Title	Cp-type criterion for quantile regression
Sub Title	
Author	高梨, 耕作(Takanashi, Kōsaku)
Publisher	Keio Economic Society, Keio University
Publication year	2018
Jtitle	Keio economic studies Vol.54, (2018.) ,p.1- 31
JaLC DOI	
Abstract	We propose a new Mallows'Cp type criterion for quantile regression (QR) which, unlike AIC or BIC, does not require parametric assumptions on the population and by construction is robust against misspecification. We show that our new Mallows type criterion for QR (QCp) is not only an asymptotically unbiased estimate of the average weighted squared error on the model average fit, but also asymptotically optimal in the sense of achieving the lowest possible weighted squared error in a class of discrete model sets. We also demonstrate that these asymptotic properties of the QCp estimator hold in finite samples with a simulation experiment.
Notes	
Genre	Journal Article
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AA00260492-20180000-0001

KEIO ECONOMIC STUDIES **54**, 1–31 (2018)

C_p-TYPE CRITERION FOR QUANTILE REGRESSION

Kōsaku TAKANASHI

Faculty of Economics, Keio University, Tokyo, Japan

First version received November 2017; final version accepted August 2018

Abstract: We propose a new Mallows' C_p type criterion for quantile regression (QR) which, unlike AIC or BIC, does not require parametric assumptions on the population and by construction is robust against misspecification. We show that our new Mallows type criterion for QR (QC_p) is not only an asymptotically unbiased estimate of the average weighted squared error on the model average fit, but also asymptotically optimal in the sense of achieving the lowest possible weighted squared error in a class of discrete model sets. We also demonstrate that these asymptotic properties of the QC_p estimator hold in finite samples with a simulation experiment.

Key words: Quantile regression, model selection, misspecification, mallows type criterion, asymptotic optimality.

JEL Classification Number: C52.

1. INTRODUCTION

We develop a new model selection criterion for quantile regression (QR) based on least absolute deviation (LAD). As ordinary least squares (OLS) estimates of the regression coefficients offer convenient summary statistics for the conditional expectation function of the model, the QR estimates can be used to infer about the conditional quantile function. Compared with OLS, the possibility of model misspecification and the importance of model selection has been less emphasized in the literature of QR, However, correct specification of the conditional quantile function is hard to find in real-world applications. It is safe to say that misspecification is the norm, not an exception. Therefore we need to devise a model selection criterion for QR. That is the purpose of our study.

Our new model selection criterion is based on Mallows' C_p proposed by Mallows (1973). We choose Mallows' C_p because, unlike Akaike information criterion (AIC; Akaike (1973)) or Bayesian information criterion (BIC; Schwarz (1978)), it does not utilize the likelihood function and we think it would be a good starting point to derive

Acknowledgments. The author appreciates helpful comments by Masao Ogaki, Taisuke Otsu and Teruo Nakatsuma. The author also wish to thanks the editor and anonymous referee for helpful comments.

a more robust model selection criterion for QR. Although the original Mallows' C_p is for model selection in OLS, we succeed in deriving a Mallows-type criterion for QR which is an asymptotically unbiased estimator of the weighted squared deviation from the true model. We also show that selecting a model with our new criterion is asymptotically optimal in the sense that the selected model asymptotically achieves the minimum weighted squared error as defined in Li (1987).

The paper is organized as follows. Section 2 presents the model selection problem we examine and introduces assumptions required to derive our main results. Section 3 introduces the new Mallows type C_p criterion for quantile regression and shows its asymptotic unbiasedness and optimality. Section 4 presents simulation evidence in support of our new criterion.

2. DESCRIPTION OF THE MODEL SELECTION PROBLEM, DEFINITION AND ASSUMPTION

2.1. *Basic Framework of Quantile Regression*

Let $\mathbf{y}_n = (y_1, \ldots, y_n)'$ be a vector of n independent responses and $X_n = \left(\mathbf{x}_1', \ldots, \mathbf{x}_n'\right)'$ be an $n \times p_n$ matrix whose ith row $\mathbf{x}_i'$ is the value of a $p_n -$ vector of explanatory variables associated with y_i . We can consider p_n , the number of regressors, grows as the sample size increases. As a general rule, a letter with subscript is used to denote observations of the corresponding random variable (e.g.y_i and y). And bold letter is used to denote a vector and capital letter denote an matrix.

ASSUMPTION 1. *The sequence $\{y_i, \mathbf{x}_i\}$ is independent and identically distributed (i.i.d.).*

The iid assumption is made for clarity and simplicity. As is now standard in the quantile regression literature, we define the asymmetric objective function $\rho_\tau : \mathbb{R} \to \mathbb{R}^+$ for given $\tau \in (0, 1)$ as

$$\rho_\tau (z) \triangleq z\varphi (z)$$

where

$$\varphi_\tau (z) \triangleq \tau - \mathbf{1} (z \leq 0)$$

also known as the "tick" of "check" function in the literature.

