \leq i \leq n$ are independent and identically distributed, with $\underline{X}_i = (X_{1i}, \ldots, X_{di})^\top$ being a random vector of $d \geq 1$ elements and Y_i being a random variable. The true qth quantile function $m(\underline{x})$ is assumed to be differentiable up to order $p+1$. This allows us to use the multivariate pth order local polynomial approximation

$$m(\underline{z}) \approx \sum_{0 \leq |\underline{r}| \leq p} \frac{1}{\underline{r}!} D^{\underline{r}} m(\underline{x}) (\underline{z} - \underline{x})^{\underline{r}},$$

where for any $\underline{r} = (r_1, \ldots, r_d),\ |\underline{r}| = \sum_{i=1}^d r_i,\ \underline{r}! = r_1! \times \cdots \times r_d!$,

$$D^{\underline{r}} m(\underline{x}) = \frac{\partial^{|\underline{r}|} m(\underline{x})}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}}, \underline{x}^{\underline{r}} = x_1^{r_1} \times \cdots \times x_d^{r_d}, \text{ and } \sum_{0 \leq |\underline{r}| \leq p} = \sum_{j=0}^{p} \sum_{r_1=0}^{j} \cdots \sum_{\substack{r_d=0 \\ |\underline{r}|=j}}^{j}.$$

2.1 Estimation of $R^1(q)$ and $\Phi(q)$

To estimate $R^1(q)$ without making any parametric restriction on m we need a nonparametric estimator. For its many well known useful properties, we adopt here the local polynomial approach. Let $K(\underline{u})$ be a density function on $\mathbb{R}^d$ and $h \equiv h_n$ a bandwidth parameter. Let $\hat{\beta}_{\underline{r}}, 0 \leq |\underline{r}| \leq p$ be the minimizer, with respect to $\beta_{\underline{r}}$ of

$$\sum_{i=1}^{n} K_h(\underline{X}_i - \underline{x}) \rho \left(Y_i; \sum_{0 \leq |\underline{r}| \leq p} \beta_{\underline{r}} (\underline{X}_i - \underline{x})^{\underline{r}} \right), \tag{2.1}$$

where $K_h(\underline{u}) = K(\underline{u}/h)$. The pth local polynomial estimator of $m(\underline{x})$ is $\hat{m}(\underline{x}) := \hat{\beta}_0(\underline{x})$. Based on this, we propose

$$\widehat{R^1(q)} \equiv \widehat{R^1_w(q)} = 1 - \frac{\sum_{i=1}^n \rho(Y_i - \hat{m}(\underline{X}_i)) w(\underline{X}_i)}{\sum_{i=1}^n \rho(Y_i - \hat{\xi}_q) w(\underline{X}_i)}, \tag{2.2}$$

as a nonparametric estimator of

$$R^1(q) \equiv R^1_w(q) = 1 - \frac{\mathrm{E}\left[\rho(Y - m(\underline{X})) w(\underline{X})\right]}{\mathrm{E}\left[\rho(Y - \xi_q) w(\underline{X})\right]}.$$

In the expressions above we incorporate the nonnegative weight function $w(\cdot)$ which is a usual strategy in kernel smoothing that aims to avoid highly uncertain estimation in regions with sparse or noisy data. Here, $\hat{\xi}_q = \arg\min_\theta n^{-1} \sum_{i=1}^n \rho(Y_i - \theta)$ is the qth sample quantile of Y.

Now, we turn to the q-inadequacy index. First, we define a population version of the q-inadequacy index incorporated with the weight function as follows:

$$\Phi(q) \;\equiv\; \Phi_w(q) \;=\; 1 - \;\frac{\mathrm{E}[\rho(Y - m(\underline{X})) w(\underline{X})]}{\mathrm{E}[\rho(Y - m(\theta^*, \underline{X})) w(\underline{X})]},$$

where $\theta^* \equiv \theta^*_w = \arg\min_{\theta \in \Theta} \mathrm{E}[\rho(Y - m(\theta, \underline{X})) w(\underline{X})]$. Note that in contrast to the Introduction, we are using a pseudo true parameter θ^*_w, which depends on the weight function. Even though it is not as natural as the one in the Introduction, we prefer this definition for technical reasons (related to the asymptotic theory of the estimator of $\Phi(q)$). To estimate $\Phi(q)$, in addition to $\hat{m}$, we need an estimator for $\theta^* \equiv \theta^*_w$, which can be defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \;\; \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i - m(\theta, \underline{X}_i)) w(\underline{X}_i),$$

where the parameter space Θ is a compact subset of $\mathbb{R}^l$. We propose to estimate $\Phi(q)$ by

$$\widehat{\Phi(q)} \equiv \widehat{\Phi_w(q)} = 1 - \frac{\sum_{i=1}^n \rho(Y_i - \hat{m}(\underline{X}_i)) w(\underline{X}_i)}{\sum_{i=1}^n \rho(Y_i - m(\hat{\theta}, \underline{X}_i)) w(\underline{X}_i)}. \tag{2.3}$$

Note that we make again use of the weight function w. This is not necessary for the parametric part of $\Phi(q)$, i.e. the denominator, but should be used in the numerator.

2.2 Bandwidth selection

A lot of research has been carried out to address the problem of data-driven bandwidth selection for nonparametric curve estimation, but much less is done related to bandwidth selection for integral functions. Intuitively, we can consider that an optimal bandwidth to estimate $R^1(q)$ is given by the minimizer of $\mathrm{E}\left[d(\widehat{R^1(q)}, R^1(q))\right]$. Here, $d(x, y)$ is a given distance like the L_1 or the L_2 distance. Even though we can easily obtain an asymptotic approximation of this quantity from Theorem 3.4 below, such an expression depends on the unknown quantities ξ_q, m and $R^1(q)$, etc. and consequently the bandwidth choice based on it turns out to be practically infeasible, because it needs many pilot estimates based on initial bandwidths or on approximate parametric models. Instead, we simply utilize the cross-validation bandwidth h for curve estimation. Doksum and Samarov (1995) and Tong and Yao (2000) also considered this cross-validation as one of the practicable options for the bandwidth choice of the estimation of their proposed integral functionals when they faced the same difficulties as ours. Following their approach, we consider the bandwidth h which minimizes the criterion

$$CV(h) = \sum_{i=1}^{n} \rho(Y_i - \hat{m}_{-i}(\underline{X}_i; h))w(\underline{X}_i)$$

where $\hat{m}_{-i}(\cdot; h)$ is the leave-one-out estimator for the conditional quantile function $m(\cdot)$, evaluated with the bandwidth h. Note that different from their criterion, which considered the case of the mean regression, the square function is replaced by the check function for quantile estimation. Furthermore, we incorporate the weight function into our cross-validation criterion.

3 Asymptotic Properties

To derive asymptotic properties of $\widehat{R^1(q)}$ and $\widehat{\Phi(q)}$, we need two building blocks. One is the Bahadur representation of $\hat{\xi}_q$ and $\hat{m}(\cdot)$. The other is the asymptotic normality of the estimator $\hat{\theta}$ of the parametric quantile regression model $m(\theta, \underline{X})$ under misspecification. These building blocks enable us to derive the Bahadur representation of $\widehat{R^1(q)}$ and $\widehat{\Phi(q)}$, which implies consistency and asymptotic normality. For convenience, we relegate all the assumptions for the asymptotic properties to Section 6.

3.1 Notations

In order to state the uniform Bahadur representation for local polynomial estimates of the conditional quantile function $m(\underline{X})$, we need to develop some notations. Let $N_i = \binom{i+d-1}{d-1}$ be the number of distinct d-tuples $\underline{r}$ with $|\underline{r}| = i$. Arrange these N_i d-tuples as a sequence in lexicographical order (with the highest priority to the last position so that $(0, \ldots, 0, i)$ is the first element in the sequence and $(i, 0, \ldots, 0)$ the last element). Let τ_i denote the 1-to-1 mapping, defined by $\tau_i(1) = (0, \ldots, 0, i), \ldots, \tau_i(N_i) = (i, 0, \ldots, 0)$. For each $i = 1, \ldots, p$, define a $N_i \times 1$ vector $\mu_i(\underline{x})$ with its kth element given by $\underline{x}^{\tau_i(k)}$ and write $\mu(\underline{x}) = (1, \mu_1(\underline{x})^\top, \ldots, \mu_p(\underline{x})^\top)^\top$, which is a column vector of length $N = \sum_{i=0}^{p} N_i$. Similarly define vectors $\beta_p(\underline{x})$ and $\underline{\beta}$ through the same lexicographical arrangement of $D^{\underline{r}} m(\underline{x})$ and $\beta_{\underline{r}}$ in (2.1) for $0 \le |\underline{r}| \le p$. Then (2.1) can be rewritten as

$$\sum_{i=1}^{n} K_h(\underline{X}_i - \underline{x}) \rho(Y_i - \mu(\underline{X}_i - \underline{x})^\top \beta). \tag{3.1}$$

Suppose the minimizer of (3.1) is denoted as $\hat{\beta}_n(\underline{x})$. Let $\hat{\beta}_p(\underline{x}) = W_p \hat{\beta}_n(\underline{x})$, where W_p is a diagonal matrix with diagonal entries equal to the lexicographical arrangement of $\underline{r}!$, $0 \le |\underline{r}| \le p$.

