Focused information criterion and model averaging based on weighted composite quantile regression

GANGGANG XU, SUOJIN WANG and JIANHUA Z. HUANG
Department of Statistics, Texas A&M University

ABSTRACT. We study the focused information criterion (FIC) and frequentist model averaging (FMA) and their application to post-model-selection inference for weighted composite quantile regression (WCQR) in the context of the additive partial linear models. With the nonparametric functions approximated by polynomial splines, we show that, under certain conditions, the asymptotic distribution of the FMA WCQR-estimator of a focused parameter is a nonlinear mixture of normal distributions. This asymptotic distribution is used to construct confidence intervals that achieve the nominal coverage probability. With properly chosen weights, the FIC-based WCQR estimators are not only robust to outliers and nonnormal residuals but also can achieve efficiency close to the maximum likelihood estimator (MLE), without assuming the true error distribution. Simulation studies and a real data analysis are used to illustrate the effectiveness of the proposed procedure.

Key words: Focused information criterion; Frequentist model averaging; Model inference; Weighted composite quantile estimator.

Running headline: Model averaging using weighted composite quantile regression

1 Introduction

While model selection becomes an important aspect of statistical modeling, the impact of model selection on the subsequent statistical inference is much less understood (Leeb & Potscher, 2005; Leeb & Potscher, 2006). Ignoring uncertainty in the model selection step can result in inaccurate or even misleading inferences. For example, some initial investigation suggests that the finite sample distribution of a post-model-selection estimator in a linear regression model can be highly nonnormal and is often multimodal and the actual coverage probability of the naive confidence interval based on the selected model can be significantly smaller than its nominal level (Leeb & Potscher, 2005; Kabaila & Leeb, 2006). A second problem with the traditional model selection procedures is that once a best model is selected, all statistical inferences are carried out using the same model, regardless of the purpose of the inference.

However, estimating a different parameter may require using a different model as the best model, which has been confirmed by many researchers (Claeskens & Hjort, 2003; Hjort & Claeskens, 2003; Claeskens et al. 2006)

In attempting to address these two problems, Claeskens & Hjort (2003) and Hjort & Claeskens (2003) proposed the focused information criterion (FIC) and the frequentist model averaging (FMA) framework, which has received much attention in the literature and has been widely used and further developed by many researchers. For example, Hjort & Claeskens (2006) studied FIC and FMA for the Cox hazard regression model. Claeskens et al. (2007) used FIC to determine the order of the autoregressive model. More recently, Claeskens & Carroll (2007) and Zhang & Liang (2011) studied the asymptotic properties of FIC and FMA in the semiparametric partially linear model.

The FIC and the FMA framework are appealing in that they provide more stable estimators with better finite sample properties. However, one limitation of existing work of FIC and FMA is that they almost exclusively focus on the likelihood based approach (e.g., Claeskens & Hjort, 2003; Hjort & Claeskens, 2003) and few, if any, have studied other types of estimators such as M-estimators, parametrically or semiparametrically. One exception is the recent work of Zhang & Liang (2011), which studied the quasi-likelihood based FIC and FMA framework. However, they did not consider the estimation efficiency of the proposed estimator. In fact, their method reduces to the least square estimation under Model (1) below, which may be unsatisfactory because the least square estimator is known to be sensitive to outliers and can be quite inefficient if the residual distribution is far from normal. Considering that one of the major motivations of FIC is to provide more accurate estimation for the focused parameter, lack of knowledge of the likelihood function can greatly compromise the effectiveness of the FIC procedure. For example, using a misspecified likelihood function can lead to estimators with higher mean squared error and unnecessarily wider confidence intervals. Therefore, it is of great interest to develop more flexible FIC estimation procedure that can maintain high efficiency when the true likelihood function is not available.

In this paper, we propose to fill this gap by proposing FIC and FMA based on the WCQR estimator (Zou & Yuan, 2008), whose efficacy are closely related to the MLE. Our main contributions in this work are two-fold: (1) to establish properties of FIC and FMA for WCQR-estimators in Model (1) and use these properties to conduct reliable post model selection inferences; and (2) to obtain highly efficient estimators without assuming the distribution of ε in Model (1). The asymptotic theory and numerical examples given in this paper show that, by carefully choosing weights, the FIC-based WCQR estimator (F-WCQR) can achieve efficiency close to the FIC-based MLE (F-MLE). In addition, the confidence

intervals based on F-WCQR estimator achieve the nominal coverage probability, which is not the case for naive confidence intervals based on model selection procedures such as AIC or BIC. While our results are of their own interest even just for linear models, we choose to present our results for the partially linear additive Model (1) not only because it is technically more challenging, but also because it offers much more flexibility by including nonparametric regressors and thus is more useful in practice.

In an independent work, Behl et al. (2013) studied the focused model selection in quantile regression under the assumption that the specific form of the conditional quantile of the response Y given covariates $\mathbf{X}$ is known up to a set of parameters. They did not address the problem of how to obtain highly efficient FMA estimator without knowing the true likelihood function. In addition, their model is essentially a parametric one while this paper considers the semiparametric additive partially linear model.

The rest of the article is organized as follows. In Section 2, we introduce some notations and conditions needed for the theoretical investigation. In section 3, we study asymptotic properties of the FIC-based WCQR-estimator, where the nonparametric components in Model (1) are approximated by polynomial splines. Simulation results are given in Section 4 and a real data example is presented in Section 5. All technical proofs are given in the Appendix and in the Supplementary Material.

2 Notations and Conditions

Let Y be the response variable and $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ and $\mathbf{X} = (X_1, \ldots, X_m)^T$ be covariates. Consider the partially linear additive regression model

$$Y_i = a_0 + \sum_{j=1}^{m} f_j(X_{ji}) + \mathbf{Z}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ is a realization of $(Y, \mathbf{X}, \mathbf{Z})$ with $\mathbf{X}_i = (X_{1i}, \ldots, X_{mi})^T$, f_j $(j = 1, \ldots, m)$ are univariate smooth functions, and ε are independent and identically distributed errors that are independent of $(\mathbf{Z}, \mathbf{X})$. For the identifiability purpose, we assume that $E\{f_j(X_j)\} = 0$ for each f_j and $E(\mathbf{Z}) = \mathbf{0}$. Without loss of generality, we assume that the support of each covariate X_j is $[0, 1]$ for $j = 1, \ldots, m$.

Model (1) has been extensively studied (e.g. He & Shi, 1996; Huang et al., 2007; Liu et al., 2011) since it enjoys the flexibility of the nonparametric regression models and in the meantime allows us to do statistical inference on the parametric components. The existence of multiple nonparametric components in (1) often makes the computation and the theoretical development of the estimation difficult. In this paper, we use polynomial splines to approximate the nonparametric components, which is computationally more efficient and easier to implement than the kernel-based backfitting procedures.

Suppose that the covariates in the parametric part $\mathbf{Z}$ can be categorized into two types: p_f covariates $\mathbf{Z}_f$ that are surely included in all submodels and p_u covariates $\mathbf{Z}_u$ that may or may not be in the model. Therefore, there are in total 2^{p_u} possible submodels. Let $(\mathbf{Z}_f, \mathbf{Z}_{u,S}, \mathbf{X})$ be the set of covariates used in the Sth submodel, and Π_S be the projection matrix such that $\Pi_S\mathbf{Z} = (\mathbf{Z}_f^T, \mathbf{Z}_{u,S}^T)^T$ and $\boldsymbol{\beta}_S = \Pi_S\boldsymbol{\beta} = (\boldsymbol{\beta}_f^T, \boldsymbol{\beta}_{u,S}^T)^T$. Following Claeskens & Hjort (2003) and Hjort & Claeskens (2003), we consider the local misspecification model where we assume that the true value of $\boldsymbol{\beta} = (\boldsymbol{\beta}_f^T, \boldsymbol{\beta}_u^T)^T$ is $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{f,0}^T, n^{-1/2}\boldsymbol{\delta}^T)^T$ with $\boldsymbol{\beta}_{f,0}$ representing the $p_f \times 1$ coefficient vector of $\mathbf{Z}_f$ and $\boldsymbol{\delta}$ being a $p_u \times 1$ vector. Such an assumption depicts scenarios where some of coefficients are very small in a regression model. Although in reality the true parameter would not change as the sample size changes, the asymptotic theory based on this type of fictitious parameters provides more accurate insights into the finite sample property of an estimator (Leeb & Potscher, 2005).

Write $r > 1$ as $r = r_0 + r_1$, where r_0 is an integer and $0 < r_1 \leq 1$. Let $\mathcal{H}_r$ be the space of functions f on $[0, 1]$ whose r_0th derivative, $f^{(r_0)}$, satisfies a Hölder Condition with exponent r_1, that is, there exists a constant $C_0 > 0$ such that

$$|f^{(r_0)}(t_1) - f^{(r_0)}(t_2)| \leq C_0|t_1 - t_2|^{r_1}$$

for $0 \leq t_1, t_2 \leq 1$. Let $f_0(\mathbf{x}) = \sum_{j=1}^m f_j(x_j)$ be the true additive function. If each component function $f_j \in \mathcal{H}_r$, then f_j can be well approximated using polynomial spline basis functions. Let $\mathcal{G}$ be a space of polynomial splines with degree d defined on $[0, 1]$. Given a nonnegative integer d and a positive integer J_n, consider a knot sequence with J_n interior knots, $0 = t_{-d} = \cdots = t_0 < t_1 < \cdots < t_{J_n} < t_{J_n+1} = \cdots = t_{J_n+d+1} = 1$ such that

$$\frac{\max_{1\leq l\leq J_n+1}(t_l - t_{l-1})}{\min_{1\leq l\leq J_n+1}(t_l - t_{l-1})}$$

is uniformly bounded. Denote $b_{l,j}(x_j)$ as the B-spline basis functions of degree d for function $f_j(x_j)$, $(l = -d, \ldots, J_n; j = 1, \ldots, m)$. Then $\sum_{j=1}^m f_j(x_j)$ can be well approximated by $\boldsymbol{b}(\mathbf{X})^T\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma}$, where $\boldsymbol{b}(\mathbf{X}) = \{b_{l,j}(x_j), l = -d, \ldots, J_n, j = 1, \ldots, m\}^T$. To ensure the identifiability, all covariates and basis functions are centered such that $\sum_{i=1}^n \mathbf{Z}_i = 0$ and $\sum_{i=1}^n \boldsymbol{b}(\mathbf{X}_i) = 0$.