We are interested in the conditional quantile function(CQF) of y given $\mathbf{x}$. The conditional quantile is defined as

$$q_\tau (\mathbf{x}) \triangleq \inf \left\{q : F_y (q \mid \mathbf{x}) \geq \tau\right\} , \tag{2.1}$$

where $F_y (q \mid \mathbf{x})$ is the distribution function for y conditional on $\mathbf{x}$.

ASSUMPTION 2. *F is assumed to have conditional density $f_y (y \mid \mathbf{x})$.*

The CQF solves the minimization problem

$$q_\tau (\mathbf{x}) = \arg\min_q \mathbb{E} \left[\rho_\tau (y - q (\mathbf{x}))\right] ,$$

and the minimum is over the set of measurable function of $\mathbf{x}$ (Fox and Rubin (1964)). The CQF satisfies

$$\mathbb{E}\left[\varphi_\tau\left(y-q\left(\mathbf{x}\right)\right)\right]=0\,.$$

The linear quantile regression(QR) vector solves the population minimization problem

$$\boldsymbol{\beta}_\tau^* \triangleq \arg\min_{\beta\in\mathbb{R}^d}\mathbb{E}\left[\rho_\tau\left(Y-\mathbf{x}'\boldsymbol{\beta}\right)\right]\,.$$

We assume integrability and uniqueness of the solution.

ASSUMPTION 3. *There exists $\boldsymbol{\beta}_\tau^*$ such that*

$$\boldsymbol{\beta}_\tau^* \triangleq \arg\min_{\beta\in\mathbb{R}^d}\mathbb{E}\left[\rho_\tau\left(Y-\mathbf{x}'\boldsymbol{\beta}\right)\right]\,,$$

equivalently

$$\mathbb{E}\left[\mathbf{x}\varphi_\tau\left(y-\mathbf{x}'\boldsymbol{\beta}_\tau^*\right)\right]=0\,. \tag{2.2}$$

We want to estimate the conditional τ-quantile vector $\mathbf{q}_\tau=(q_{1\tau},\dots,q_{n\tau})$. We use the least absolute deviation (LAD) to estimate the conditional quantile. The QR process $\hat{\boldsymbol{\beta}}_\tau$ is formally defined as

$$\hat{\boldsymbol{\beta}}_\tau=\arg\min_\beta\sum_{i=1}^n\rho_\tau\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)$$

and estimate the conditional quantile as

$$\hat{q}_{i\tau}=\mathbf{x}_i'\hat{\boldsymbol{\beta}}_\tau\,.$$

REMARK 4. *Assumption 3 is more general setup than the traditional quantile regression models. Our setup is the case where all models are potentially misspecified but each model has the pseudo-true parameter β_τ^* satisfying (2.2). Our purpose described below is to select the most close model to the "true" conditional qunatile. Therefore, the consistency (in the traditional meaning) is out of scope in this article.*

2.2. *The Model Selection Problem*

In the standard LAD framework (which is not adopted here), one assumes that the conditional quantile regression model is correctly specified. That is, for some $\boldsymbol{\beta}_\tau^*$, one has $\mathbb{E}\left[\varphi_\tau\left(y-\mathbf{x}'\boldsymbol{\beta}_\tau^*\right)\right]=0$. Furthermore, to achieve identification, one assumes that $\boldsymbol{\beta}_\tau^*$ is the unique solution to these equations. The parameter $\boldsymbol{\beta}_\tau^*$ is then called the "true" value of $\boldsymbol{\beta}$. In this case, the standard LAD estimator $\hat{\boldsymbol{\beta}}_\tau$ of $\boldsymbol{\beta}_\tau^*$ is defined to minimize

$$\sum_{i=1}^n\rho_\tau\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}\right)\qquad,\text{over}\boldsymbol{\beta}\in\mathbb{R}^p\,.$$

The LAD estimator $\hat{\boldsymbol{\beta}}_n$ is consistent for $\boldsymbol{\beta}_\tau^*$ under minimal (and well-known) additional assumptions.

Often in empirical applications, however, researchers find that the J test of overidentifying restrictions (see Hansen (1982)) rejects the null hypothesis that all moment conditions are correct. Thus, it seems useful to consider statistical inference in the case where not all of the moment conditions are necessarily correct. That is what we do here. We presume that the researcher does not know a priori which regression variables are correct. (Otherwise he would discard the incorrect variables and be faced with the standard situation considered in the literature.)

Following Andrews (1999), we define the vector of selection variables. We let $h \in \mathbb{R}^{p_n}$ denote a *model selection vector*. By definition, h is a vector of zeros and ones. If the jth element of h is a one, then the jth variable is included. If the jth element is a zero, then it is not included. Let

$$H_{p_n} = \left\{ h \in \mathbb{R}^{p_n} : h_j = 0 \; or \; 1 \; \forall 1 \leq j \leq p_n, \; where \; h = \left(h_1, \ldots, h_{p_n}\right)' \right\} .$$

Let $\dim(h)$ denote the number of variables selected by h i.e., $\dim(h) = \sum_j h_j$ for $h \in H_{p_n}$. Thus, $\boldsymbol{\beta}(h)$ and $\mathbf{x}(h)$ is $\dim(h)$ vector of selection variables that are specified by $h \in H_{p_n}$. The model selection vector h selects not only a finite number of parameters but also the regression variables x_j.