Let $\varphi(t) = 2qI\{t \ge 0\} + (2q-2)I\{t < 0\}$ be the piecewise derivative of the check function $\rho(t)$. Define $G(t, \underline{x}) = \mathrm{E}\{\varphi(Y - t) | \underline{X} = \underline{x}\}$ and $g(\underline{x}) = (\partial/\partial t) G(t, \underline{x})$. Then, $g(\underline{x}) = -2 f_{\varepsilon|\underline{X}}(0|\underline{x})$, where $f_{\varepsilon|\underline{X}}(\cdot|\underline{x})$ is the conditional density of $\varepsilon = Y - m(\underline{X})$ given $\underline{X} = \underline{x}$. Let $\nu_{\underline{i}} = \int K(\underline{u}) \underline{u}^{\underline{i}} d\underline{u}$. For $f(\cdot)$, the probability density function of $\underline{X}$, define $\nu_{n\underline{i}}(\underline{x}) = \int K(\underline{u}) \underline{u}^{\underline{i}} g(\underline{x} + h\underline{u}) f(\underline{x} + h\underline{u}) d\underline{u}$. For $0 \le j, k \le p$, let $S_{j,k}$ and $S_{n,j,k}(\underline{x})$ be two $N_j \times N_k$ matrices with their (l, m) elements respectively given by

$$[S_{j,k}]_{l,m} = \nu_{\tau_j(l) + \tau_k(m)}, \quad [S_{n,j,k}(\underline{x})]_{l,m} = \nu_{n, \tau_j(l) + \tau_k(m)}(\underline{x}).$$

From this, we can define the $N \times N$ matrices S_p and $S_{n,p}(\underline{x})$ by

$$S_p = \begin{pmatrix} S_{0,0} & S_{0,1} & \cdots & S_{0,p} \\ S_{1,0} & S_{1,1} & \cdots & S_{1,p} \\ \vdots & \ddots & \vdots & \\ S_{p,0} & S_{p,1} & \cdots & S_{p,p} \end{pmatrix}, \quad S_{n,p}(\underline{x}) = \begin{pmatrix} S_{n,0,0}(\underline{x}) & S_{n,0,1}(\underline{x}) & \cdots & S_{n,0,p}(\underline{x}) \\ S_{n,1,0}(\underline{x}) & S_{n,1,1}(\underline{x}) & \cdots & S_{n,1,p}(\underline{x}) \\ \vdots & \ddots & \vdots & \\ S_{n,p,0}(\underline{x}) & S_{n,p,1}(\underline{x}) & \cdots & S_{n,p,p}(\underline{x}) \end{pmatrix}.$$

For $|S_p| \ne 0$, define

$$\beta_n^*(\underline{x}) = -\frac{1}{nh^d} W_p S_{n,p}^{-1}(\underline{x}) H_n^{-1} \sum_{i=1}^{n} K_h(\underline{X}_i - \underline{x}) \varphi(Y_i, \mu(\underline{X}_i - \underline{x})^\top W_p^{-1} \beta_p(\underline{x})) \mu(\underline{X}_i - \underline{x}),$$

where H_n is a diagonal matrix with diagonal entries $h^{|\underline{r}|}$, $0 \leq |\underline{r}| \leq p$ in the aforementioned lexicographical order. Note that $\beta_n^*(\underline{x})$ is the leading term in the Bahadur representation of $\hat{\beta}_p(\underline{x}) - \beta_p(\underline{x})$ as it can be found in Kong et al. (2010). We would like to point out that the definition of $\beta_n^*(\underline{x})$ in Kong et al. (2010) is not correct and $\mu(\underline{X}_i - \underline{x})^\top \beta_p(\underline{x})$ should in fact be replaced by $\mu(\underline{X}_i - \underline{x})^\top W_p^{-1} \beta_p(\underline{x})$ in their definition.

3.2 Bahadur representation of the functions involving $\hat{\xi}_q$ and $\hat{m}(\cdot)$

Lemma 3.1 *Suppose that (A4),(A5) and (A9) hold. Then,*

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i - \hat{\xi}_q) w(\underline{X}_i) - \mathrm{E}[\rho(Y - \xi_q) w(\underline{X})] \\ = \; & \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i - \xi_q) w(\underline{X}_i) - \mathrm{E}[\rho(Y - \xi_q) w(\underline{X})] + o_p(n^{-1/2}). \end{aligned}$$

The proof is similar to the proof of Lemma 3.3 below (using the Bahadur representation for $\hat{\xi}_q$ instead of the one for $\hat{m}(\cdot)$), and is therefore omitted. To derive the Bahadur representation of $\widehat{R^1(q)}$, we need to investigate the asymptotic behavior of $\sum_{i=1}^n \rho(Y_i - \hat{m}(\underline{X}_i)) w(\underline{X}_i)$. To this purpose, we need the Bahadur representation of $\hat{m}(\cdot)$ given in Lemma 3.2.

Lemma 3.2 *Let $\underline{e}_1$ be a $N \times 1$ vector with its first element given by 1 and all others 0. Suppose (A1)-(A9) hold and $h \asymp n^{-\kappa}$ with $\kappa > 1/(2p+2+d)$. Then with probability one we have,*

$$\hat{m}(\underline{x}) - m(\underline{x}) = -\underline{e}_1^\top \frac{H_n^{-1}}{nh^d} S_{np}^{-1}(\underline{x}) \sum_{i=1}^{n} K_h(\underline{X}_i - \underline{x}) \varphi(\varepsilon_i) \mu(\underline{X}_i - \underline{x}) + R_n,$$

where $R_n = o_p((nh^d)^{-1/2})$ uniformly in $\underline{x} \in \mathcal{D}$ and $\mathcal{D}$ is the compact support of the weight function $w(\cdot)$.

Lemma 3.2 can be proved by using Corollary 3.3, 5.8 and 5.10 in Kong et al. (2010) and the fact that $\sup_{\underline{x} \in \mathcal{D}} |\mathrm{E}\beta_n^*(\underline{x})| = O(h^{p+1})$ almost surely, which is obtained by the similar arguments as Propostion 4 in Masry (1996). The condition $\kappa > 1/(2p+2+d)$ implies that h decreases to zero faster than the optimal order for curve estimation. In order to obtain $\sqrt{n}$-convergence of $\widehat{R^1(q)}$, undersmoothing is used, which is common when estimating some unknown parameter resulting from a sample average of some nonparametric estimator. A similiar approach can be found in Van Keilegom and Wang (2010) and Doksum and Samarov (1995).

Lemma 3.3 *Suppose that (A1)-(A9) hold,* $p > d/2-1$ *and* $h \asymp n^{-\kappa}$ *with* $1/(2p+2+d) < \kappa < 1/(2d)$. *Then, we have*

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\rho(Y_i - \hat{m}(\underline{X}_i))w(\underline{X}_i) - \mathrm{E}[\rho(Y - m(\underline{X}))w(\underline{X})] \\ = \; & \frac{1}{n}\sum_{i=1}^{n}\rho(Y_i - m(\underline{X}_i))w(\underline{X}_i) - \mathrm{E}[\rho(Y - m(\underline{X}))w(\underline{X})] + o_p(n^{-1/2}). \end{aligned}$$

The proof of Lemma 3.3 is given in Section 6.

3.3 Bahadur representation of $\widehat{R^1(q)}$

We can now establish the following Bahadur representation of $\widehat{R^1(q)}$ from the expansions in Lemma 3.1 and Lemma 3.3 and the fact that $\hat{a}/\hat{b} = a/b + \hat{b}^{-1}\left[\hat{a} - a - (\hat{b} - b)(a/b)\right]$.