For a given positive integer K, define $\xi_k = G^{-1}(\tau_k)$, $(k = 1, \ldots, K)$ with $G(\cdot)$ being the cumulative distribution function (CDF) of ε and let $\Gamma(\mathbf{x}) = E(\mathbf{Z}|\mathbf{X} = \mathbf{x})$ and $\boldsymbol{\Sigma} = \mathrm{var}\{\mathbf{Z} - \Gamma(\mathbf{X})\}$. Let $||\cdot||$ be the Euclidean norm of a vector and use $a_n \ll b_n$ to mean $b_n^{-1}a_n \to 0$ as $n \to \infty$. The following conditions are assumed to achieve our theoretical results.

Condition 1: For $j = 1, \ldots, m$, $f_j \in \mathcal{H}_r$ and the density function of X_j is bounded away from zero and infinity on $[0, 1]$.

Condition 2: $\Gamma(\mathbf{x}) = \Gamma_1(x_1) + \cdots + \Gamma_m(x_m)$ with each element $\Gamma_j(x_j) \in \mathcal{H}_r$ and $E||\mathbf{Z}||^3 < \infty$. Both $\mathrm{var}(\mathbf{Z})$ and $\mathbf{\Sigma}$ are invertible almost surely.

Condition 3: The number of knots J_n satisfies $n^{1/(2r)} \ll J_n \ll n^{1/2}$.

Condition 4: The density function of ε, $g(\varepsilon)$, is positive and Lipschitz in neighborhoods of $\xi_1, \ldots, \xi_K$.

Let $\mathbf{B}_n = \{\boldsymbol{b}(\mathbf{X}_1), \ldots, \boldsymbol{b}(\mathbf{X}_n)\}^T$, $\mathbf{P}_n = \mathbf{B}_n(\mathbf{B}_n^T\mathbf{B}_n)^-\mathbf{B}_n^T$. Then under Condition 2, it is straightforward to show that, as $n \to \infty$,

$$\widehat{\mathbf{\Sigma}}_n = \frac{1}{n}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)(\mathbf{I} - \mathbf{P}_n)(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T \to \mathbf{\Sigma} \quad \text{in probability,}$$

where $(\mathbf{B}_n^T\mathbf{B}_n)^-$ is the generalized inverse of $\mathbf{B}_n^T\mathbf{B}_n$. Condition 2 was also used in for example, He & Shi (1996) and Liu et al. (2011).

3 Model averaging based on the WCQR

The WCQR estimator of the Sth submodel $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}_{S,wcqr})$ is obtained by minimizing

$$\ell(\boldsymbol{a}, \boldsymbol{\gamma}, \boldsymbol{\beta}_S) = \sum_{k=1}^{K}\sum_{i=1}^{n} w_k \rho_{\tau_k}(Y_i - a_k - \boldsymbol{b}(\mathbf{X}_i)^T\boldsymbol{\gamma} - (\Pi_S\mathbf{Z}_i)^T\boldsymbol{\beta}_S), \tag{2}$$

where $\rho_{\tau_k}(s) = \tau_k I(s > 0) + (1 - \tau_k)I(s \le 0)$ is the quantile regression check functions (Koenker & Bassett, 1978) corresponding to the $100\tau_k$th quantile with $\tau_k \in (0, 1)$. The vector $\boldsymbol{w} = (w_1, \ldots, w_K)^T$ is a vector of some positive pre-selected weights and $\boldsymbol{a} = (a_1, \ldots, a_K)^T$ is a vector of intercepts.

The WCQR estimator has been used in the literature by several authors in the non FIC setting. Zou & Yuan (2008) and Kai et al. (2011) considered a simpler version of WCQR estimators with equal weights (w_k's are all 1's) for variable selection in linear regression model and semiparametric varying-coefficient models, respectively. Bradic et al. (2011) proposed to use the WCQR estimators to conduct variable selection in ultra-high dimensional regression models. But none of these works has considered reliable post model selection inference and model averaging.

The following lemma shows that under local misspecification setup, if the number of knots is chosen in an appropriate manner, all submodels would yield consistent estimates for both parametric and nonparametric components.

LEMMA 1. *Under Conditions 1–4, we have that*

$$\left|\left|\hat{\boldsymbol{\beta}}_{S,wcqr} - \begin{pmatrix} \boldsymbol{\beta}_{f,0} \\ 0 \end{pmatrix}\right|\right| = O_p(n^{-1/2}J_n^{1/2}), \quad ||\hat{\boldsymbol{a}} - \boldsymbol{a}|| = O_p(n^{-1/2}J_n^{1/2})$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(\mathbf{X}_i) - f_0(\mathbf{X}_i))^2 = O_p(n^{-1}J_n),$$

where $\hat{f}_n(\mathbf{X}) = \boldsymbol{b}(\mathbf{X})^T\hat{\boldsymbol{\gamma}}$ *and* $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}_{S,wcqr})$ *is the minimizer of (2).*

A proof of Lemma 1 is given in the Appendix.

The next theorem is the main result of this paper, which states that under the local misspecification assumption, the parametric estimate is root-n consistent with an asymptotically normal distribution for all submodels, but the bias is not negligible if a submodel is used instead of the full model. This result leads to two important corollaries that justify the proposed methodology in this paper.

Let $b = \sum_{k=1}^{K} w_k g(\xi_k)$ and the derivative $\rho'_{\tau_k}(s) = \phi_{\tau_k}(s) = \tau_k - I(s < 0)$ for $k = 1, \ldots, K$, and then define quantities

$$\mathbf{Z}_i^* = \mathbf{Z}_i - \sum_{j=1}^{n} B(\mathbf{X}_j)^T(\mathbf{B}_n^T\mathbf{B}_n)^- B(\mathbf{X}_i)\mathbf{Z}_j, \mathbf{G}_n = \frac{1}{bn^{1/2}}\sum_{k=1}^{k}\sum_{i=1}^{n} w_k\phi_{\tau_k}(\varepsilon_i - \xi_k)\mathbf{Z}_i^*.$$

THEOREM 1. *Under Conditions 1–4, minimizing (2) gives*

$$\begin{aligned} n^{1/2}\left\{\hat{\boldsymbol{\beta}}_{S,wcqr} - \begin{pmatrix} \boldsymbol{\beta}_{f,0} \\ 0 \end{pmatrix}\right\} &= (\Pi_S\boldsymbol{\Sigma}\Pi_S^T)^{-1}\Pi_S\mathbf{G}_n + (\Pi_S\boldsymbol{\Sigma}\Pi_S^T)^{-1}\Pi_S\boldsymbol{\Sigma}\begin{pmatrix} 0 \\ \boldsymbol{\delta} \end{pmatrix} + o_p(1) \\ &\xrightarrow{d} (\Pi_S\boldsymbol{\Sigma}\Pi_S^T)^{-1}\Pi_S\mathbf{G}\sigma_{wcqr}(\boldsymbol{w}) + (\Pi_S\boldsymbol{\Sigma}\Pi_S^T)^{-1}\Pi_S\boldsymbol{\Sigma}\begin{pmatrix} 0 \\ \boldsymbol{\delta} \end{pmatrix}, \end{aligned} \tag{3}$$

with $\mathbf{G}_n \xrightarrow{d} \mathbf{G} \sim N(0, \boldsymbol{\Sigma})$*, where "*$\xrightarrow{d}$*" denotes convergence in distribution and the variance*

$$\sigma^2_{wcqr}(\boldsymbol{w}) = \frac{\sum_{k,l=1}^{K} w_k w_l\{\min(\tau_k, \tau_l) - \tau_k\tau_l\}}{\{\sum_{k=1}^{K} w_k g(\xi_k)\}^2}. \tag{4}$$

A proof of Theorem 1 is given in the Appendix.