For inference purposes, a class of models, indexed by $h \in H_{p_n}$, is to characterize the relation between the quantile response $\mathbf{q}_\tau$, and the explanatory variables. In this paper, we use a class of Least Absolute Deviation (LAD) estimators using the regressor matrix $\{X_n(h)\}_{h \in H_{p_n}}$,for each $h \in H_{p_n}$. We have a subvector $\mathbf{x}_i(h)$ of $\mathbf{x}_i$ and do the LAD estimation:

$$\hat{\boldsymbol{\beta}}_\tau(h) = \arg\min_\beta \sum_{i=1}^n \rho_\tau\left(y_i - \mathbf{x}_i'(h)\boldsymbol{\beta}\right)$$

and estimate the conditional quantile as

$$\hat{q}_{i\tau}(h) = \mathbf{x}_i'(h)\hat{\boldsymbol{\beta}}_\tau(h) .$$

If H_{p_n} contains more than one model, then we need to select a model from H_{p_n} using the given X_n and the data vector $\mathbf{y}_n$. The following is a typical example.

EXAMPLE 5 (Model Selection:). *Suppose that $p_n = p$ for all n and $\mathbf{q}_n = X_n\boldsymbol{\beta}$ with an unknown $p-$vector $\boldsymbol{\beta}$. Write $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$ and $X_n = (X_{n1}, \; X_{n2})$. It is suspected that the sub-vector $\boldsymbol{\beta}_2 = 0$, i.e., X_{n2} is actually not related to $\mathbf{q}_n$. Then we may propose the following two models:*

$$Model\ 1: \mathbf{q}_n = X_{n1}\boldsymbol{\beta}_1$$

$$Model\ 2: \mathbf{q}_n = X_n\boldsymbol{\beta} .$$

In this case, $H_{p_n} = \{1, 2\}$. More generally, we can consider models

$$\mathbf{q}_n(h) = X_n(h)\boldsymbol{\beta}(h) ,$$

where h is a subset of $\{1, \ldots, p\}$ and $\boldsymbol{\beta}(h)$contains the components of β that are indexed by the integers in h. In this case H_{p_n} consists of some distinct subsets of $\{1, \ldots, p\}$. If H_{p_n} contains all nonempty subsets of $\{1, \ldots, p\}$, then the number of models in H_{p_n} is $2^p - 1$.

In the previous example, regression vector is specified only by a finite number of parameters. This assumption is reasonable for obtaining an estimate of a testing procedure, but, for the true structure of the population, it is not so easily justified. Even if $q_\tau(\mathbf{x})$ is continuous on some finite interval, we cannot avoid dealing with an infinite series expansion, polynomial expansion, orthogonal expansion and so on. Therefore it is rather natural to specify $q_\tau(\mathbf{x})$ using infinitely many parameters.

EXAMPLE 6 ("Cut-off choice" of series estimation of nonparametric regression:). *Suppose that we wish to select the best approximation to the true median response surface from a class of linear models. Note that the approximation is exact if the response surface is actually linear and is in H_{p_n}. The proposed models are $\mathbf{q}_n = X_n(h)\boldsymbol{\beta}(h)$, $h \in H_{p_n}$, where $X_n(h)$ is a sub-matrix of X_n and $\boldsymbol{\beta}(h)$ is a sub-vector of a p_n-vector $\boldsymbol{\beta}(h)$ whose components have to be estimated. As a more specific example, we consider the situation where we try to approximate a one-dimensional curve by a polynomial, i.e., $\mathbf{q}_n = X_n(h)\boldsymbol{\beta}(h)$ with the i-th row of $X_n(h)$ being $\left(1, t_i, t_i^2, ..., t_i^{h-1}\right)'$, $i = 1, ..., n$. In this case $H_{p_n} = \{h_k, k = 1, ..., p_n\}$ and $h_k = \{1, ..., h\}$ corresponds to a polynomial of order h used to approximate the true model. The largest possible order of the polynomial may increase as n increases, since the more data we have, the more terms we can afford to use in the polynomial approximation.*

Note that H_{p_n} may not contain a correct model (Example 6). A correct model is not necessarily the best model, since there may be several correct models in H_{p_n} (Example 5) and there may be an incorrect model having a smaller loss than the best correct model (Example 6). Here, we allow the maximal dimension of model set H_{p_n}, p_n, to increase to infinitry with n in order to reduce approximation errors.

Different loss functions correspond to different optimal model. In this article the object of interest is q_τ, the conditional true τ-quantile of the distribution y. In the forecasting literature; e.g., Giacomini and Komunjer (2005), they use asymmetric loss to provides the best linear predictor for a response variable. This interpretation is not very satisfying, however, since prediction under asymmetric loss is typically not the object of interest in empirical work.