Theorem 3.4 *Suppose that (A1)-(A9) hold,* $p > d/2-1$ *and* $h \asymp n^{-\kappa}$ *with* $1/(2p+2+d) < \kappa < 1/(2d)$. *Let* $R^1(q) = 1 - \mathrm{E}[\rho(Y - m(\underline{X}))w(\underline{X})]/\mathrm{E}[\rho(Y - \xi_q)w(\underline{X})]$. *Then, we have*

$$\sqrt{n}(\widehat{R^1(q)} - R^1(q)) \;\; = \;\; \frac{1}{\sqrt{n}}(1 - R^1(q))\sum_{i=1}^{n}(e_i - u_i) + o_p(1),$$

where

$$e_i = \frac{\rho(Y_i - \xi_q)w(\underline{X}_i) - \mathrm{E}[\rho(Y - \xi_q)w(\underline{X})]}{\mathrm{E}[\rho(Y - \xi_q)w(\underline{X})]} \text{ and } u_i = \frac{\rho(Y_i - m(\underline{X}_i))w(\underline{X}_i) - \mathrm{E}[\rho(Y - m(\underline{X}))w(\underline{X})]}{\mathrm{E}[\rho(Y - m(\underline{X}))w(\underline{X})]}.$$

Theorem 3.4 implies that $n^{1/2}(\widehat{R_1(q)} - R_1(q))$ is asymptotically normal with mean zero and variance $\sigma^2 \equiv (1 - R_1(q))^2\mathrm{Var}(e_1 - u_1)$. When $R_1(q) = 0$ or 1, this variance is zero and our result implies degenerate normality. We should investigate in that case the next term in the expansion to obtain a meaningful distributional convergence result. Note that to guarantee the optimal root-n rate of convergence for $\widehat{R^1(q)}$, the condition $p > d/2 - 1$ is needed. It implies that the order of the local approximation should increase as the dimension of $\underline{X}$ increases. Similar restrictions in nonparametric estimation can be found in e.g. Linton (1995), Powell and Stoker (1996) and El Ghouch et al. (2010).

3.4 Bahadur representation of $\widehat{\Phi(q)}$

For some technical reasons, we will restrict our attention to linear models as a candidate parametric family for Φ. In fact, Lemma 3.5 below has only been proved in the literature for linear models, and its extension to nonlinear models is beyond the scope of this paper. First, we will state the

asymptotic normality result of the parameter estimator of the linear quantile regression model under misspecification, which is a little bit modified to accommodate the weight function $w(\cdot)$ in the estimation.

Lemma 3.5 *(Modification of Theorem 1 in Kim and White (2003)) Suppose that (A4) and (B1)-(B4) hold. θ^* is a "pseudo-true" parameter from Assumption (B1), $\underline{X}_e = (1, \underline{X}^\top)^\top$ and $f_{\varepsilon^*|\underline{X}_e}(\cdot|\underline{x}_e)$ is the conditional density of $\varepsilon^* = Y - \theta^{*\top}\underline{X}_e$ given $\underline{X}_e = \underline{x}_e$. Let $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1^\top)$ be defined as*

$$\hat{\theta} = \underset{\theta=(\theta_0,\theta_1^\top)\in\Theta}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i - \theta_0 - \theta_1^\top \underline{X}_i) w(\underline{X}_i).$$

Then,

$$\sqrt{n}(\hat{\theta} - \theta^*) = \sqrt{n}((\hat{\theta}_0, \hat{\theta}_1^\top) - (\theta_0^*, \theta_1^{*\top})) \xrightarrow{d} N(\underline{0}, C),$$

where $C = Q^{-1}VQ^{-1}$ and

$$Q = \mathrm{E}\left[2 f_{\varepsilon^*|\underline{X}_e}(0|\underline{X}_e)\underline{X}_e\underline{X}_e^\top w(\underline{X}_e)\right], \quad V = \mathrm{E}\left[\varphi(\varepsilon^*)^2 \underline{X}_e\underline{X}_e^\top w(\underline{X}_e)^2\right].$$

The proof of Lemma 3.5 is nothing but a slight modification of that of Theorem 1 in Kim and White (2003), and so it is omitted.

Theorem 3.6 *Suppose that (A1)-(A9) and (B1)-(B4) hold, $p > d/2 - 1$ and $h \asymp n^{-\kappa}$ with $1/(2p + 2 + d) < \kappa < 1/(2d)$. Let $\Phi(q) = 1 - \mathrm{E}[\rho(Y - m(\underline{X}))w(\underline{X})]/\mathrm{E}[\rho(Y - \theta_0^* - \theta_1^{*\top}\underline{X})w(\underline{X})]$. Then, we have*

$$\begin{aligned}\sqrt{n}(\widehat{\Phi(q)} - \Phi(q)) &= \frac{1}{\sqrt{n}}(1 - \Phi(q)) \sum_{i=1}^{n} \left[\frac{\rho(Y_i - \theta_0^* - \theta_1^{*\top}\underline{X}_i)w(\underline{X}_i) - \mathrm{E}[\rho(Y - \theta_0^* - \theta_1^{*\top}\underline{X})w(\underline{X})]}{\mathrm{E}[\rho(Y - \theta_0^* - \theta_1^{*\top}\underline{X})w(\underline{X})]} \right. \\ &\qquad \left. - \frac{\rho(Y_i - m(\underline{X}_i))w(\underline{X}_i) - \mathrm{E}[\rho(Y - m(\underline{X}))w(\underline{X})]}{\mathrm{E}[\rho(Y - m(\underline{X}))w(\underline{X})]} \right] + o_p(1).\end{aligned}$$

The proof of Theorem 3.6 is given in Section 6. As in Theorem 3.4, Theorem 3.6 implies that $\widehat{\Phi(q)}$ converges to a zero mean normal distribution as $n \to \infty$.

4 Simulation Results

We illustrate our proposed estimators through four examples below. A Gaussian kernel has been used throughout. For a multivariate kernel, we simply use the product kernel and we let all components of each bandwidth vector be equal.

Example 1. Consider the model

$$Y_i = m_1(X_i) + \lambda\, m_2(X_i) + \tau\varepsilon_i = 6 + 2X_i + \lambda\, \sin(|3\pi X_i + \pi|) + \tau\varepsilon_i,$$

where $\{X_i\}$ and $\{\varepsilon_i\}$ are two independent random series, the X_i's are independent with common distribution $U[-\epsilon, 1+\epsilon]$ and the ε_i's are independent and standard normal. We set $w(x) = I(0 \le x \le 1)$ and ϵ was chosen so that $w(x)$ corresponds to 95% of the inner sample range of the data. We are interested in measuring the contribution of the covariate X and the discrepancy of the linear model $m(\theta, x) = \theta_0 + \theta_1 x$ for estimating a specific quantile of Y. For this purpose, we vary the values of λ and τ. Table 1 displays the values of τ and λ used to generate the data together with the corresponding values of $R^1(0.5)$ and $\Phi(0.5)$. As shown in Table 1, if we increase τ, then the influence of ε on the quantile of Y increases and X becomes more and more irrelevant ($R^1(q) \searrow 0$). On the other hand, when $\lambda = 0$, the linear model $m(\theta, x)$ is correct so $\Phi(q) = 0$, but as λ increases, $m(\theta, x)$ becomes more and more inadequate ($\Phi(q) \nearrow 1$). Table 2 shows biases and standard deviations of $\widehat{R^1(0.5)}$ and $\widehat{\Phi(0.5)}$, multiplied by 100, based on local linear estimators of $m(\cdot)$ in Monte Carlo trials with 100 replications and sample size $n = 100, 200$ and 400. From Table 2, we can see that biases and standard deviations decrease as n increases. We also have done the same simulations for different values of q and have calculated confidence intervals for $R^1(q)$ and $\Phi(q)$, but the results are not presented here for the sake of brevity. We observed that the increase of sample size improves bias and standard deviations for different values of q as when $q = 0.5$. To construct confidence intervals, we utilize the obtained asymptotic normality, see Theorem 3.4 and 3.6. We used a "naive" estimator of the asymptotic variance based on the plug-in principle (without modifying the bandwidth parameter). We observed that the empirical coverage probabilities approximate the nominal coverage confidence levels as n increases. But the observed coverage probabilities were not good especially when $d = 3$. One of the possible reasons for this phenomenon is that our bandwidth choice is not optimal for the estimation of $R^1(q)$ and $\Phi(q)$ and also for the asymptotic variances of them. Moreover, as was pointed out in Doksum and Samarov (1995), the asymptotic theory tends to underestimate the actual variability.

Example 2. (Relative importance of subsets of the candidate covariates) Consider the model

$$Y_i = 3X_{1i} + 0.3X_{2i} + 0.2\exp(X_{3i}) + (0.4 - 0.3X_{2i})\varepsilon_i,$$

where $\{\underline{X}_i = (X_{1i}, X_{2i}, X_{3i})\}$ and $\{\varepsilon_i\}$ are independent, and the ε_i's are independent and standard normal. The components of $\underline{X}_i$ are $U[-\epsilon, 1+\epsilon]$, independent of each other and ϵ is chosen so that

the weight function $w(\underline{x}) = \prod_{j=1}^{3} I(0 \le x_j \le 1)$ corresponds to 95% of the inner sample range of the data.