3.1 Rationale behind using the WCQR

When the true underlying distribution of ε is known, we can replace the loss function defined in (2) by the log-likelihood function. Denote the resulting estimator of the linear coefficients of the Sth submodel

as $\hat{\boldsymbol{\beta}}_{S,mle}$. Using the result of Claeskens & Carroll (2007), we can show that the conclusion of Theorem 1 still holds except that $\sigma^2_{wcqr}(\boldsymbol{w})$ is replaced by

$$\sigma^2_{mle} = \frac{1}{\int (g'/g)^2 g\, dt} = \int_0^1 \int_0^1 \frac{\{\min(s,t) - st\}}{g(G^{-1}(s))g(G^{-1}(t))}\, d\Omega(s)\, d\Omega(t), \tag{5}$$

with

$$\Omega'(t) = (g'/g)'(G^{-1}(t)) \Big/ \int_{-\infty}^{\infty} (g'/g)'(x)\, dG(x). \tag{6}$$

A first look at (4) and (5) does not suggest any relationship between $\sigma^2_{wcqr}(\boldsymbol{w})$ and σ^2_{mle}. To see their connection, consider the quantity

$$\sigma^2_{wclr}(\boldsymbol{w}) = \sum_{k,l=1}^{K} \frac{w_k w_l \{\min(\tau_k, \tau_l) - \tau_k \tau_l\}}{g(\xi_k) g(\xi_l)} \tag{7}$$

subject to the constraints that $\sum_{k=1}^K w_k = 1$ and $w_k \geq 0$ for $k = 1, \ldots, K$. We can see that (7) can be considered as a discrete version of (5). If the density function $g(\cdot)$ is uni-modal, then (6) is always positive and by taking $w_k = d\Omega(\tau_k)$, we have $\lim_{K\to\infty} \sigma^2_{wclr}(\boldsymbol{w}) = \sigma^2_{mle}$. On the other hand, letting $\boldsymbol{w}^* = (w_1^*, \ldots, w_K^*)^T$ with $w_k^* = \{\sum_{j=1}^K g(\xi_j) w_j\}^{-1} w_k g(\xi_k)$, it is straightforward to show that

$$\sigma^2_{wcqr}(\boldsymbol{w}) = \sigma^2_{wclr}(\boldsymbol{w}^*).$$

Notice that the relationship between $\boldsymbol{w}$ and $\boldsymbol{w}^*$ is one-to-one, which implies that

$$\min_{\boldsymbol{w}\geq 0} \sigma^2_{wcqr}(\boldsymbol{w}) = \min_{\boldsymbol{w}\geq 0, \sum_{k=1}^K w_k = 1} \sigma^2_{wclr}(\boldsymbol{w}).$$

From the above discussion, for a finite K, it is reasonable to believe that the natural choice of

$$\boldsymbol{w}^{opt} = \min_{\boldsymbol{w}\geq 0} \sigma^2_{wcqr}(\boldsymbol{w}) \tag{8}$$

can yields WCQR estimators with efficiency close to the MLE estimator. This conjecture is confirmed by our empirical findings in the simulation studies, where the relative efficiencies of the FIC-based WCOR estimator to the FIC-based MLE are mostly more than 85% for six different common error distributions; see the simulation section. A formal theoretical investigation would be an interesting future research topic but shall not be pursued in this paper.

3.2 Focused Information Criterion based on WCQR

We are interested in estimating a specific function of the regression coefficients $\mu_{true} = \mu(\boldsymbol{\beta}_0) = \mu(\boldsymbol{\beta}_{f,0}, n^{-1/2}\boldsymbol{\delta})$, which has continuous partial derivatives in a neighborhood of $(\boldsymbol{\beta}_{f,0}^T, \mathbf{0}^T)^T$. In the Sth

submodel, μ_{true} is estimated by plugging in $\hat{\boldsymbol{\beta}}_S$ for those covariates in the model and 0 for covariates that are not in the model, that is, $\hat{\mu}_S = \mu(\Pi_S^T\hat{\boldsymbol{\beta}}_S)$. Let $\hat{\mu}_{S,wcqr} = \mu(\Pi_S^T\hat{\boldsymbol{\beta}}_{S,wcqr})$. Then the following corollary is an immediate result of Theorem 1.

COROLLARY 1. *Under Conditions 1–4, we have that*

$$\begin{aligned} n^{1/2}(\hat{\mu}_{S,wcqr} - \mu_{true}) &= \mu_{\boldsymbol{\beta}}^T\mathbf{H}_S\mathbf{G}_n + \mu_{\boldsymbol{\beta}}^T(\mathbf{H}_S\boldsymbol{\Sigma} - \mathbf{I})\begin{pmatrix} 0 \\ \boldsymbol{\delta} \end{pmatrix} + o_p(1) \\ &\xrightarrow{d} \Lambda_S \equiv \mu_{\boldsymbol{\beta}}^T\mathbf{H}_S\mathbf{G}\sigma_{wcqr}(\boldsymbol{w}) + \mu_{\boldsymbol{\beta}}^T(\mathbf{H}_S\boldsymbol{\Sigma} - \mathbf{I})\begin{pmatrix} 0 \\ \boldsymbol{\delta} \end{pmatrix}, \end{aligned} \tag{9}$$

where $\mathbf{H}_S = \Pi_S^T(\Pi_s\boldsymbol{\Sigma}\Pi_S^T)^{-1}\Pi_S$ *and* $\mu_{\boldsymbol{\beta}} = \frac{\partial\mu(\boldsymbol{\beta}_f,\boldsymbol{\beta}_u)}{\partial\boldsymbol{\beta}}|_{\boldsymbol{\beta}_f=\boldsymbol{\beta}_{f,0},\boldsymbol{\beta}_u=0}$.

The proof is a direct application of the delta method and is thus omitted here. A straightforward calculation yields that

$$E(\Lambda_{S,wcqr}^2) = \mu_{\boldsymbol{\beta}}^T\left\{\sigma_{wcqr}^2(\boldsymbol{w})\mathbf{H}_S\boldsymbol{\Sigma}\mathbf{H}_S + (\mathbf{H}_S\boldsymbol{\Sigma} - \mathbf{I})\begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{\delta}\boldsymbol{\delta}^T \end{pmatrix}(\mathbf{H}_S\boldsymbol{\Sigma} - \mathbf{I})^T\right\}\mu_{\boldsymbol{\beta}}. \tag{10}$$

Claeskens & Hjort (2003) proposed the FIC as an unbiased estimator of $E(\Lambda_{S,wcqr}^2)$, which is the limit of the second moment of $n^{1/2}(\hat{\mu}_{S,wcqr} - \mu_{true})$. One remaining problem is that $\boldsymbol{\delta}$ is unknown. Denote the estimator of $\boldsymbol{\delta}$ by the full model as $\hat{\boldsymbol{\delta}}_f$. From Theorem 1, we can see that

$$\hat{\boldsymbol{\delta}}_f = \Pi_u\boldsymbol{\Sigma}^{-1}\mathbf{G}_n + \boldsymbol{\delta} + o_p(1) \xrightarrow{d} N(\boldsymbol{\delta}, \mathbf{K}_u), \tag{11}$$

where $\Pi_u = [0_{p_u\times p_f}, \mathbf{I}_{p_u}]$ and $\mathbf{K}_u = \Pi_u\boldsymbol{\Sigma}^{-1}\Pi_u^T$, which implies that $\hat{\boldsymbol{\delta}}_f\hat{\boldsymbol{\delta}}_f^T - \mathbf{K}_u$ is an unbiased estimator of $\boldsymbol{\delta}\boldsymbol{\delta}^T$. As in Claeskens & Hjort (2003), we define the FIC score of the Sth submodel as

$$\text{FIC}_S = \mu_{\boldsymbol{\beta}}^T\left\{\sigma_{wcqr}^2(\boldsymbol{w})\mathbf{H}_S\boldsymbol{\Sigma}\mathbf{H}_S + (\mathbf{H}_S\boldsymbol{\Sigma} - \mathbf{I})\begin{pmatrix} 0 & 0 \\ 0 & \hat{\boldsymbol{\delta}}_f\hat{\boldsymbol{\delta}}_f^T - \mathbf{K}_u \end{pmatrix}(\mathbf{H}_S\boldsymbol{\Sigma} - \mathbf{I})^T\right\}\mu_{\boldsymbol{\beta}}. \tag{12}$$

By the definition of the FIC score, it can be used to select the best model for a specific parameter of interest $\mu(\boldsymbol{\beta}_{f,0}, n^{-1/2}\boldsymbol{\delta})$. Note that for different $\mu(\boldsymbol{\beta}_{f,0}, n^{-1/2}\boldsymbol{\delta})$ the best model can be different, which is one of appealing properties of the FIC.

3.3 Frequentist model averaging based on WCQR

Choosing the best submodel using the FIC score can be an effective model selection strategy. As an alternative to the model selection, model averaging has the potential to provide estimators that are more

stable and with smaller mean squared error (Zhang & Liang, 2011). While most of the existing works in model averaging are developed in a Bayesian framework, Hjort & Claeskens (2003) proposed a frequentist model averaging method based on the FIC. Let $\mathcal{S}$ be a subset of submodels. Then the averaged estimator of μ can be defined as

$$\hat{\mu}_{wcqr}(\hat{\boldsymbol{\delta}}_f) = \sum_{S\in\mathcal{S}} c(S|\hat{\boldsymbol{\delta}}_f)\hat{\mu}_{S,wcqr}, \tag{13}$$

where the positive weights $c(S|\hat{\boldsymbol{\delta}}_f)$ sum to 1. Model selection can be viewed as a special type of model averaging by assigning 1 to one of the submodels and 0 to others. The following corollary shows the asymptotic property of $\hat{\mu}_{wcqr}(\hat{\boldsymbol{\delta}}_f)$.

COROLLARY 2. *Under Conditions 1–4, if the weight functions $c(S|\hat{\boldsymbol{\delta}}_f)$ have at most a countable number of discontinuities, then*

$$\begin{aligned} n^{1/2}(\hat{\mu}_{wcqr}(\hat{\boldsymbol{\delta}}_f) - \mu_{true}) &= \mu_{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}^{-1}\mathbf{G}_n + \mu_{\boldsymbol{\beta}}^T\{\mathbf{Q}(\hat{\boldsymbol{\delta}}_f) - \mathbf{I}\}\begin{pmatrix} 0 \\ \hat{\boldsymbol{\delta}}_f \end{pmatrix} + o_p(1) \\ &\xrightarrow{d} \Lambda \equiv \mu_{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}^{-1}\mathbf{G}\sigma_{wcqr}(\boldsymbol{w}) + \mu_{\boldsymbol{\beta}}^T\{\mathbf{Q}(\Delta) - \mathbf{I}\}\begin{pmatrix} 0 \\ \Delta \end{pmatrix}, \end{aligned} \tag{14}$$

where $\mathbf{Q}(\cdot) = \sum_{S\in\mathcal{S}} c(S|\cdot)\mathbf{H}_S\boldsymbol{\Sigma}$ *and* $\Delta \sim N(\boldsymbol{\delta}, \mathbf{K}_u)$.