In the linear model selection using OLS fitting literature, the mean squared error loss is used. OLS estimates provide the minimum mean-squared error linear approximation to a conditional expectation function of any shape. The approximation properties of OLS have been emphasized by White (1980). On the other hand, QR is the best linear approximation to the conditional quantile function using a weighted mean-squared error loss function, much as OLS regression provides a minimum mean-squared error fit to the conditional expectation function. The approximation properties of QR have been shown by Angrist et al. (2006).

In the following subsection, we define the loss function for liner model selection using QR estimation process.

2.3. *Definition of Loss Function*

As mentioned above, QR is the best linear approximation to the conditional quantile function using a weighted mean squared error loss function (cf. Angrist et al. (2006), Theorem 1,2). We define the quantile regression specification error as

$$\Delta_{\tau i}(h) = \left(q_{\tau i} - \mathbf{x}_i'(h)\boldsymbol{\beta}^*(h)\right), \qquad \hat{\Delta}_{\tau i}(h) = \left(q_{\tau i} - \mathbf{x}_i'(h)\hat{\boldsymbol{\beta}}(h)\right),$$

$$\boldsymbol{\Delta}_{\tau n}(h) = \begin{bmatrix} \Delta_{\tau 1}(h) \\ \vdots \\ \Delta_{\tau n}(h) \end{bmatrix}, \qquad \hat{\boldsymbol{\Delta}}_n(h) = \begin{bmatrix} \hat{\Delta}_{\tau 1}(h) \\ \vdots \\ \hat{\Delta}_{\tau n}(h) \end{bmatrix}.$$

Angrist et al. (2006) have shown the following theorem.

THEOREM 7. *Suppose that*
(*i*) *the conditional density* $f_Y(y \mid X)$*exists a.s.* (*ii*) $\mathbb{E}[Y]$, $\mathbb{E}[q]$ *and* $\mathbb{E}[\mathbf{x}'\mathbf{x}]$ *are finite.* (*iii*) $\boldsymbol{\beta}_\tau$ *uniquely solves. Then*

$$\beta_\tau^* = \arg\min_\beta \mathbb{E}\left[w_\tau\left(\mathbf{x}, \boldsymbol{\beta}_\tau^*\right)\left(q_\tau - \mathbf{x}'\boldsymbol{\beta}\right)^2\right]$$

where

$$w_\tau\left(\mathbf{x}, \boldsymbol{\beta}^*\right) = \int_0^1 (1-u)\, f_e\left(u\left(q_\tau - \mathbf{x}'\boldsymbol{\beta}_\tau^*\right) \mid \mathbf{x}\right) du \geq 0$$

where the weight $w_\tau\left(\mathbf{x}, \boldsymbol{\beta}^*\right)$ *is a function of* $\mathbf{x}$ *only, so we write* $w_\tau\left(\mathbf{x}, \boldsymbol{\beta}_\tau^*\right)$ *as* $w_\tau(\mathbf{x})$.

Proof. See Angrist et al. (2006), Theorem 1,2. □

Theorem states that the population QR coefficient vector β_τ^* minimizes the expected weighted mean squared approximation error, i.e., the square of the difference between the true CQF and a linear approximation, with weighting function $w_\tau(\mathbf{x})$. The weights are given by the average density of the response variable over a line from the point of approximation, $\mathbf{x}'\boldsymbol{\beta}$, to the true conditional quantile, q.

We assume in this paper that the models in H_n are linear models and the LAD fitting is used under each proposed model. After observing the vector $\mathbf{y}_n$, our concern is to select a model h from H_n so that the weighted squared error loss

$$\begin{aligned} L_n^2(h) &= \frac{1}{n}\sum_{i=1}^n w_\tau\left(\mathbf{x}_i(h)\right)\left(q_i - \mathbf{x}_i'(h)\hat{\boldsymbol{\beta}}_\tau(h)\right)^2 \\ &= \frac{1}{n}\left\|\mathbf{q}_{\tau n} - \hat{\mathbf{q}}_{\tau n}(h)\right\|_{W_\tau(h)}^2 \end{aligned}$$

be as small as possible. The notations are following way. $\|\alpha\|_{W_\tau(h)} = \langle W_\tau(h)\boldsymbol{\alpha}, \boldsymbol{\alpha}\rangle^{1/2}$ is the semi-norm defined by any vector $\boldsymbol{\alpha}$ in Hilbert space l_2. $\hat{\mathbf{q}}_{\tau n}(h) = X_n(h)\hat{\boldsymbol{\beta}}_\tau(h)$ is the LAD estimator of $\mathbf{q}_{\tau n}$ under model h. $W_\tau(h)$ is the weight $n \times n$ matrix as

$$W_\tau(h) = \begin{bmatrix} w_\tau\left(\mathbf{x}_1(h)\right) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_\tau\left(\mathbf{x}_n(h)\right) \end{bmatrix}.$$

And $\mathbf{e}_n = (e_1, e_2, \ldots, e_n)'$ and $X_n(h) = (\mathbf{x}_1(h), \mathbf{x}_2(h), \ldots, \mathbf{x}_n(h))'$.