Table 3 shows $R^1(q)(q = 0.1, \ldots, 0.5)$ and η^2 values for all the subsets of X_1, X_2 and X_3. From the values of η^2, we can see that in mean regression both X_2 and X_3 are insignificant covariates with no difference between them. However, from the values of $R^1(q)$ for different q we can see that in quantile regression X_2, which is related not only to the drift term but also to the diffusion term of the model, has a larger effect on high conditional quantiles of Y than X_3. This fact becomes more clear when we compare the relative importance between the subsets (X_1, X_2) and (X_1, X_3). The comparison suggests that if we would like to reduce the number of covariates, we can choose either (X_1, X_2) or (X_1, X_3) in the median or mean regression case, while we should choose (X_1, X_2) instead of (X_1, X_3) when we want to estimate high conditional quantiles of Y. We simulated 100 random samples of size 200 to investigate how well the estimator $\widehat{R^1(q)}$ reflects this property of the true value $R^1(q)$. We have used local linear estimators. Table 4 shows medians, biases and standard deviations of $\widehat{R^1(0.5)}$ and $\widehat{R^1(0.2)}$ especially for the subsets which contain X_1. From the results in Table 4, we can see that the estimator can guide us to preferring X_2 over X_3 for the estimation of high conditional quantiles. Moreover, we observe that the number of cases where $\hat{R}^1(X_1, X_2)$ is larger than $\hat{R}^1(X_1, X_3)$ is 90 out of 100 when $q = 0.2$, while when $q = 0.5$ the number of such cases is only 37. The same lesson can be learned from Figure 1, which shows boxplot summaries of $\widehat{R^1(0.5)}$ and $\widehat{R^1(0.2)}$ of the subsets (X_1, X_2) and (X_1, X_3).

Example 3. (Discrepancy of linear models) Consider the model

$$Y_i = 3X_{1i} + \exp(X_{2i}) + (0.9 - 0.3\exp(X_{2i}))\varepsilon_i.$$

All other ingredients of this model are the same as for the model in Example 2, except that ϵ is chosen so that the weight function $w(\underline{x}) = \prod_{j=1}^{2} I(0 \le x_j \le 1)$ corresponds to 95% of the inner sample range of the data. We consider four surrogate parametric linear models for the given model, $Y_i = \theta_0 + \theta_1 X'_{1i} + \theta_2 X'_{2i} + \varepsilon'_i$, and a different pair (X'_1, X'_2) is given to each linear model according to Table 5. All surrogate linear models have a certain degree of discrepancy except Model $S1$. Table 6 shows true values, medians, biases and standard deviations of $\widehat{\Phi(0.5)}$ and $\widehat{\Phi(0.2)}$ for four surrogate linear models based on local linear estimators of $m(\cdot)$ in Monte Carlo trials with 100 replications and sample size $n = 200$. First, from the values of Φ given in Table 6, we can derive the discrepancy order among the four models: $S1 < S2 < S3 < S4$. Second, Table 6 and Figure 2 show that our proposed estimator $\widehat{\Phi(q)}$ can detect the discrepancy order based on the sample in a reasonable sense.

For example, out of 100 repetitions, our estimator can find the exact discrepancy order among the four models 93 times.

Example 4. (The necessity of the requirement $p > d/2 - 1$) To guarantee the optimal convergence rate of $\widehat{R^1(q)}$ and $\widehat{\Phi(q)}$, the condition $p > d/2 - 1$ is needed. We revisited the model in Example 2 to provide an illustration for the requirement $p > d/2 - 1$. Since $d = 3$ in the model, we should use a local linear smoother ($p = 1$) or a higher-order local smoother ($p \geq 2$). To illustrate the necessity of the requirement, we calculated the estimators $\hat{R}^1(X_1, X_2)$ and $\hat{R}^1(X_1, X_3)$ when $q = 0.2$, based on the local constant fit and the local linear fit of $m(\cdot)$ in Monte Carlo trials with 100 replications and sample size $n = 200$. Figures 3 and 4 show boxplot summaries of the selected subsets $[(X_1, X_2, X_3), (X_1, X_2), (X_1, X_3)]$. We can see that the estimators based on the local linear fit clearly outperform those based on the local constant fit especially in terms of bias. If we compare the scatter plot of $\hat{R}^1(X_1, X_2)$ and $\hat{R}^1(X_1, X_3)$ for those two estimators when $q = 0.2$ and $q = 0.5$ (Figure 5), we can also see that the estimators based on the local linear fit clearly show the preference of X_2 over X_3 for the estimation of high conditional quantiles of Y, while those based on the local constant fit do not. But note that the superiority of the estimator based on the local linear fit can only be observed if enough data are available. For applications with high dimensional $\underline{X}$-vectors, locally high order fits involve estimating many local parameters and require a reasonably large bandwidth in order to include sufficiently many data points. So under the deficiency of data points, the locally low order fit with a smaller bandwidth will provide competitive performance.

5 Conclusions and Remarks

In this work we have investigated two local indices: the qth explanatory power $R^1(q)$ of a given covariate and the qth inadequacy index $\Phi(q)$ of a given quantile parametric (linear) model. In the following we summarize some key points.

- Using the local polynomial approximation method we were able to construct nonparametric estimators that are $\sqrt{n}-$consistent and asymptotically normal. To get free of the curse of dimensionality we assumed that the true regression function m is sufficiently smooth in the sense that the order of the local approximation p needed in the Taylor approximation is larger then $d/2 - 1$, with d being the dimension of the covariate vector. However, from a practical point of view, we have observed that increasing d, inevitably deteriorates the performance of our

estimators for a fixed finite simple size. For this reason and in order to virtually increase the local sample size we used a Gaussian kernel (which has unbounded support) in all our simulations. This reduces significantly the mean squared error compared to the case when we use a bounded kernel, e.g. the Epanechnikov kernel; we don't show the results here for the sake of brevity.

- Like for any nonparametric estimator, the bandwidth parameter is crucial here and critically affects the rate of convergence. Globally, the theoretical asymptotic results match well the simulation results although the practical data-driven bandwidth that we used is surely not the optimal one. The problem of selecting a good bandwidth needs to be investigated more both theoretically and practically.

- To construct confidence intervals for $R^1(q)$ and $\Phi(q)$, we utilized the obtained asymptotic normality, see Theorem 3.4 and 3.6. To this end we used a "naive" estimator of the asymptotic variance based on the plug-in principle without modifying the bandwidth parameter. But the empirical coverage probabilities were not good especially when $d = 3$. One of the possible reasons for this phenomenon is that our bandwidth choice is not optimal for the estimation of $R^1(q)$ and $\Phi(q)$ and also for the asymptotic variances of them.

- To use the qth inadequacy index with a nonlinear model, the theory for a parametric estimation of a misspecified nonlinear quantile function under random design needs to be developed.

6 Assumptions and Proofs

6.1 Assumptions for the Bahadur representation of $\widehat{R^1(q)}$

Let V be an open convex set in $\mathbb{R}^d$.

(A1) All partial derivatives of $m(\underline{x})$ up to order $p+1$ exist and are continuous for all $\underline{x} \in V$, and there exists a constant $C > 0$ such that $|D^{\underline{r}} m(\underline{x})| \leq C$ for all $\underline{x} \in V$ and $|\underline{r}| = p + 1$.

(A2) The marginal density of $\varepsilon = Y - m(\underline{X})$ is bounded and $\mathrm{E}\{\varphi(\varepsilon)|\underline{X}\} = 0$.

(A3) For all e in a neighborhood of zero, the conditional density $f_{\varepsilon|\underline{X}}(e|\underline{x})$ of $\varepsilon = Y - m(\underline{X})$ given $\underline{X} = \underline{x}$ satisfies

$$|f_{\varepsilon|\underline{X}}(e|\underline{x}_1) - f_{\varepsilon|\underline{X}}(e|\underline{x}_2)| \leq K_e \|\underline{x}_1 - \underline{x}_2\|,$$

where K_e is a positive constant depending on e. Further, the conditional density is positive for $e = 0$ for all values of $\underline{x} \in V$, and its first partial derivative w.r.t e, $D^1 f_{\varepsilon|\underline{X}}(e|\underline{x})$ is bounded for all $\underline{x} \in V$ and e in a neighborhood of zero.