Using (11), by plugging $\boldsymbol{\delta} = \hat{\boldsymbol{\delta}}_{full} - \Pi_u\boldsymbol{\Sigma}^{-1}\mathbf{G}_n$ back to (9), the proof is a direct application of the delta method and is thus omitted here.

It is worth pointing out that Corollary 2 holds for a general choice of weight functions $c(S|\hat{\boldsymbol{\delta}}_f)$, but a good choice of $c(S|\hat{\boldsymbol{\delta}}_f)$ would help reduce the mean squared error of $n^{1/2}(\hat{\mu} - \mu_{true})$. Therefore, even though the total number of submodels 2^{p_u} can be large, we do not have to consider all submodels. Corollary 2 still holds by assigning weight 0 to those left out models. As proposed in Hjort & Claeskens (2003), an example of weight functions is

$$c(S|\hat{\boldsymbol{\delta}}_f) = \exp\left(-\frac{1}{2}\kappa\frac{\text{FIC}_S}{\sigma^2_{wcqr}(\boldsymbol{w})\mu_{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}^{-1}\mu_{\boldsymbol{\beta}}}\right) \Big/ \sum_{S'\in\mathcal{S}} \exp\left(-\frac{1}{2}\kappa\frac{\text{FIC}_{S'}}{\sigma^2_{wcqr}(\boldsymbol{w})\mu_{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}^{-1}\mu_{\boldsymbol{\beta}}}\right), \tag{15}$$

where $\kappa \geq 0$ is a tuning parameter. While κ close to 0 would give uniform weights, large κ gives the FIC-based weights.

Corollary 2 shows that the limiting distribution of $n^{1/2}(\hat{\mu}_{wcqr}(\hat{\boldsymbol{\delta}}_f) - \mu_{true})$ is a nonlinear mixture of normal distributions, which confirms the phenomenon that the post model selection or the model averaging estimator can be highly non-normal in the local misspecification setting. Therefore, naive

confidence intervals based on the asymptotic normal distribution can have poor coverage probabilities. To fix this, in the spirit of Zhang & Liang (2011), define the lower and upper bounds of a $(1-\alpha)100\%$ confidence interval of μ_{true} as

$$\hat{\mu}_{wcqr}(\hat{\boldsymbol{\delta}}_f) - \mu_{\boldsymbol{\beta}}^T\{\mathbf{Q}(\hat{\boldsymbol{\delta}}_f) - \mathbf{I}\}\begin{pmatrix} 0 \\ n^{-1/2}\hat{\boldsymbol{\delta}}_f \end{pmatrix} \pm n^{-1/2} z_{\alpha/2}\hat{\tau}, \tag{16}$$

where $\hat{\tau}^2$ is a consistent estimator of $\mu_{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}^{-1}\mu_{\boldsymbol{\beta}}\sigma^2_{wcqr}(\boldsymbol{w})$ and $z_{\alpha/2}$ is the $(1-\alpha/2)100\%$th standard normal quantile.

3.4 Sparsity function and variance estimation

Since the true residual density is usually unknown, it is necessary to estimate $g(\xi_k) = g(G^{-1}(\tau_k)) = 1/s(\tau_k)$. Here $s(\tau) = dG^{-1}(\tau)/d\tau$ is called the quantile density function. From the large amount of the literature on estimating $s(\tau)$ (e.g. Siddiqui, 1960; Sheather & Maritz, 1983; Koenker & Machado, 1999), we adopt the procedure described in Section 2.5 in Koenker & Machado (1999), where $s(\tau)$ is estimated by

$$\hat{s}_n(\tau) = \frac{\hat{G}_n^{-1}(\tau + h_n) - \hat{G}_n^{-1}(\tau - h_n)}{2h_n}, \tag{17}$$

where $\hat{G}_n^{-1}(\cdot)$ is an estimate of $G^{-1}(\cdot)$ and h_n is a bandwidth. In this paper, the full model is used to estimate (17) and the bandwidth is taken as suggested by Hall & Sheather (1988):

$$h_n = n^{-1/3} z_\alpha^{2/3}\{1.5\psi(\Psi^{-1}(\tau))/(2(\Psi^{-1}(\tau))^2 + 1)\}^{1/3}, \tag{18}$$

where $\Psi(\cdot)$ and $\psi(\cdot)$ are standard normal cumulative distribution function and density function, respectively.

Having obtained an estimator of $s(\tau_k)$'s, say $\hat{s}(\tau_k)$'s, simply plugging them into (4) would lead to under estimation of σ^2_{wcqr} due to the large number of parameters involved in the estimation of Model (1). This phenomenon is well-known for the least square estimator when the variance is estimated using mean of squared residuals. For this reason, we propose to use

$$\hat{\sigma}^2_{wcqr}(\boldsymbol{w}) = \frac{n}{n-p-K}\sum_{k,l=1}^{K} \hat{s}(\tau_k)\hat{s}(\tau_k) w_k w_l\{\min(\tau_k, \tau_l) - \tau_k\tau_l\}, \tag{19}$$

where p is the total number of parameters in the model, K is the number of quantiles.

In general, using K less than 10 quantiles is adequate for WCQR estimator to approach the efficiency of the MLE estimator (Bradic et al., 2011). In this paper, we found in our numerical examples

that there is not much gain in efficiency for K over 5. For a moderate K, the variance estimator defined in (19) appears to provide satisfactory performance in terms of the coverage probability of the confidence interval described in (16) as illustrated in our simulation studies.

4 Simulation study

In this section, we study the finite sample performance of the proposed procedure under several error distributions. Artificial data were generated from the model

$$Y_i = f_1(X_{1,i}) + f_2(X_{2,i}) + \mathbf{Z}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the true parameter $\boldsymbol{\beta} = \{3.0, 1.5, n^{-1/2}c_0(2, 1, 0)\}^T$ with c_0 taking values $1, 2, 3$. Functions $f_1(x) = \sin(2\pi x)$ and $f_2(x) = 5x^4 + 3x^2 - 2$ are taken from Zhang & Liang (2011). Covariates $X_{1,i}$ and $X_{2,i}$ are independently generated from the uniform distribution on $[0, 1]$ and $Z_{1,i}, \ldots, Z_{5,i}$ are jointly normally distributed with mean 0 and variance 1. The correlation between $Z_{i,j}$ and $Z_{i,l}$ is $0.5^{|j-l|}$. Z_1 and Z_2 are always present in all submodels while other three Z_3, Z_4, Z_5 may or may not present, which gives $2^3 = 8$ models to select from or to average across. The focused parameters are (i) $\mu_1 = \boldsymbol{\beta}_1$, (ii) $\mu_2 = \boldsymbol{\beta}_2$ and (iii) $\mu_3 = 0.8\boldsymbol{\beta}_1 + 0.05\boldsymbol{\beta}_2 - 0.5\boldsymbol{\beta}_3 + 0.1\boldsymbol{\beta}_4 + 0.09\boldsymbol{\beta}_5$.

The cubic B-spline basis functions were used to approximate the two nonparametric functions, where J_n equally spaced internal knots in $[0, 1]$ were used for both functions. As in Zhang & Liang (2011), the number of internal knots J_n was selected in the range $[2/3N_r, 4/3N_r]$, where $N_r = [n^{1/5.5}] + 1$ with $[x]$ representing the integer part of x. The optimal knot number J_n^{opt} was selected using the 5-fold cross validation based on the median absolute prediction error (MAPE) defined as the median of $\{|Y_i - \hat{Y}_i|, i = 1, \ldots, n_v\}$, where $\hat{Y}_i$ is the fitted value for case i and n_v is the size of the validation sample.

Six different residual distributions were considered in the simulation: the normal $N(0, 3)$, the t-distribution with 3 degrees of freedom (t_3), a scale mixture of normal distribution $0.1N(0, 25) + 0.9N(0, 1)$ (MN_s), a location mixture of normal distribution $0.1N(3, 2) + 0.9N(-1/3, 2)$ (MN_l), a skewed normal distribution $SN(-2.19, \sqrt{8}, 4)$ (Azzalini & Capitanio, 2003) (SN) and a skewed t distribution (Azzalini & Capitanio, 2003) with 4 degrees of freedom $ST(-1.68, \sqrt{3}, 4, 4)$ (ST). All parameters were set so that the residual distribution has a mean 0 and a variance around 3 for fair comparisons. Random numbers from skewed distributions were generated using the R package SN.

Three FIC-based estimators were studied: the ordinary least square estimator (F-OLS), the weighted composite quantile regression estimator (F-WCQR) and the maximum likelihood estimator (F-MLE).