2.4. Mallows's C_p Type Criterion

To assess the proposed models, we use the weighted squared error loss. Our concern is to select an h from H_{p_n} so that the average weighted squared error

$$\begin{aligned} L_n^2(h) &= \frac{1}{n}\sum_{i=1}^{n} w_\tau(\mathbf{x}_i(h))\left(q_{\tau i} - \mathbf{x}_i'(h)\hat{\boldsymbol{\beta}}_\tau(h)\right)^2 \\ &= \frac{1}{n}\left\|\mathbf{q}_{\tau n} - \hat{\mathbf{q}}_{\tau n}(h)\right\|_{W_\tau(h)}^2 \end{aligned}$$

or the statistical L_2 risk

$$R_n^2(h) = \mathbb{E}\left[L_n^2(h)\right]$$

may be as small as possible.

The scenario is very similar to an estimation problem. We are not able to assess the finite sample average weighted squared error. Mallows's C_p type criterion is an unbiased estimate of the loss $L_n^2(h)$. The original version of C_p based on least squares estimation is an estimate of mean non-weighted squared error. Let L_{OLS}^2 be the average non-weighted squared error,

$$L_{OLS}^2(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[y_i \mid \mathbf{x}_i\right] - \mathbf{x}_i'(h)\hat{\boldsymbol{\beta}}_{OLS}(h)\right)^2$$

where $\hat{\boldsymbol{\beta}}_{OLS}(h) = \left(X_n'(h)X_n(h)\right)^{-1}X_n'(h)\mathbf{y}_n$. The original C_p criterion satisfies

$$\mathbb{E}\left[C_p(h)\right] = \mathbb{E}\left[L_{OLS}^2(h)\right] + \text{constant.}$$

The Mallows's C_p criterion may be used to select the quantile regression model h. Define

$$\hat{h}_n^{OLS} = \arg\min_{h\in H_{pn}} C_p(h)$$

the empirical Mallows's C_p selected quantile regression model.

Our purpose in this paper is to introduce the Mallows type C_p criterion which is based on LAD estimate and based on an estimate of average weighted squared error loss function. Next section develops our new criterion.

2.5. Basic Assumptions

We now state the basic assumptions under which the results below hold. This assumption holds quite generally.

We assume that the true model is the homoskedastic linear model having countably infinite regressors $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots)$.

ASSUMPTION 8. *The true model is the homoskedastic linear regression, write*

$$y_i = q_{i\tau} + e_i \tag{2.3}$$

$$q_{i\tau} = \sum_{j=1}^{\infty} x_{ji}\beta_j , \qquad i = 1, 2, \ldots, n , \tag{2.4}$$

and we assume that the random errors e_i are identically independently distributed and independent of $\mathbf{x}_i$, $\forall i$ *with conditional density* $f_e(\cdot \mid \mathbf{x})$ *and*

$$\inf \{\epsilon : F_e(\epsilon \mid \mathbf{x}) \geq \tau\} , = 0 , \tag{2.5}$$

$$\mathbb{E}[e_i \mid \mathbf{x}_i] = m . \tag{2.6}$$

Note that there are infinite number of regressors, so all model are misspecified.

ASSUMPTION 9. *We assume*

$$\mathbb{E}\left[q_{i\tau}^2\right] < \infty$$

and $q_{i\tau} = \sum_{j=1}^{\infty} x_{ji}\beta_j , \qquad i = 1, 2, \ldots, n,$ *converges in the mean square*[1].

We impose the following quantile version of the orthogonality condition.

ASSUMPTION 10. *For each* $h \in H_{p_n}$ *(*H_{p_n} *is the set of models considered by the researcher), there exist* $\beta_\tau^*(h)$ *such that*

$$\mathbb{E}\left[\mathbf{x}_i(h)\,\varphi_\tau\left(y_i - \mathbf{x}_i'(h)\,\boldsymbol{\beta}_\tau^*(h)\right)\right] = 0 . \tag{2.7}$$

Given the pseudo-true parameters $\boldsymbol{\beta}_\tau^*(h)$, assuming that $\mathbb{E}\left[\varphi_\tau\left(y_i - \mathbf{x}_i'(h)\,\boldsymbol{\beta}_\tau^*(h)\right) \mid \mathbf{x}_i\right] = 0$, which is stronger assumption, is equivalent to assuming that the conditional quantile model is correctly specified.

We assume the asymptotic representations of LAD-estimates (Bahadur form).