(A4) The weight function $w(\underline{x})$ is continuous, and its support $\mathcal{D} \subset V$ is compact and has nonempty interior.

(A5) $K(\cdot)$ has a compact support, say $[-1,1]^{\bigotimes d}$ and $|H_{\underline{j}}(\underline{u}) - H_{\underline{j}}(\underline{v})| \leq C\|u - v\|$ for all j with $0 \leq |\underline{j}| \leq 2p+1$, where $H_{\underline{j}}(\underline{u}) = \underline{u}^{\underline{j}} K(\underline{u})$.

(A6) The marginal density $f(\underline{x})$ of $\underline{X}$ is positive and bounded with bounded first order derivatives on V.

(A7) $nh^{d+2(p+1)}/\log n = O(1)$ as $h \to 0$.

(A8) The conditional density $f_{\underline{X}|Y}$ of $\underline{X}$ given Y exists and is bounded.

(A9) The distribution function of Y, $F_Y(\cdot)$ has bounded second derivative in a neighborhood of ξ_q and $f_Y(\xi_q) > 0$ where f_Y is the marginal density function of Y.

The assumptions given here are basically a simplified adaptation of those in Kong et al. (2010) to the case of *i.i.d* quantile regression except (A3), (A4) and (A9), which are assumed for the Bahadur representation of the functions involving $\hat{\xi}_q$ and $\hat{m}(\cdot)$ in Subsection 3.2. (A3) is similar to Condition 3 in Chaudhuri et al. (1997), which estimated a parameter using the local polynomial estimate of the conditional quantile function as in our paper. (A9) is for the Bahadur representation of the sample quantile $\hat{\xi}_q$. For more details, we refer to Kiefer (1967).

6.2 Assumptions for the Bahadur represenatation of $\widehat{\Phi(q)}$

The assumptions given below are the same as in Kim and White (2003), except the slight modifications for the weight function $w(\cdot)$.

(B1) (Orthogonality condition) There exists a vector $\theta_* = (\theta_0^*, \theta_1^{*\top})$ such that

$$\mathrm{E}[\underline{X}_e \varphi(Y - \theta^{*\top} \underline{X}_e) w(\underline{X}_e)] = 0,$$

and $\theta^* = (\theta_0^*, \theta_1^{*\top})$ is an interior point of Θ, where Θ is a compact set of $\mathbb{R}^{d+1}$ and $\underline{X}_e = (1, \underline{X}^\top)^\top$.

(B2) $\mathrm{E}[\|\underline{X}_e\|^3 w(\underline{X}_e)] < \infty$ and $\mathrm{E}[\|\underline{X}_e\|^2 w(\underline{X}_e)^2] < \infty$.

(B3) For all $\underline{x}_e \in \mathcal{D}$, $f_{\varepsilon^*|\underline{X}_e}(\cdot|\underline{x}_e)$ is positive at zero, bounded and Lipschitz continuous where $f_{\varepsilon^*|\underline{X}_e}(\cdot|\underline{x}_e)$ is the conditional density of $\varepsilon^* = Y - \theta^{*\top}\underline{X}_e$ given $\underline{X}_e = \underline{x}_e$. Moreover, the marginal density of ε_* is bounded.

(B4) $Q = \mathrm{E}\left[2f_{\varepsilon^*|\underline{X}_e}(0|\underline{X}_e)\underline{X}_e\underline{X}_e^\top w(\underline{X}_e)\right]$ and $V = \mathrm{E}\left[\varphi(\varepsilon^*)^2\underline{X}_e\underline{X}_e^\top w(\underline{X}_e)^2\right]$ are positive definite.

6.3 Proof of Lemma 3.3

Recall that for any x, y,

$$\rho(x-y) - \rho(x) = (-y)\varphi(x) + 2(y-x)[I(y > x > 0) - I(y < x < 0)].$$

Let $\hat{d}(\underline{X}_i) = \hat{m}(\underline{X}_i) - m(\underline{X}_i)$. Then, we have

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n w(\underline{X}_i)\rho(Y_i - \hat{m}(\underline{X}_i)) \\
= \; & \frac{1}{n}\sum_{i=1}^n w(\underline{X}_i)\rho(Y_i - m(\underline{X}_i)) - \frac{1}{n}\sum_{i=1}^n w(\underline{X}_i)(\hat{m}(\underline{X}_i) - m(\underline{X}_i))\varphi(\varepsilon_i) \\
& -\frac{2}{n}\sum_{i=1}^n w(\underline{X}_i)(Y_i - \hat{m}(\underline{X}_i)) \times \left\{I(\hat{d}(\underline{X}_i) > \varepsilon_i > 0) - I(\hat{d}(\underline{X}_i) < \varepsilon_i < 0)\right\} \\
= \; & \frac{1}{n}\sum_{i=1}^n w(\underline{X}_i)\rho(Y_i - m(\underline{X}_i)) + A + B \quad (say).
\end{aligned}$$

From this decomposition, it follows that it is enough to show that both A and B are of order $o_p(n^{-1/2})$. First, we will show that $A = o_p(n^{-1/2})$. Using Lemma 3.2,

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n w(\underline{X}_i)(\hat{m}(\underline{X}_i) - m(\underline{X}_i))\varphi(\varepsilon_i) \\
& = -\frac{1}{n}\sum_{i=1}^n w(\underline{X}_i)\varphi(\varepsilon_i)\underline{e}_1^\top \frac{H_n^{-1}}{nh^d} S_{np}^{-1}(\underline{X}_i)\sum_{j=1}^n K_h(\underline{X}_j - \underline{X}_i)\varphi(\varepsilon_j)\mu(\underline{X}_j - \underline{X}_i) \\
& \quad + \frac{1}{n}\sum_{i=1}^n w(\underline{X}_i)\varphi(\varepsilon_i) \times o_p(1) \\
& = -\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n w(\underline{X}_i)\underline{e}_1^\top S_{np}^{-1}(\underline{X}_i)\frac{K_h(\underline{X}_j - \underline{X}_i)}{h^d} H_n^{-1}\mu(\underline{X}_j - \underline{X}_i)\varphi(\varepsilon_i)\varphi(\varepsilon_j) + o_p(n^{-1/2}) \\
& \equiv -V_n + o_p(n^{-1/2}).
\end{aligned}$$

For the second equality, we used the fact that $\mathrm{E}[w(\underline{X})\varphi(\varepsilon)] = 0$. Consider U_n which is a U-statistic with kernel depending on n:

$$U_n = \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n}\xi_n(\underline{Z}_i, \underline{Z}_j)$$

with

$$\begin{aligned}\underline{Z}_i &= (\underline{X}_i, \varepsilon_i),\ \xi_n(\underline{Z}_i, \underline{Z}_j) = \eta_n(\underline{Z}_i, \underline{Z}_j) + \eta_n(\underline{Z}_j, \underline{Z}_i) \\ \eta_n(\underline{z}_1, \underline{z}_2) &= w(\underline{x}_1)\underline{e}_1^\top S_{np}^{-1}(\underline{x}_1)\frac{K_h(\underline{x}_2 - \underline{x}_1)}{h^d} H_n^{-1}\mu(\underline{x}_2 - \underline{x}_1)\varphi(\varepsilon_1)\varphi(\varepsilon_2),\end{aligned}$$

where $\underline{z}_k = (\underline{x}_k, \varepsilon_k),\ k = 1, 2$. Applying the fact that the smallest eigenvalue of $S_{np}(\underline{x})$ is bounded away from zero, as $n \to \infty$, uniformly over $\underline{x} \in \mathcal{D}$ and $0 < \varphi^2(\cdot) \leq \max\{(2q)^2, (2q-2)^2\}$, we have that

$$\left| V_n - \left(\frac{n-1}{2n}\right) U_n \right| = \left| \frac{1}{n^2} \sum_{i=1}^{n} w(\underline{X}_i)\underline{e}_1^\top S_{np}^{-1}(\underline{X}_i)\frac{K(0)}{h^d}\varphi^2(\varepsilon_i) \right| \leq C/(nh^d).$$

If $h \asymp n^{-\kappa}$ with $1/(2p+2+d) < \kappa < 1/(2d)$, then we have, uniformly in $\underline{x} \in \mathcal{D}$,

$$V_n - \left(\frac{n-1}{2n}\right) U_n = o_p(n^{-1/2}).$$

Finally, we will show that

$$U_n = o_p(n^{-1/2}).$$

Note that

$$\mathrm{E}[\xi_n(\underline{Z}_i, \underline{Z}_j)] = \mathrm{E}[\eta_n(\underline{Z}_i, \underline{Z}_j)] = \mathrm{E}[\xi_n(\underline{Z}_i, \underline{Z}_j)|\underline{Z}_i] = \mathrm{E}[\eta_n(\underline{Z}_i, \underline{Z}_j)|\underline{Z}_i] = 0.$$