The FIC weights across different submodels are as defined in (15) with $\kappa = 2$. The F-OLS estimator can be viewed as an application of the results of Zhang & Liang (2011) to the estimation of Model (1). For comparisons, we also computed the mean squared error of the focused parameter estimators and the empirical coverage probability of confidence intervals based on the best model selected by the AIC and BIC using the true likelihood function. For the F-WCQR estimator, $\{1/(K+1), \ldots, K/(K+1)\}$ quantiles were used with $K = 5$. We also tried several other choices of K such as $K = 7$ and $K = 10$ but found that the estimation efficiency does not increase much for $K > 5$. The variance of the F-WCQR estimator used to construct confidence interval defined in (16) was estimated using (19) with optimal weights defined in (8). As suggested by a referee, we numerically examined the sensitivity of our approach with respect to the choice of the bandwidth in (17) by manually increasing or decreasing the suggested bandwidth (18) by 10%. The resulting estimators are denoted as F-WCQR$^+$ and F-WCQR$^-$, respectively. The variance of the F-OLS estimator was estimated using the usual unbiased variance estimator based on residuals. For all MLE based confidence intervals, the variance was obtained using the true residual likelihood function.

Tables 1–3 summarize the empirical relative efficiency (RE) of each estimator defined as the ratio of the mean squared error of F-MLE to that of each estimator. In addition, we also present the empirical coverage probabilities (CPs) of a nominal 95% confidence interval. All results are based on 1000 simulation runs. The AIC and BIC based confidence intervals are often overly optimistic, which is reflected in the lower coverage probability of μ_2 and μ_3 than the nominal 95% level. The smallest CPs of AIC and BIC are only 84.4% and 76.5%. On the contrary, all three FIC-based confidence intervals (F-OLS, F-WCQR, F-MLE) have coverage probabilities much closer to 95%, with the smallest CPs as 94.4%, 93.5%, and 91.8%, respectively.

Except for the normal case, the F-WCQR estimator appears to always outperform the F-OLS estimator and is at least 80% as efficient as the F-MLE estimator. Using the least square loss function can be viewed as the case where we mis-specify the residual distribution as the normal distribution when it is not. Therefore, the lower efficiency of the F-OLS estimator also indicates that if the likelihood function is misspecified, the resulting F-MLE estimator may have poor performance. In this sense, the F-WCQR estimator is safer in that it does not make distributional assumptions on the residuals and thus does not suffer from the consequence of mis-specification of the likelihood.

Another observation is that even though the AIC and BIC estimators make use of the knowledge of the true error distribution, their relative efficiency can still be lower than the F-WCQR estimators, sometimes even lower than the F-OLS estimator. This indicates that the uncertainty in the model selection

step could result in loss of efficiency in the subsequent estimation and prediction. On the other hand, the FIC-based model averaging estimators are more stable and efficient because when they averages across a set of submodels the weights of submodels are automatically adjusted for a specific focused parameter. Note that in some cases, the AIC and BIC estimators are more efficient than the F-MLE estimator, which confirms the findings in Yuan and Yang (2005) and Zhang et al. (2012) that selection by AIC and BIC can perform better than FMA in terms of the prediction accuracy.

Finally, we did not observe any appreciable differences among F-WCQR, F-WCQR$^+$ and F-WCQR$^-$ estimators, in terms of either the coverage probabilities or the relative efficiency. This is an indication that our approach is robust to moderate fluctuations in the bandwidth used in (17).

5 Real data analysis

In this section, we employ the proposed F-WCQR procedure to analyze the beta-carotene level data set. This data set is publicly available at `lib.stat.cmu.edu/datasets/Plasma_Retinol`. We used a sample of 273 male subjects in this data set to study the relationships between the plasma beta-carotene level (numeric) and covariates including age (numeric), smoking status (Never, Former, Current), quetelet index (numeric), vitamin use (Fairly often, Not often, No), calories (numeric), fat (numeric), fiber (numeric), number of alcohol drinks (numeric), cholesterol (numeric), and dietary beta-carotene (numeric). For a more detailed description of the study, please refer to Nierenberg et al. (1989).

This data set was analyzed in Kai et al. (2011). For easy comparison, we adopt the same parameterizations as theirs. Other than the categorical variables "smoking status" and "vitamin use", all covariates are standardized. The "dietary beta-carotene" is used as the nonparametric component in (1). There are in total 11 covariates in the parametric part, including 4 dummy variables for those two categorical variables. The first 200 observations were used as the training data set to fit the model and to conduct model averaging and the remaining 73 observations were used to evaluate the performance of model in terms of the MAPE $=$ median$\{|\hat{Y}_i - Y_i|, i = 1, \ldots, 73\}$. To use F-WCQR for prediction, the intercept term a_0 in (1) was estimated by $\hat{a}_0 = K^{-1}\sum_{k=1}^{K}\hat{a}_k$ where $\hat{a}_k$'s are minimizer of (2). The same estimator was used in Kai et al. (2011).

For the F-WCQR estimator, $\{1/(K+1), \ldots, K/(K+1)\}$ quantiles were used with $K = 9$ and $J_n = 4$ internal knots was selected by the 5-fold cross-validation for the nonparametric part. Further cross validation showed that the prediction performance does not improve much using $K > 9$. A preliminary examination of the full model using WCQR indicated that only two covariates are statistically significant:

"quetelet index" and "fiber". Therefore, we assume these two covariates are always in the model. The rest 9 variables may or may not be in the model, which gives $2^9 = 512$ submodels in total. However, we do not need to estimate all these submodels. Instead, we randomly select 32 out of these 512 models to do the model averaging. The results are summarized in Table 4, where $\hat{\boldsymbol{\beta}}^{FIC}_{WCQR}$ stands for the F-WCQR estimator. For comparisons, the penalized least square estimator ($\hat{\boldsymbol{\beta}}^{OSE}_{LS}$) and the penalized equal weights composite quantile regression estimator ($\hat{\boldsymbol{\beta}}^{OSE}_{CQR}$) were taken from the statistical analysis on the same data set in Kai et al. (2011). Observe that only confidence intervals corresponding to the F-WCQR estimates of "quetelet index" and "fiber" do not contain 0, which indicates that they might be the only two significant variables.

To see the effect of using only randomly selected 32 out of 512 submodels, we repeated the F-WCQR procedure using 100 different sets of submodels, each of which is of size 32. The MAPEs for these 100 F-WCQR estimators have a mean of 48.10 and SD of 1.61, indicating that the variation due to the use of a random subset of submodels is relatively small and can be considered as negligible. This empirical evidence further indicates that the F-WCQR procedure performs better on prediction than the two methods proposed in Kai et al. (2011) in this application.

6 Concluding remarks

In this paper, We study the focused information criterion (FIC) and frequentist model averaging (FMA) based on the weighted composite quantile regression estimator under the context of the additive partly linear regression model. With properly chosen weights, the FIC-based WCQR estimators not only are robust to outliers and nonnormal residuals but also can achieve efficiency close to the maximum likelihood estimator (MLE), without assuming the true error distribution. We used the polynomial spline to approximate the nonlinear function since it is more computationally efficient than the kernel methods. We studied the asymptotic distribution of the averaged WCQR-estimator under the local mis-specification frame work and used it to construct valid confidence interval that takes into account the model selection uncertainty.

This work is of importance because the existing work of FIC and FMA has been largely focused on the likelihood approach and when the likelihood information is absent the effectiveness of using the FIC approach can be compromised. This is undesirable because the primary goal of FIC is to improve the estimation efficiency of the focused parameter. By proposing the FIC-based WCQR estimator, we are able to develop a not only flexible but also highly efficient estimation approach that can maintain the

advantage of the FIC approach when the likelihood information is missing for the additive partly linear regression model. Our method can also be applied to the linear model without additional effort.

Acknowledgment This research was partially supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). Huang's research was also partially supported by NSF (DMS-0907170, DMS-1007618, DMS-1208952), NCI (CA57030). Part of this work was carried out while Wang was a visiting Professor at KAUST.

References

Azzalini, A., and Capitanio, A. (2003), "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution," *J. R. Stat. Soc. Ser. B*, 65, 367–389.

Behl, P., Claeskens, G., and Dette, H. (2013), "Focused model selection in quantile regression," *Statist. Sinica*, to appear.

Bradic, J., Fan, J., and Wang, W. (2011), "Penalized Composite quasi-likelihood for ultrahigh dimensional variable selection," *J. R. Stat. Soc. Ser. B*, 73, 325–349.

Claeskens, G., and Carroll, R. (2007), "An asymptotic theory for model selection inference in general semiparametric problems," *Biometrika*, 94, 1–17.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2006), "Variable selection for logistic regression using a prediction-focused information criterion," *Biometrics*, 62, 972–979.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007), "Prediction-focused model selection for autoregressive models," *Aus. J. Statist.*, 49, 359–379.

Claeskens, G., and Hjort, N. (2003), "The focused information criterion (with discussion)," *J. Amer. Statist. Assoc.*, 98, 900–945.

Hall, P., and Sheather, S. (1988), "On the distribution of a studentized quantile," *J. R. Stat. Soc. Ser. B*, 50, 381–391.

He, X., and Shi, P. (1994), "Convergence rate of B-spline estimators of nonparametric conditional quantile functions," *J. Nonparam. Statist.*, 3, 299–308.

He, X., and Shi, P. (1996), "Bivariate tensor-product B-Splines in a partly linear model," *J. Mult. Anal.*, 58, 162–181.

Hjort, N. L., and Claeskens, G. (2003), "Frequentist model average estimators," *J. Amer. Statist. Assoc.*, 98, 879–899.

Hjort, N. L., and Claeskens, G. (2006), "Focused information criteria and model averaging for Cox's hazard regression model," *J. Amer. Statist. Assoc.*, 101, 1449–1464.

Huang, J. Z., Zhang, L., and Zhou, L. (2007), "Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines," *Scand. J. Statist.*, 34, 451–477.

Kabaila, P., and Leeb, H. (2006), "On the large-sample minimal coverage probability of confidence intervals after model selection," *J. Amer. Statist. Assoc.*, 101, 619–629.