ASSUMPTION 11. *Parameters* $\hat{\boldsymbol{\beta}}_\tau(h)$ *are the form*

$$\hat{\boldsymbol{\beta}}_\tau(h) = \boldsymbol{\beta}_\tau^*(h) + \frac{1}{n}\sum_{i=1}^{n} T_\tau^{(1)}(\mathbf{x}_i(h)) + RE(h) \tag{2.8}$$

where $T^{(1)}$ *is the influence function and the matrix form is given by*

$$T_\tau^{(1)}(\mathbf{x}_i(h)) = \frac{1}{2} J_{\tau h}^{-1}\varphi_\tau\left(y_i - \mathbf{x}_i'(h)\,\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i(h)$$

where $J_{\tau h} \triangleq \mathbb{E}\left[f_e(\Delta_\tau(h) \mid \mathbf{x})\,\mathbf{x}(h)\,\mathbf{x}'(h)\right]$ *and* $RE(h)$ *is remainder term.*

We assume that the estimator has the following accuracy.

ASSUMPTION 12. *Remainder term satisfies*

$$\lim_{n\to\infty}\sup_{h\in H_n}\mathbb{E}\left[\|X(h)\,RE(h)\|^2\right] < \infty . \tag{2.9}$$

[1] This means

$$\lim_{k\to\infty}\mathbb{E}\left[\left(\sum_{j=1}^{\infty} x_{ji}\beta_j - \sum_{j=1}^{k} x_{ji}\beta_j\right)^2\right] = 0 .$$

3. MODEL SELECTION CRITERION

3.1. *Model Selection Criterion*

In the later of this paper, We will omit subscript τ. The Mallows type criterion for quantile regression "QC_p" is

$$QC_p(h) = \left\| \mathbf{y}_n - X_n(h)\hat{\boldsymbol{\beta}}(h) - m \right\|_{W(h)}^2$$
$$+ \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i(h))\,\rho_\tau\left(y_i - \mathbf{x}_i'(h)\boldsymbol{\beta}^*(h)\right) \cdot \mathbf{x}_i'(h)\,J_h^{-1}\mathbf{x}_i(h)$$

where $m = \mathbb{E}[e]$. We assume the moment of $e{:}m = \mathbb{E}[e]$ and $\boldsymbol{\beta}^*(h)$ is known. We discuss below the replacement of m with an estimate. The Mallows QC_p criterion may be used to select the quantile regression model h. Define

$$\hat{h}_n = \arg\min_{h \in H_{pn}} QC_p(h)$$

the empirical Mallows QC_p selected quantile regression model.

3.2. *Asymptotic Unbiasedness*

We present two justifications for the QC_p criterion. Our first is the classic observation that QC_p is an asymptotic unbiased estimate of the expected weighted squared error plus a constant. Proofs of the following lemmas are in Appendix.

LEMMA 13. *We have*

$$\mathbb{E}\left[QC_p(h)\right] = R_n^2(h) + \mathbb{E}\left[w(\mathbf{x}(h))\right]\sigma^2 + O_p\left(n^{-1}\right).$$

In practice, $m = \mathbb{E}[e]$ is unknown, so QC_p needs to be computed with a sample estimate. One choice is $\hat{m}_K = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i(K)\hat{\boldsymbol{\beta}}(K)\right)$, where K corresponds to a full model. If K grow with sample size n, $\hat{m}_K$ is consistent for m, which is valid as shown next.

LEMMA 14. *If $K \to \infty$ and $K/n \to 0$ as $n \to \infty$, then $\hat{m}_K \underset{p}{\to} m$ as $n \to \infty$.*

Furthermore, the unknown parameter β^* through $\sum_{i=1}^{n}\rho_\tau\left(y_i - \mathbf{x}_i'(h)\boldsymbol{\beta}^*(h)\right)$ must be estimated based on observed data. Let $\sum_{i=1}^{n}\rho_\tau\left(y_i - \mathbf{x}_i'(h)\hat{\boldsymbol{\beta}}(h)\right)$ be the consistent estimators of $\sum_{i=1}^{n}\rho_\tau\left(y_i - \mathbf{x}_i'(h)\boldsymbol{\beta}^*(h)\right)$. Then, we need to verify

$$\sup_{h \in \mathcal{H}} \frac{\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\rho_\tau\left(y_i - \mathbf{x}_i'(h)\hat{\boldsymbol{\beta}}(h)\right) - \rho_\tau\left(y_i - \mathbf{x}_i'(h)\boldsymbol{\beta}^*(h)\right)\right\}\mathbf{x}_i'(h)\,J_h^{-1}\mathbf{x}_i(h)\right|}{nR_n^2(h)} \to 0 \tag{3.1}$$

$$\sup_{h\in\mathcal{H}} \frac{\left|\left(\hat{m}-m\right)\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(\Delta_i(h)\right)\right|}{nR_n^2\left(h\right)} \to 0 \tag{3.2}$$

LEMMA 15. *If* $\hat{m} \underset{p}{\to} m$, n, $K \to \infty$ *and* $\frac{K}{n} \to 0$ *then*

$$P\left\{\sup_{h\in\mathcal{H}} \frac{\left|\left(\hat{m}-m\right)\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(q_i - \mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\right|}{nR_n^2\left(h\right)} > \delta\right\} \to 0$$