Conditioning on $(\underline{X}_1, \underline{X}_2)$, we have

$$\begin{aligned}\mathrm{E}[\xi_n^2(\underline{Z}_1, \underline{Z}_2)] &\leq 4\mathrm{E}[\eta_n^2(\underline{Z}_1, \underline{Z}_2)] \\ &= 4\mathrm{E}\Big[w^2(\underline{X}_1)\varphi(\varepsilon_1)^2\varphi(\varepsilon_2)^2(\underline{e}_1^\top S_{np}^{-1}(\underline{X}_1)H_n^{-1}\mu(\underline{X}_2 - \underline{X}_1))^2(K_h(\underline{X}_2 - \underline{X}_1)/h^d)^2\Big] \\ &= 64q(1-q)\mathrm{E}\Big[w^2(\underline{X}_1)(\underline{e}_1^\top S_{np}^{-1}(\underline{X}_1)H_n^{-1}\mu(\underline{X}_2 - \underline{X}_1))^2(K_h(\underline{X}_2 - \underline{X}_1)/h^d)^2\Big].\end{aligned}$$

Applying the fact that the smallest eigenvalue of $S_{np}(\underline{x})$ is bounded away from zero, as $n \to \infty$, uniformly over $\underline{x} \in \mathcal{D}$, and that each component of $H_n^{-1}\mu(\underline{X}_2 - \underline{X}_1)K_h(\underline{X}_2 - \underline{X}_1)$ is bounded by $\max_{\underline{x}\in\mathcal{D}} w(\underline{x}) < \infty$, we get $\mathrm{E}[\xi_n^2(\underline{Z}_1, \underline{Z}_2)] = O(1/h^d) = o(n)$. If we denote P_n as the projection of U_n, Powell's projection theorem for U-statistics [see, Lemma A.3 in Ahn and Powell (1993) or Lemma 3.1 in Powell et al. (1989)] now gives

$$U_n = P_n + o_p(n^{-1/2}) = o_p(n^{-1/2}),$$

because $P_n = 0$. Now we will prove that $B = o_p(n^{-1/2})$.

Let $I(w) = \{i : \underline{X}_i \in \mathcal{D},\ i = 1, \ldots, n\}$. Then,

$$\begin{aligned}
|B| &\leq \frac{2}{n}\sum_{i=1}^{n} w(\underline{X}_i)(|\hat{d}(\underline{X}_i)| + |\varepsilon_i|)I(|\varepsilon_i| < |\hat{d}(\underline{X}_i)|) \\
&\leq \frac{4}{n}\sum_{i=1}^{n} w(\underline{X}_i)|\hat{d}(\underline{X}_i)|\ I(|\varepsilon_i| < |\hat{d}(\underline{X}_i)|) \\
&\leq 4 \max_{i\in I(w)} |\hat{d}(\underline{X}_i)|\ \frac{1}{n}\sum_{i=1}^{n} w(\underline{X}_i)I(|\varepsilon_i| < |\hat{d}(\underline{X}_i)|) \\
&\leq 4 \max_{i\in I(w)} |\hat{d}(\underline{X}_i)|\ \max_{\underline{x}\in\mathcal{D}} w(\underline{x})\frac{1}{n}\sum_{i=1}^{n} I(|\varepsilon_i| < \max_{j\in I(w)} |\hat{d}(\underline{X}_j)|).
\end{aligned}$$

From the fact that

$$\sup_{a\in\mathbb{R}}\left|\frac{1}{n}\sum_{i=1}^{n} I(|\varepsilon_i| < a) - P(|\varepsilon| < a)\right| = O_p(n^{-1/2}),$$

it follows that

$$\begin{aligned}
|B| &\leq 4 \max_{i\in I(w)} |\hat{d}(\underline{X}_i)| \max_{\underline{x}\in\mathcal{D}} w(\underline{x})\ \left\{P\left(|\varepsilon| < \max_{j\in I(w)} |\hat{d}(\underline{X}_j)|\right) + O_p(n^{-1/2})\right\}. \\
&= 4 \max_{i\in I(w)} |\hat{d}(\underline{X}_i)| \max_{\underline{x}\in\mathcal{D}} w(\underline{x}) \left\{F_\varepsilon\left(\max_{i\in I(w)} |\hat{d}(\underline{X}_i)|\right) - F_\varepsilon\left(-\max_{i\in I(w)} |\hat{d}(\underline{X}_i)|\right)\right\} \\
&\quad +4 \max_{i\in I(w)} |\hat{d}(\underline{X}_i)| \max_{\underline{x}\in\mathcal{D}} w(\underline{x}) \times O_p(n^{-1/2}).
\end{aligned}$$

By (A2), we have

$$|B| \leq C \max_{i\in I(w)} |\hat{d}(\underline{X}_i)| \max_{i\in I(w)} |\hat{d}(\underline{X}_i)| + 4 \max_{i\in I(w)} |\hat{d}(\underline{X}_i)| \max_{\underline{x}\in\mathcal{D}} w(\underline{x}) \times O_p(n^{-1/2}).$$

If $p + 1 > d/2$ and $h \asymp n^{-\kappa}$ with $1/(2p + 2 + d) < \kappa < 1/(2d)$, then,

$$\max_{i\in I(w)} |\hat{d}(\underline{X}_i)| = o_p(n^{-1/4}).$$

Hence, $B = o_p(n^{-1/2})$.

6.4 Proof of Theorem 3.6

The proof is similar to the proof of Theorem 3.4. The only major difference is that we need to show that

$$\frac{1}{n}\sum_{i=1}^{n} w(\underline{X}_i)\rho(Y_i - \hat{\theta}_0 - \hat{\theta}_1^\top \underline{X}_i) = \frac{1}{n}\sum_{i=1}^{n} w(\underline{X}_i)\rho(Y_i - \theta_0^* - \theta_1^{*\top} \underline{X}_i) + o_p(n^{-1/2}).$$

Let $\hat{d}(\underline{X}_i) = (\hat{\theta}_0 - \theta_0^*) + (\hat{\theta}_1 - \theta_1^*)^\top \underline{X}_i$ and $\varepsilon_i^* = Y_i - \theta_0^* - \theta_1^{*\top} \underline{X}_i$. Consider the following decomposition of $\rho(\cdot)$:

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} w(\underline{X}_i)\rho(Y_i - \hat{\theta}_0 - \hat{\theta}_1^\top \underline{X}_i) - \frac{1}{n}\sum_{i=1}^{n} w(\underline{X}_i)\rho(Y_i - \theta_0^* - \theta_1^{*\top} \underline{X}_i) \\
= \; & -\frac{1}{n}\sum_{i=1}^{n} w(\underline{X}_i)((\hat{\theta}_0 - \theta_0^*) + (\hat{\theta}_1 - \theta_1^*)^\top \underline{X}_i)\varphi(\varepsilon_i^*) \\
& -\frac{2}{n}\sum_{i=1}^{n} w(\underline{X}_i)(\varepsilon_i^* - \hat{d}(\underline{X}_i)) \times \Big\{ I(\hat{d}(\underline{X}_i) > \varepsilon_i^* > 0) - I(\hat{d}(\underline{X}_i) < \varepsilon_i^* < 0) \Big\} \\
= \; & C + D \;\; (say).
\end{aligned}$$

First, we will show that $C = o_p(n^{-1/2})$.

$$C = -(\hat{\theta}_0 - \theta_0^*, (\hat{\theta}_1 - \theta_1^*)^\top)\frac{1}{n}\sum_{i=1}^{n} \underline{X}_{ei} w(\underline{X}_{ei})\varphi(\varepsilon_i^*).$$

Since $\mathrm{E}[\underline{X}_e \varphi(\varepsilon_*) w(\underline{X}_e)] = 0$ (orthogonality condition),

$$\frac{1}{n}\sum_{i=1}^{n} \underline{X}_{ei} w(\underline{X}_{ei})\varphi(\varepsilon_i^*) = O_p(n^{-1/2}). \tag{6.1}$$