Kai, B., Li, R., and Zou, H. (2011), "New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models," *Ann. Statist.*, 39, 305–332.

Koenker, R., and Bassett, G. (1978), "Regression quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and Machado, J. (1999), "Goodness of fit and related inference processes for quantile regression," *J. Amer. Statist. Assoc.*, 94, 1296–1310.

Leeb, H., and Potscher, B. M. (2005), "Model selection and inference: facts and fiction," *Economet. Theory*, 21, 21–59.

Leeb, H., and Potscher, B. M. (2006), "Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results," *Economet. Theory*, 22, 69–97.

Liu, X., Wang, L., and Liang, H. (2011), "Estimation and variable selection for semiparametric additive partial linear model," *Statist. Sinica*, 21, 1225–1248.

Nierenberg, D., Stukel, T., Baron, J., Dain, B., and Greenberg, E. (1989), "Determinants of plasma levels of Beta-carotene and Retinol," *American Journal of Epidemiology*, 130, 511–521.

Schumaker, L. L. (1981), *Spline Functions: Basic Theory*, New York: Wiley.

Sheather, S., and Maritz, J. (1983), "An estimate of the asymptotic standard error of the sample median," *Aus. J. Statist.*, 25, 109–122.

Siddiqui, M. (1960), "Distribution of quantiles from a bivariate population," *Journal of Research of the National Bureau of Standards*, 64B, 145–150.

Yuan, Z., and Yang, Y. (2005), "Combining linear regression models: When and how?" *J. Amer. Statist. Assoc.*, 100, 1202–1214.

Zhang, X., and Liang, H. (2011), "Focused information criterion and model averaging for generalized additive partial linear models," *Ann. Statist.*, 39, 174–200.

Zhang, X., Wan, A. T. K., and Zhou, Z. (2012), "Focused information criterion and model averaging in a Tobit model with a non-zero threshold," *J. Business and Economic Statist.*, 30, 132–142.

Zou, H., and Yuan, M. (2008), "Composite quantile regression and the oracle model selection theory," *Ann. Statist.*, 36, 1108–1126.

Ganggang Xu, Department of Statistics, Texas A&M University, College Station, TX 77483, USA.
Email: gang@stat.tamu.edu

Appendix: Technical proofs

For simplicity, we use equally spaced knots. In this case, since each component function $f_j \in \mathcal{H}_r$, there exists a function $\tilde{f}_j \in \mathcal{G}$ such that $||\tilde{f}_j - f_j||_\infty \leq C_{1,j} J_n^{-r}$, where $C_{1,j}$ is some positive constant. The proof can be found in, for example, Schumaker (1981). For $f_0(\mathbf{x}) = \sum_{j=1}^m f(x_j)$, a simple application of triangle inequality yields that there exists a coefficient vector $\boldsymbol{\gamma}_0$ such that $\tilde{f}_0(\mathbf{X}) = \boldsymbol{\gamma}_0^T \boldsymbol{b}(\mathbf{X})$ satisfying

$$||\tilde{f}_0 - f_0||_\infty \leq C_2 J_n^{-r}. \tag{A.1}$$

First we introduce some notation. Define quantities

$$\mathbf{H}_{1n}^2 = n\Pi_s \widehat{\boldsymbol{\Sigma}}_n \Pi_s^T, \quad \mathbf{H}_{2n}^2 = J_n \mathbf{B}_n^T \mathbf{B}_n, \quad \mathbf{H}_n^2 = diag\{\mathbf{H}_{1n}^2, \mathbf{H}_{2n}^2\},$$

$$\boldsymbol{\delta}_n = \begin{pmatrix} 0 \\ n^{-1/2}\boldsymbol{\delta} \end{pmatrix}, \quad \hat{\boldsymbol{\theta}}_n = \begin{pmatrix} \hat{\boldsymbol{\theta}}_{1n} \\ \hat{\boldsymbol{\theta}}_{2n} \end{pmatrix}, \quad \hat{\boldsymbol{\theta}}_{1n} = \mathbf{H}_{1n} \left\{ \hat{\boldsymbol{\beta}}_{S,wcqr} - \begin{pmatrix} \boldsymbol{\beta}_{c,0} \\ 0 \end{pmatrix} \right\},$$

$$\hat{\boldsymbol{\theta}}_{2n} = J_n^{-1/2} \mathbf{H}_{2n} \{ (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + (\mathbf{B}_n^T \mathbf{B}_n)^- \sum_{k=1}^n \boldsymbol{b}(\mathbf{X}_k) \mathbf{Z}_k^T (\Pi_s^T \mathbf{H}_{1n}^- \hat{\boldsymbol{\theta}}_{1n}) \},$$

$$\mathbf{T}_i = \begin{pmatrix} \mathbf{T}_{1i} \\ \mathbf{T}_{2i} \end{pmatrix}, \quad \mathbf{T}_{1i} = \mathbf{H}_{1n}^- \Pi_s \mathbf{Z}_i^*, \quad \mathbf{T}_{2i} = \mathbf{H}_{2n}^- \boldsymbol{b}(\mathbf{X}_i) J_n^{1/2}.$$

Further, let $R_{ni} = \boldsymbol{b}(\mathbf{X}_i)^T \boldsymbol{\gamma}_0 - f_0(\mathbf{X}_i)$ be the bias of the spline approximation and $\hat{\eta}_k = n^{1/2}(\hat{a}_k - \xi_k)$, $k = 1, \ldots, K$. Then the loss of the estimators $(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}_{S,wcqr})$ becomes

$$\ell(\hat{\boldsymbol{a}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}_{S,wcqr}) = \sum_{k=1}^K \sum_{i=1}^n w_k \rho_{\tau_k} (\varepsilon_i - \xi_k - n^{-1/2}\hat{\eta}_k - \mathbf{T}_{1i}^T \hat{\boldsymbol{\theta}}_{1n} - \mathbf{T}_{2i}^T \hat{\boldsymbol{\theta}}_{2n} + \mathbf{Z}_i^T \boldsymbol{\delta}_n - R_{ni}). \tag{A.2}$$

Recall that $\rho'_{\tau_k}(s) = \phi_{\tau_k}(s) = \tau_k - I(s<0)$. Let

$$\begin{aligned} U_{i,k}(\eta_k, \boldsymbol{\theta}) &= \rho_{\tau_k}(\varepsilon_i - \xi_k - n^{-1/2}\eta_k - \mathbf{T}_i^T\boldsymbol{\theta} + \mathbf{Z}_i^T\boldsymbol{\delta}_n - R_{ni}) \\ &\quad - \rho_{\tau_k}(\varepsilon_i - \xi_k + \mathbf{Z}_i^T\boldsymbol{\delta}_n - R_{ni}) + (n^{-1/2}\eta_k + \mathbf{T}_i^T\boldsymbol{\theta})\phi_{\tau_k}(\varepsilon_i - \xi_k), \\ V_{i,k}(\eta_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \rho_{\tau_k}(\varepsilon_i - \xi_k - n^{-1/2}\eta_k - \mathbf{T}_{1i}^T\boldsymbol{\theta}_1 - \mathbf{T}_{2i}^T\boldsymbol{\theta}_2 + \mathbf{Z}_i^T\boldsymbol{\delta}_n - R_{ni}) \\ &\quad - \rho_{\tau_k}(\varepsilon_i - \xi_k - n^{-1/2}\eta_k - \mathbf{T}_{2i}^T\boldsymbol{\theta}_2 + \mathbf{Z}_i^T\boldsymbol{\delta}_n - R_{ni}) + \mathbf{T}_{1i}^T\boldsymbol{\theta}_1\phi_{\tau_k}(\varepsilon_i - \xi_k). \end{aligned}$$

Through out the proofs, the expectation $E(\cdot)$ is taken with respect to ε unless otherwise stated.

LEMMA A.1. *Under Conditions 1–4, for any sequence $\{L_n\}$ satisfying $1 \le L_n \le J_n^{\boldsymbol{\delta}_0/10}$ for some $0 < \boldsymbol{\delta}_0 < r$ such that $J_n^{2+\boldsymbol{\delta}_0} = o(n)$, one has*

$$\sup_{(||\boldsymbol{\theta}||^2+||\boldsymbol{\eta}||^2)^{1/2}\le L_n J_n^{1/2}} \left| J_n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n} w_k(U_{i,k}(\eta_k,\boldsymbol{\theta}) - E\{U_{i,k}(\eta_k,\boldsymbol{\theta})\})\right| = o_p(1), \tag{A.3}$$

and for any $M_1 > 0$, $M_2 > 0$,

$$\begin{aligned} \sup_{||\boldsymbol{\theta}_1||\le M_1;(||\boldsymbol{\theta}_2||^2+||\boldsymbol{\eta}||^2)^{1/2}\le M_2 J_n^{1/2}} \Bigg| J_n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n} & w_k(V_{i,k}(\eta_k,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \\ & - E\{V_{i,k}(\eta_k,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\})\Bigg| = o_p(1), \end{aligned} \tag{A.4}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)^T$.

The proof is given in the Supplementary Material.