3.3. *Asymptotic Optimality*

Our second justification is that this method of QC_p is asymptotically optimal in the sense that the fitted estimates asymptotically achieve the minimum possible squared error in a class of LAD estimators. Note that minimizing $L_n^2\left(h\right)$ is not related to the consistency of $\hat{q}_\tau\left(\hat{h}_n\right)$ as an estimator of q_τ, i.e., $L_n^2\left(\hat{h}_n\right) \overset{p}{\to} 0$. In fact, it may not be worthwhile to discuss the consistency of $\hat{q}_\tau\left(\hat{h}_n\right)$, since in our circumstance, there is no consistent estimator of q_τ (e.g., $\min L_n\left(\hat{h}_n\right) \overset{p}{\nrightarrow} 0$). The purpose of model selection to to minimize the loss $L_n^2\left(h\right)$. The essential asymptotic requirement for a selection procedure is

$$\frac{L_n^2\left(\hat{h}_n\right)}{\inf_{h\in H_n} L_n^2\left(h\right)} \overset{p}{\to} 1 \tag{3.3}$$

i.e., $\hat{h}_n$ is asymptotically as efficient as $\inf_{H_{pn}} h$ in terms of the loss $L_n^2\left(h\right)$.

Li (1987) established the asymptotic optimality of Mallows' C_p criterion for OLS estimate under reasonable conditions. But, ordinary Mallows' C_p criterion does not work in LAD estimate for QR model. This is because the penalty term on Mallows' C_p is for estimation noise of OLS estimation. We need to modify the penalty term on Mallows C_p criterion. However, since QR estimation criterion function $\sum_{i=1}^{n} \rho_\tau\left(y_i - \mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\right)$ dose not have an algebraic solution, modification on penalty need asymptotic analysis. The following result is an analogue of Theorem 2.1 of Li (1987), who showed the asymptotic optimality of Mallows' criterion for model selection.

The primary goal of this paper is to demonstrate that under reasonable conditions, these procedures are asymptotically optimal in the sense (3.3). Thus using these procedures, statisticians may do as well as if they knew the true $\mathbf{q}_n$ (but are restricted to the use of the LAD estimators $\hat{\mathbf{q}}_n$). Appendix proves the asymptotic optimality of QC_p criterion under condition that

$$\sup_{h\in H_{pn}} \mathbb{E}\left[\mathbf{x}'\left(h\right)J_h^{-1}\mathbf{x}\left(h\right)\right] < \infty \tag{3.4}$$

$$\mathbb{E}\left[e^2\right] < \infty \tag{3.5}$$

$$\lim_{n\to\infty} \sum_{h\in H_{pn}} {}^1\!/\!\left(\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right) \to 0 \tag{3.6}$$

$$\sup_{h\in H_n} \mathbb{E}\left[\Delta^2(h)\right] < \infty \tag{3.7}$$

To explain condition (3.6), which Li (1987) referred to this condition as "reasonable", if all model h is bounded, then very likely $\Delta^2(h)$ is bounded away from 0. If every model h, either $h \to \infty$ or, if $\Delta(h) \to 0$, it dose so slower than $1/\sqrt{n}$.

we observe that condition (3.6) implies the condition (A.3) in Li (1987).

LEMMA 16. *We have*

$$\sup_{h\in H_n}\left|\frac{nR_n^2(h)}{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X_n(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}-1\right| \to 0\,.$$

Proof. The risk $R^2(h)$ is

$$\begin{aligned} nR_n^2(h) = {} & \mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X_n(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right] \\ & -2\mathbb{E}\left[\left\langle W(h)\,\boldsymbol{\Delta}_n(h),\ X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\rangle\right] \end{aligned}$$

To see that (3.6) implies $nR_n^2(h) \to \infty$, we need to verify that

$$\sup_{h\in\mathcal{H}}\left|\frac{\mathbb{E}\left[\left\langle W(h)\,\boldsymbol{\Delta}_n(h),\ X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\rangle\right]}{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}\right| \to 0. \tag{3.8}$$

By assumption (2.9), we have

$$\begin{aligned} & \mathbb{E}\left[\left\langle W(h)\,\boldsymbol{\Delta}_n(h),\ X_n(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\rangle\right] \\ = & \mathbb{E}\left[\left\langle W(h)\,\boldsymbol{\Delta}_n(h),\ X_n(h)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}Q_h^{-1}\varphi\left(y_i-\mathbf{x}_i'(h)\,\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i(h)\right\rangle\right] \\ & +\mathbb{E}\left[\left\langle W(h)\,\boldsymbol{\Delta}_n(h),\ X_n(h)\,RE\right\rangle\right]. \end{aligned}$$