Hence, (6.1) combined with Lemma 3.5 produces the desired result. Similar to the term B, we can show that the term D is $o_p(n^{-1/2})$. Let $I(w) = \{i : \underline{X}_i \in \mathcal{D},\ i = 1, \ldots, n \;\text{ where }\; \underline{X}_{ei} = (1, \underline{X}_i)\}$. Then,

$$\begin{aligned}
|D| \;\; \leq \;\; & \frac{2}{n}\sum_{i=1}^{n} w(\underline{X}_i)(|\hat{d}(\underline{X}_i)| + |\varepsilon_i^*|) I(|\varepsilon_i^*| < |\hat{d}(\underline{X}_i)|) \\
\leq \;\; & \frac{4}{n}\sum_{i=1}^{n} w(\underline{X}_i)|\hat{d}(\underline{X}_i)| \; I(|\varepsilon_i^*| < |\hat{d}(\underline{X}_i)|) \\
\leq \;\; & 4 \max_{i \in I(w)} |\hat{d}(\underline{X}_i)| \max_{\underline{x} \in \mathcal{D}} w(\underline{x}) \frac{1}{n}\sum_{i=1}^{n} I(|\varepsilon_i^*| < \max_{i \in I(w)} |\hat{d}(\underline{X}_i)|).
\end{aligned}$$

By the classical Donsker theorem,

$$|D| \leq 4 \max_{i \in I(w)} |\hat{d}(\underline{X}_i)| \max_{\underline{x} \in \mathcal{D}} w(\underline{x}) \{P(|\varepsilon^*| < \max_{i \in I(w)} |\hat{d}(\underline{X}_i)|) + O_p(n^{-1/2})\}.$$

Since $\mathcal{D}$ is bounded, $\max_{i \in I(w)} |\hat{d}(\underline{X}_i)| = O_p(n^{-1/2})$ from Lemma 3.5. This result and the boundedness of the marginal density of ε^* proves that $D = o_p(n^{-1/2})$.

References

H. Ahn and J. L. Powell. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58:3–29, 1993.

G. Chamberlain. Quantile regression, censoring, and the structure of wages. In *Advances in econometrics, Sixth World Congress, Vol. I (Barcelona, 1990)*, volume 23 of *Econom. Soc. Monogr.*, pages 171–209. Cambridge Univ. Press, Cambridge, 1994.

P. Chaudhuri, K. Doksum, and A. Samarov. On average derivative quantile regression. *The Annals of Statistics*, 25:715–744, 1997.

K. Doksum and A. Samarov. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, 23:1443–1473, 1995.

A. El Ghouch, M. G. Genton, and T. Bouezmarni. Measuring the discreapancy of a parametric model via local polynomial smoothing. Technical report, Université catholique de Louvain, 2010.

J. Kiefer. On Bahadur's representation of sample quantiles. *The Annals of Mathematical Statistics*, 38:1323–1342, 1967.

T.-W. Kim and H. White. Estimation, inference, and specification testing for possibly misspecified quantile regression. *Advances in Econometrics*, 17:107–132, 2003.

R. Koenker and J. A. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94:1296–1310, 1999.

E. Kong, O. Linton, and Y. Xia. Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory*, 26:1529–1564, 2010.

O. Linton. Second order approximation in the partially linear regression model. *Econometrica*, 63: 1079–1112, 1995.

E. Masry. Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series and Analysis*, 17:571–599, 1996.

J. W. Mckean and G. L. Sievers. Coefficient of determination for least absolute deviation analysis. *Statistics & Probability Letters*, 5:49–54, 1987.

J. L. Powell and T. M. Stoker. Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, 75:291–316, 1996.

J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57:1403–1430, 1989.

H. Tong and Q. Yao. Nonparametric estimation of ratios of noise to signal in stochastic regression. *Statistica Sinica*, 10:751–770, 2000.

I. Van Keilegom and L. Wang. Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electronic Journal of Statistics*, 4:133–160, 2010.

	R^1				Φ			
τ \ λ	0.00	0.80	1.50	2.50	0.00	0.80	1.50	2.50
0.10	84.20	87.96	92.00	94.77	0.00	82.77	90.74	94.43
0.30	56.10	65.90	76.57	84.44	0.00	53.17	73.01	83.42
0.50	36.07	48.16	62.45	74.62	0.00	33.31	57.38	73.02

Table 1: $R^1(0.5)$ and $\Phi(0.5)$, multiplied by 100, when $w(x) = I(0 \leq x \leq 1)$.

			R^1				Φ			
	n	λ	0.00	0.80	1.50	2.50	0.00	0.80	1.50	2.50
$\tau = 0.1$	100	Bias	0.76	1.55	1.25	0.88	2.51	2.30	1.33	0.96
		STDV	1.86	1.71	0.99	0.62	4.46	1.95	1.08	0.66
	200	Bias	0.32	0.96	0.76	0.55	1.66	1.01	0.67	0.46
		STDV	0.97	1.00	0.67	0.44	2.85	1.41	0.89	0.56
	400	Bias	0.12	0.58	0.48	0.39	0.90	0.79	0.54	0.39
		STDV	0.76	0.81	0.53	0.38	2.06	0.98	0.54	0.37
$\tau = 0.3$	100	Bias	1.91	2.83	2.16	1.83	2.70	3.13	2.29	1.63
		STDV	4.13	4.00	2.76	1.86	7.06	5.62	3.43	2.23
	200	Bias	0.77	1.65	1.25	0.95	1.75	1.91	1.49	1.13
		STDV	2.95	2.72	1.79	1.16	3.07	4.14	2.36	1.44
	400	Bias	0.29	0.61	0.55	0.50	0.91	1.13	0.88	0.71
		STDV	2.01	2.00	1.35	0.93	2.09	2.48	1.47	0.93
$\tau = 0.5$	100	Bias	2.50	3.58	3.08	2.48	3.32	5.15	3.75	2.94
		STDV	6.02	5.56	4.25	2.93	5.36	6.70	4.99	3.17
	200	Bias	1.31	2.37	2.40	1.75	1.54	2.23	1.76	1.26
		STDV	4.38	3.50	2.55	1.92	2.43	4.26	2.85	2.05
	400	Bias	0.41	1.29	1.21	0.84	0.90	1.52	1.33	0.93
		STDV	2.61	2.49	2.00	1.44	2.09	2.98	2.04	1.38

Table 2: Biases and standard deviations of $\widehat{R^1(0.5)}$ and $\widehat{\Phi(0.5)}$, multiplied by 100, based on the local linear estimators of $m(\cdot)$ in Monte Carlo trials with 100 replications and sample size $n = 100, 200$ and 400.

	X_1	X_2	X_3	X_1, X_2	X_2, X_3	X_1, X_3	X_1, X_2, X_3
$R^1(0.5)$	70.91	0.79	0.86	72.35	1.13	72.90	74.54
$R^1(0.4)$	69.98	0.94	0.92	72.20	1.31	71.79	74.35
$R^1(0.3)$	68.41	0.96	0.89	71.60	1.40	70.00	73.75
$R^1(0.2)$	65.80	1.10	1.01	70.43	1.72	67.24	72.61
$R^1(0.1)$	61.40	1.69	1.28	68.37	2.55	62.74	70.64
η^2	89.50	0.86	1.11	90.45	1.97	90.69	91.64

Table 3: $R^1(q)(q = 0.1, \ldots, 0.5)$ and η^2 values, multiplied by 100, for all the subsets of X_1, X_2 and X_3.

		X_1	X_1, X_2	X_1, X_3	X_1, X_2, X_3
$\widehat{R^1(0.5)}$	Median	0.7145	0.7310	0.7343	0.7581
	Bias	0.0043	0.0088	0.0078	0.0141
	STDV	0.0224	0.0201	0.0196	0.0222
$\widehat{R^1(0.2)}$	Median	0.6711	0.7115	0.6826	0.7371
	Bias	0.0138	0.0078	0.0144	0.0136
	STDV	0.0343	0.0263	0.0328	0.0259

Table 4: Medians, biases and standard deviations of $\widehat{R^1(0.5)}$ and $\widehat{R^1(0.2)}$ for the subsets which contain X_1 based on the local linear estimators of $m(\cdot)$ in Monte Carlo trials with 100 replications and sample size $n = 200$.

Model	$S1$	$S2$	$S3$	$S4$
(X_1', X_2')	$X_1, \exp(X_2)$	$X_1, \sqrt{\lvert X_2 \rvert}$	X_1^2, X_2	$X_1, \sin(\pi X_2)$

Table 5: Specifications of the covariates of four parametric linear models.