LEMMA A.2. *Under Condtions 1–4, for any sequence $\{L_n\}$ satisfying $1 \le L_n \le J_n^{\boldsymbol{\delta}_0/10}$ for some $0 < \boldsymbol{\delta}_0 < r$ such that $J_n^{2+\boldsymbol{\delta}_0} = o(n)$, one has that as $n \to \infty$, for any constant $M > 0$,*

$$\begin{aligned} P\Bigg(\inf_{||\boldsymbol{\theta}||^2+||\boldsymbol{\eta}||^2=1} J_n^{-1}\Bigg| \sum_{k=1}^{K}\sum_{i=1}^{n} & w_k[E\{U_{i,k}(L_nJ_n^{1/2}\eta_k, L_nJ_n^{1/2}\boldsymbol{\theta})\} \\ & - L_nJ_n^{1/2}(n^{-1/2}\eta_k + \mathbf{T}_i^T\boldsymbol{\theta})\phi_{\tau_k}(\varepsilon_i - \xi_k)]\Bigg| > M\Bigg) \to 1. \end{aligned} \tag{A.5}$$

The proof is given in the Supplementary Material.

Proof of Lemma 1. Let $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\theta}})^T = (\hat{\eta}_1, \ldots, \hat{\eta}_K, \hat{\boldsymbol{\theta}})^T$ be the minimizer of (A.2). To prove Lemma 1, it suffices to show that $||\hat{\boldsymbol{\eta}}|| = O_p(J_n^{1/2})$ and $||\hat{\boldsymbol{\theta}}|| = O_p(J_n^{1/2})$. Using (A.3) and (A.5), we have that for any $\epsilon > 0$, there exists a constant $L_\epsilon \in (0,1)$ such that for all n

$$\begin{aligned} P\Bigg\{ \inf_{(||\boldsymbol{\theta}||^2+||\boldsymbol{\eta}||^2)^{1/2} > L_\epsilon J_n^{1/2}} & \sum_{k=1}^{K}\sum_{i=1}^{n} w_k\rho_{\tau_k}(\varepsilon_i - \xi_k - n^{-1/2}\eta_k - \mathbf{T}_i^T\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\delta}_n - R_{ni}) \\ & > \sum_{k=1}^{K}\sum_{i=1}^{n} w_k\rho_{\tau_k}(\varepsilon_i - \xi_k + \mathbf{Z}_i\boldsymbol{\delta}_n - R_{ni})\Bigg\} > 1 - \epsilon. \end{aligned}$$

On the other hand, $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\theta}})$ by definition is the minimizer of the left part of the above inequality. Hence we conclude that $P(||\hat{\boldsymbol{\eta}}||^2 + ||\hat{\boldsymbol{\theta}}||^2 \leq L_\epsilon^2 J_n) > 1 - \epsilon$, which implies that $||\hat{\boldsymbol{\eta}}|| = O_p(J_n^{1/2})$ and $||\hat{\boldsymbol{\theta}}|| = O_p(J_n^{1/2})$. The proof of Lemma 1 is completed. □

LEMMA A.3. *Under Conditions 1–4, for any $L > 0$ and $M > 0$,*

$$\sum_{k=1}^{K}\sum_{i=1}^{n} w_k E\{V_{i,k}(\eta_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\} = \frac{b}{2}\boldsymbol{\theta}_1^T\boldsymbol{\theta}_1 - bn^{1/2}\boldsymbol{\theta}_1^T\mathbf{H}_{1n}^{-}\Pi_s\hat{\boldsymbol{\Sigma}}\begin{pmatrix} 0 \\ \boldsymbol{\delta} \end{pmatrix} + r_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2),$$

where $b = \sum_{k=1}^{K} w_k g(\xi_k)$ and $\sup_{||\boldsymbol{\theta}_1|| \leq M, ||\boldsymbol{\theta}_2||^2 + ||\boldsymbol{\eta}||^2 \leq L^2 J_n} |r_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)| = o_p(1)$.

The proof is given in the Supplementary Material.

Proof of Theorem 1. Recall that $b = \sum_{k=1}^{K} w_k g(\xi_k)$ and define quantities

$$\tilde{\boldsymbol{\theta}}_{1n} = b^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n} w_k \phi_{\tau_k}(\varepsilon_i - \xi_k)\mathbf{T}_{1i} + n^{1/2}\mathbf{H}_{1n}^{-}\Pi_s\hat{\boldsymbol{\Sigma}}\begin{pmatrix} 0 \\ \boldsymbol{\delta} \end{pmatrix},$$

$\mathbf{D} = n^{1/2}\mathbf{H}_{1n}^{-}\Pi_s\boldsymbol{\Sigma}\begin{pmatrix} 0 \\ \boldsymbol{\delta} \end{pmatrix}$, and for $i = 1, \ldots, n$, $k = 1, \ldots, K$,

$$W_{i,k}(\eta_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \rho_{\tau_k}(\varepsilon_i - \xi_k - n^{-1/2}\eta_k - \mathbf{T}_{1i}^T\boldsymbol{\theta}_1 - \mathbf{T}_{2i}^T\boldsymbol{\theta}_2 + \mathbf{Z}_i^T\boldsymbol{\delta}_n - R_{ni}).$$

Denote $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\theta}}_{1n}^T, \hat{\boldsymbol{\theta}}_{2n}^T)^T$ as the minimizer of (A.2). For any $L > 0$ and $\Delta > 0$, by an application of the triangle inequality, we have

$$\begin{aligned}
&\sup_{||\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\theta}}_{1n}|| = \Delta, \boldsymbol{\theta}_1 \in R^p} I(||\tilde{\boldsymbol{\theta}}_{1n}|| \leq L, ||\hat{\boldsymbol{\theta}}_{2n}||^2 + ||\hat{\boldsymbol{\eta}}||^2 \leq L^2 J_n) \\
&\qquad \times \left| \sum_{k=1}^{K}\sum_{i=1}^{n} w_k\{W_{i,k}(\hat{\eta}_k, \boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{2n}) - W_{i,k}(\hat{\eta}_k, \tilde{\boldsymbol{\theta}}_{1n}, \hat{\boldsymbol{\theta}}_{2n}) - \frac{1}{2}g(\xi_k)\Delta^2\} \right| \\
&\leq 2 \sup_{||\boldsymbol{\theta}_1|| \leq L + \Delta, ||\hat{\boldsymbol{\theta}}_{2n}||^2 + ||\hat{\boldsymbol{\eta}}||^2 \leq L^2 J_n} \left| \sum_{k=1}^{K}\sum_{i=1}^{n} w_k\{W_{i,k}(\hat{\eta}_k, \boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{2n}) - W_{i,k}(\hat{\eta}_k, \mathbf{0}, \hat{\boldsymbol{\theta}}_{2n}) \right. \\
&\qquad \left. + \mathbf{T}_i^T\boldsymbol{\theta}_1\phi_{\tau_k}(\varepsilon_i - \xi_k) - \frac{g(\xi_k)\boldsymbol{\theta}_1^T(\boldsymbol{\theta}_1 - 2\mathbf{D})}{2}\} \right| \\
&\qquad = o_p(1).
\end{aligned} \tag{A.6}$$

The second last inequality is obtained by plugging in $\Delta^2 = (\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\theta}}_{1n})^T(\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\theta}}_{1n})$ and the last equality follows from (A.4) in Lemma A.1 and Lemma A.3 by noting that $\frac{1}{2}b\boldsymbol{\theta}_1^T(\boldsymbol{\theta}_1 - 2\mathbf{D}) + r_n(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is

exactly the right-hand side of the expression in Lemma A.3. Letting L go to ∞, by Lemma 1 we have that $P(||\tilde{\boldsymbol{\theta}}_{1n}|| \leq L, ||\hat{\boldsymbol{\theta}}_{2n}||^2 + ||\hat{\boldsymbol{\eta}}||^2 \leq L^2 J_n) \to 1$. Hence Equation (A.6) implies that

$$\sup_{||\boldsymbol{\theta}_1 - \tilde{\boldsymbol{\theta}}_{1n}|| = \Delta, \boldsymbol{\theta}_1 \in R^p} \left| \sum_{k=1}^{K} \sum_{i=1}^{n} w_k \{ W_{i,k}(\hat{\eta}_k, \boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{2n}) - W_{i,k}(\hat{\eta}_k, \tilde{\boldsymbol{\theta}}_{1n}, \hat{\boldsymbol{\theta}}_{2n}) - \frac{1}{2} g(\xi_k) \Delta^2 \} \right| \xrightarrow{p} 0$$

for any $\Delta > 0$. Note that $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\theta}}_{1n}, \hat{\boldsymbol{\theta}}_{2n})$ is the minimizer of

$$\sum_{k=1}^{K} \sum_{i=1}^{n} w_k W_{i,k}(\eta_k, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

Following arguments similar to that at the end of the proof of Theorem 3.2 in He & Shi (1996), we can show that $\hat{\boldsymbol{\theta}}_{1n} = \tilde{\boldsymbol{\theta}}_{1n} + o_p(1)$. Finally, Equation (3) follows from the Slutsky Theorem, the Central Limit Theorem and the definitions of $\hat{\boldsymbol{\theta}}_{1n}$ and $\tilde{\boldsymbol{\theta}}_{1n}$. □

Table 1: Coverage probabilities and relative efficiency with sample size $n = 200$

c_0	Method	Normal						t_3					
		CPs			RE			CPs			RE		
		μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
1	F-OLS	94.4	95.5	95.2	100.0	100.0	100.0	94.7	96.0	95.2	52.8	58.1	60.1
	F-WCQR	94.6	95.8	94.8	88.9	90.1	89.6	95.9	96.0	95.8	92.1	93.8	94.9
	F-WCQR$^+$	95.0	96.6	95.2	88.8	90.5	90.0	95.7	96.2	96.0	91.7	94.8	95.3
	F-WCQR$^-$	94.7	95.3	93.8	89.3	90.0	90.4	95.0	95.4	94.8	91.7	92.0	94.3
	F-MLE	94.5	95.5	94.8	100.0	100.0	100.0	94.5	94.2	94.0	100.0	100.0	100.0
	AIC	94.1	92.7	88.9	99.3	87.5	78.4	94.3	92.4	87.8	98.2	86.3	77.2
	BIC	94.2	90.5	89.1	99.2	82.9	82.3	94.5	89.3	85.3	98.4	75.9	73.7
2	F-OLS	94.4	95.5	95.2	100.0	100.2	100.0	94.7	96.0	95.2	53.3	59.7	64.1
	F-WCQR	94.7	95.8	94.7	89.6	92.9	91.2	95.3	96.1	95.4	93.3	94.1	95.1
	F-WCQR$^+$	95.1	96.5	95.2	89.3	92.6	91.5	95.6	96.2	96.2	92.9	94.4	95.8
	F-WCQR$^-$	94.6	95.7	94.0	89.6	91.9	91.6	94.9	95.2	95.1	93.4	92.2	95.3
	F-MLE	94.5	95.5	94.8	100.0	100.0	100.0	94.5	94.3	94.1	100.0	100.0	100.0
	AIC	94.4	92.7	85.7	99.3	93.7	80.2	94.1	92.1	86.8	99.0	98.1	87.0
	BIC	94.3	87.7	78.9	98.6	75.3	65.2	93.8	88.3	79.5	97.5	86.5	68.4
3	F-OLS	94.4	95.5	95.2	100.0	100.3	100.0	94.7	96.0	95.2	53.9	58.2	62.4
	F-WCQR	94.8	95.8	94.9	90.7	92.8	91.2	95.6	95.9	95.5	93.7	92.5	94.3
	F-WCQR$^+$	95.0	96.7	95.3	90.3	92.7	91.4	95.8	96.2	96.3	93.3	92.5	94.7
	F-WCQR$^-$	94.6	95.8	94.1	90.5	91.8	91.5	95.2	95.5	95.2	94.2	90.9	95.2
	F-MLE	94.5	95.5	94.8	100.0	100.0	100.0	94.4	94.2	94.0	100.0	100.0	100.0
	AIC	94.5	93.9	89.0	99.7	106.9	96.3	94.0	93.9	91.6	100.3	110.1	109.1
	BIC	94.2	90.4	79.0	99.0	89.7	69.0	94.1	91.7	83.8	99.6	97.7	78.0

Table 2: Coverage probabilities and relative efficiency with sample size $n = 200$

c_0	Method	MN_s CPs			MN_s RE			MN_l CPs			MN_l RE		
		μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
1	F-OLS	94.7	94.2	95.6	35.7	37.6	42.7	94.2	94.6	93.0	86.6	89.1	90.0
	F-WCQR	95.8	95.9	95.0	83.9	82.1	86.2	94.4	93.9	94.0	88.3	87.6	90.4
	F-WCQR$^+$	96.6	96.2	95.9	83.8	81.1	87.3	94.8	94.6	94.6	88.4	87.6	91.1
	F-WCQR$^-$	95.7	95.7	95.0	84.6	83.4	86.7	93.7	92.4	94.1	87.9	87.0	91.1
	F-MLE	95.3	95.3	94.4	100.0	100.0	100.0	94.1	94.1	93.8	100.0	100.0	100.0
	AIC	95.0	93.4	86.4	97.8	91.1	75.0	94.3	90.5	88.9	98.3	89.5	77.2
	BIC	94.8	89.2	82.3	98.2	75.4	68.3	94.2	88.4	87.9	98.3	82.5	80.6
2	F-OLS	94.7	94.2	95.6	35.8	39.4	46.0	94.2	94.6	93.0	86.9	89.2	91.6
	F-WCQR	96.3	95.9	94.8	84.7	82.5	87.4	94.0	94.3	94.2	88.8	88.4	92.3
	F-WCQR$^+$	96.6	96.3	95.8	83.8	81.6	88.3	94.6	94.5	94.7	88.9	88.4	93.0
	F-WCQR$^-$	95.8	95.7	94.6	84.9	83.7	87.5	93.8	92.8	94.0	88.6	87.9	93.4
	F-MLE	95.3	95.2	94.4	100.0	100.0	100.0	93.9	94.1	93.7	100.0	100.0	100.0
	AIC	95.1	93.6	87.6	96.9	104.5	92.3	94.5	91.1	85.6	99.0	95.6	80.5
	BIC	94.7	91.0	79.0	96.9	93.5	68.7	94.1	87.0	78.6	97.8	80.5	64.8
3	F-OLS	94.7	94.2	95.6	36.2	36.7	42.4	94.2	94.6	93.0	87.2	88.5	91.4
	F-WCQR	96.6	96.1	95.0	84.7	82.5	87.4	94.4	94.8	94.2	89.6	88.2	92.4
	F-WCQR$^+$	96.9	96.4	95.5	84.7	79.3	85.8	95.0	94.5	94.7	89.6	87.8	92.5
	F-WCQR$^-$	95.4	95.6	95.0	85.8	81.4	85.6	93.7	93.0	93.8	89.4	87.9	93.2
	F-MLE	95.2	95.2	94.3	100.0	100.0	100.0	94.0	93.9	93.6	100.0	100.0	100.0
	AIC	95.2	94.7	92.4	99.6	110.3	112.2	94.1	91.9	88.9	99.7	107.7	94.0
	BIC	95.1	93.1	86.3	97.7	102.6	84.4	94.0	88.9	79.8	98.9	94.4	69.3

Table 3: Coverage probabilities and relative efficiency with sample size $n = 200$

c_0	Method	SN						ST					
		CPs			RE			CPs			RE		
		μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3	μ_1	μ_2	μ_3
1	F-OLS	94.4	94.2	94.4	72.5	70.9	75.4	95.1	95.0	95.4	35.9	42.5	42.8
	F-WCQR	94.7	93.5	94.2	83.6	84.6	83.1	94.6	94.4	94.1	85.5	86.0	88.3
	F-WCQR$^+$	95.2	94.0	94.6	83.9	84.1	84.7	95.1	94.5	94.9	86.9	86.7	89.8
	F-WCQR$^-$	93.9	93.0	93.6	82.8	83.9	81.7	94.4	93.5	93.9	85.9	86.5	87.8
	F-MLE	94.0	93.3	94.4	100.0	100.0	100.0	93.3	92.3	92.2	100.0	100.0	100.0
	AIC	93.9	90.6	87.1	97.4	86.9	74.6	93.5	89.3	85.5	97.5	92.8	75.1
	BIC	93.5	88.4	86.7	98.6	80.5	78.5	92.9	84.9	81.0	96.9	78.4	69.5
2	F-OLS	94.4	94.2	94.4	72.6	72.8	78.3	95.1	95.0	95.4	36.4	43.5	47.0
	F-WCQR	95.1	93.9	94.3	84.3	86.0	84.6	94.6	93.9	94.5	86.5	86.0	90.0
	F-WCQR$^+$	95.3	93.9	94.8	85.2	85.5	86.6	94.7	95.0	94.8	88.4	86.0	90.9
	F-WCQR$^-$	93.9	93.1	94.1	83.2	85.5	83.3	94.1	94.0	93.8	87.1	86.5	89.8
	F-MLE	94.0	93.3	94.1	100.0	100.0	100.0	93.0	92.0	91.8	100.0	100.0	100.0
	AIC	93.4	90.4	84.4	97.8	97.7	79.3	92.9	91.1	87.0	98.3	107.6	95.7
	BIC	93.0	85.3	77.6	96.7	79.0	64.1	92.5	88.3	76.5	97.4	94.3	68.7
3	F-OLS	94.4	94.2	94.4	73.0	71.9	76.8	95.1	95.0	95.4	37.3	39.4	43.5
	F-WCQR	95.2	94.0	94.6	86.1	84.9	83.9	94.6	94.3	94.8	89.0	83.1	88.9
	F-WCQR$^+$	95.6	94.0	95.2	83.9	84.1	84.7	95.0	94.7	94.9	89.9	82.5	89.0
	F-WCQR$^-$	94.5	93.5	94.2	85.4	84.8	83.1	94.0	94.0	94.1	88.9	83.6	88.7
	F-MLE	94.2	93.1	94.0	100.0	100.0	100.0	93.4	91.8	92.2	100.0	100.0	100.0
	AIC	92.9	92.1	89.7	98.5	111.1	101.4	93.3	92.2	90.0	101.3	110.8	112.8
	BIC	92.6	88.8	78.5	96.6	95.0	67.5	93.1	90.4	85.0	100.8	102.2	86.6

Table 4: Inference for plasma beta-carotene level data

	$\hat{\boldsymbol{\beta}}_{LS}^{OSE}$	$\hat{\boldsymbol{\beta}}_{CQR}^{OSE}$	$\hat{\boldsymbol{\beta}}_{WCQR}^{FIC}$
Age	0	0	8.47(-2.67,23.58)
Quetelet index	0	0	-20.65(**-32.43, -7.61**)
Calories	-100.47	0	-9.43(-67.92, 9.33)
Fat	52.60	0	0.80(-14.45, 50.44)
Fiber	87.51	29.89	18.65(**4.86, 42.30**)
Alcohol	44.61	0	5.97(-0.86, 24.14)
Cholesterol	0	0	-1.83(-20.00, 12.96)
Smoking status (never)	51.71	0	17.42(-3.65, 72.80)
Smoking status (former)	72.48	0	4.72(-17.18, 63.34)
Vitamin use (fairly often)	130.39	30.21	-0.25(-17.17, 41.44)
Vitamin use (not often)	0	0	7.35(-5.34, 58.59)
MAPE	111.28	58.11	48.10(1.61)