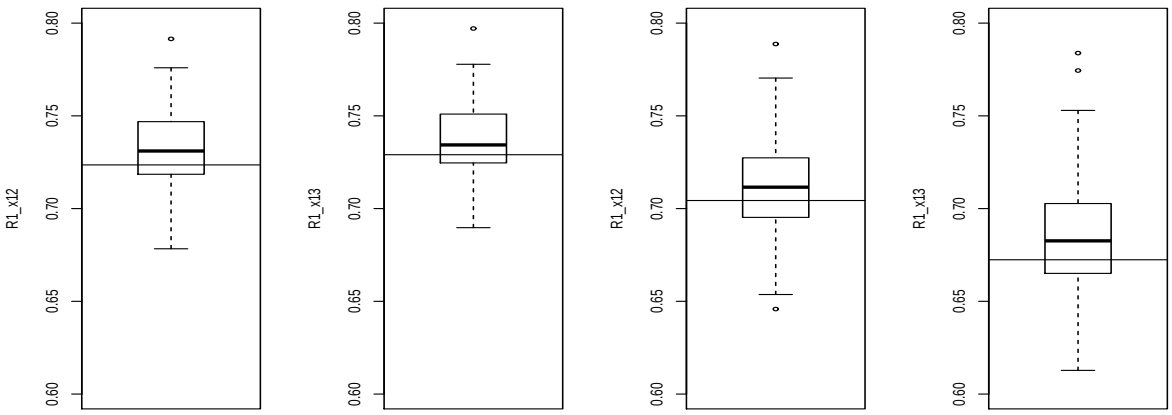


Figure 1: Boxplots of $\widehat{R^1(0.5)}$ (left two panels) and $\widehat{R^1(0.2)}$ (right two panels) for the subsets (X_1, X_2) and (X_1, X_3). The horizontal solid line in each plot represents the corresponding true value.

		$S1$	$S2$	$S3$	$S4$
$\widehat{\Phi(0.5)}$	True value	0.0000	0.0851	0.1561	0.4026
	Median	0.0138	0.1003	0.1746	0.4145
	Bias	0.0228	0.0205	0.0197	0.0113
	STDV	0.0342	0.0437	0.0433	0.0409
$\widehat{\Phi(0.2)}$	True value	0.0000	0.1047	0.1691	0.4242
	Median	0.0187	0.1247	0.1892	0.4403
	Bias	0.0321	0.0303	0.0295	0.0197
	STDV	0.0475	0.0575	0.0617	0.0430

Table 6: True values, medians, biases and standard deviations of $\widehat{\Phi(0.5)}$ and $\widehat{\Phi(0.2)}$ for four surrogate linear models based on the local linear estimators of $m(\cdot)$ in Monte Carlo trials with 100 replications and sample size $n = 200$.

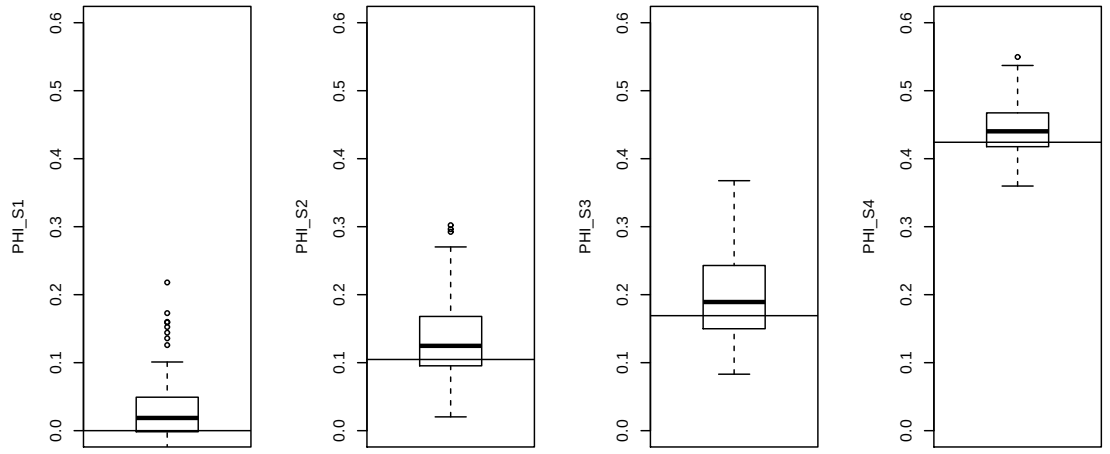


Figure 2: Boxplots of $\widehat{\Phi(0.2)}$ for the four surrogate parametric models. The horizontal solid line in each plot represents the corresponding true value.

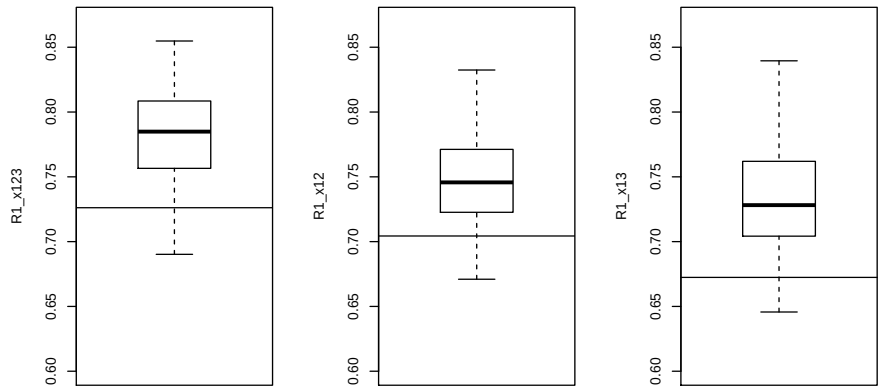


Figure 3: Boxplots of $\widehat{R^1(0.2)}$ for the selected models using the local constant fit ($p = 0$). The horizontal solid line in each plot represents the corresponding true value.

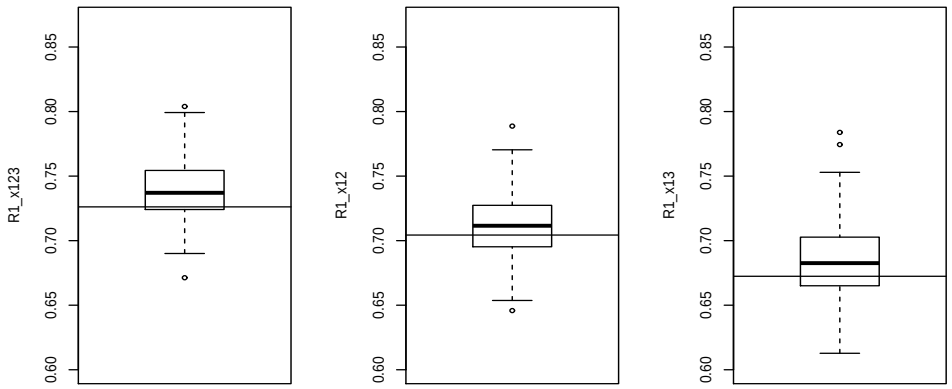


Figure 4: Boxplots of $\widehat{R^1(0.2)}$ for the selected models using the local linear fit ($p = 1$). The horizontal solid line in each plot represents the corresponding true value.

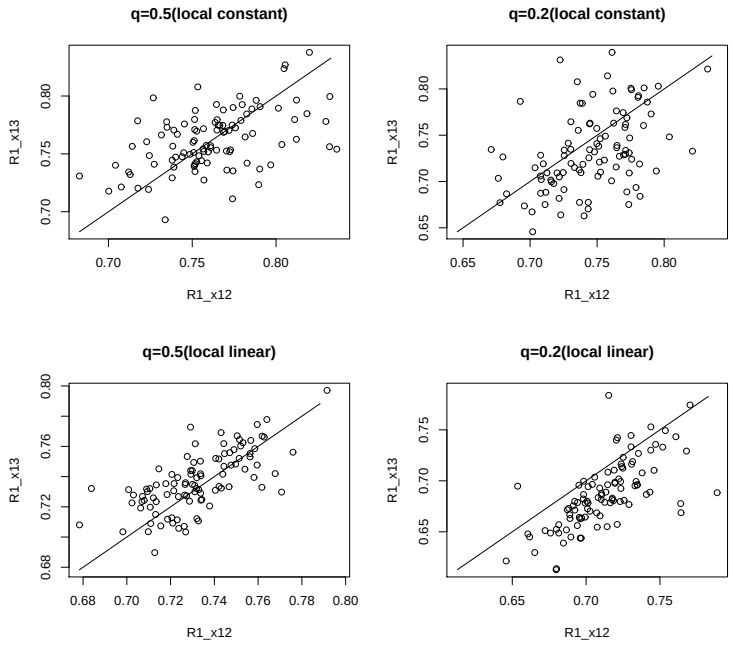


Figure 5: Scatter plot of $\hat{R}^1(X_1, X_2)$ versus $\hat{R}^1(X_1, X_3)$ using the local constant (upper panel) and local linear (lower panel) fit when $q = 0.5$ (left) and 0.2 (right). The solid line in each plot represents the bisector.