By assumption (3.4),(3.7), we have

$$\begin{aligned} & \mathbb{E}\left[\left\langle W(h)\,\boldsymbol{\Delta}_n(h),\ X_n(h)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}Q_h^{-1}\varphi\left(y_i-\mathbf{x}_i'(h)\,\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i(h)\right\rangle\right] \\ \underset{\text{cross}}{=} & \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\mathbb{E}\left[w_\tau\left(\mathbf{x}_i(h)\right)\Delta_i(h)\,\mathbf{x}_i'(h)\,Q_h^{-1}\varphi\left(y_i-\mathbf{x}_i'(h)\,\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i(h)\right] \\ < & \infty. \end{aligned}$$

^{2} By Cauchy-Schwarz inequality, we have

$$\sup_{h\in\mathcal{H}}\left|\frac{\mathbb{E}\left[\langle W(h)\,\mathbf{\Delta}_n,\ X(h)\,RE\rangle\right]}{\mathbb{E}\left[\|\mathbf{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}\right|$$

$$\leq\sup_{h\in\mathcal{H}}\left|\frac{\mathbb{E}\left[\|\mathbf{\Delta}_n\|^2_{W(h)}\right]^{1/2}\mathbb{E}\left[\|X(h)\,RE(h)\|^2_{W(h)}\right]^{1/2}}{\mathbb{E}\left[\|\mathbf{\Delta}_n\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}\right|$$

$$\leq\sup_{h\in\mathcal{H}}\left|\frac{\mathbb{E}\left[\|X(h)\,RE(h)\|^2_{W(h)}\right]^{1/2}}{\left(\mathbb{E}\left[\|\mathbf{\Delta}_n\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right)^{1/2}}\right|$$

and by assumption (2.9) $\mathbb{E}\left[\|X(h)\,RE(h)\|^2_{W(h)}\right]<\infty$. Thus (3.8) is established. □

Now we are ready to prove the main result of this paper - the asymptotic optimality of QC_p. Formally:

PROPOSITION 17. *As $n\to\infty$, under conditions (3.4)(3.5)(3.6)(3.7), QC_p is asymptotically optimal; i.e.,*

$$\frac{L_n^2\left(\hat{h}_n\right)}{\inf_{h\in H_n}L_n^2(h)}\xrightarrow{p}1\,.$$

The proof is in appendix. The following sub section is devoted to the remark on other criteria.

3.4. *Remark on Other Criteria*

The fact that AIC and BIC rely on the likelihood which defines the class of models means that these methods suffer from possible misspecification. Since the BIC paradigm is developed under the assumption that the "true" model is in fact within the class of models under consideration, this paradigm may be far off the mark if that is not the case. In practice it is more appropriate to think of any models as mere approximations, and the "true" model is too complex to be precisely approximated by anything in the class of models.

However, there is an argument that is favor of BIC regardless of the true model's complexity, which is in line with Rissanen (1986), where a BIC-like criterion is shown to be optimal from an information theoretic point of view. It is true that statistical models are mostly used in areas where the existence of a true model is doubtful. But, there is ample reason to choose a simple parsimonious model that might be untrue, even if a true model does exist. The goal of statistical analysis in this situation is to extract information rather than to identify the true model. In other words, the parsimony

[2] Cross term vanishes.

principle should be applied not only to candidates for the true model, but the true model itself as well.

The conclusion, of course, depends on the choice of a loss function since it has a tremendous bearing on the asymptotic properties of the corresponding model selection criterion. If one uses the information theoretic argument of Rissanen (1986) or the accumulated prediction error, BIC can be shown to be loss-efficient. In any case, we need to be careful to define the object of interest (such as prediction error, weighted mean-squared approximation error, stochastic complexity, etc) when we evaluate model selection criteria.

In the model selection problem for QR, if we define the object of interest as one step ahead forecast, AIC is an optimal procedure. If we define the object of interest as accumulated prediction error or consistency under correctly specified condition, BIC is a best one. If the object of interest is the best linear predictor for a response variable under asymmetric loss, FPE (Burman and Nolan (1995)) is an effective one. These object of interest are not very satisfying, however, since prediction is typically not the object of interest in typical empirical studies on economics.

The fact that LAD estimation is as easy to compute as OLS regression coefficients and that QR provides a meaningful and well-understood summary statistic for the conditional quantile undoubtedly contributes to the recent popularity of QR as an empirical tool. In view of the possibility of interpretation under misspecification, QR estimated by LAD implicitly provides a weighted minimum distance approximation to the true conditional quantile function. Therefore, we should choose the weighted squared error as a loss function for QR model selection estimated by LAD.

It is useful to compare the QR fit estimated by LAD to an explicit minimum distance (MD) fit similar to that discussed by Chamberlain (1994). The MD estimator for QR is the vector $\tilde{\beta}\,(\tau)$ that solves

$$\tilde{\boldsymbol{\beta}}\,(\tau) = \arg\min_{\beta} \frac{1}{n} \sum_{j=1}^{J} \left(\hat{q}_\tau \left(\mathbf{x}_j\right) - \mathbf{x}_j' \boldsymbol{\beta} \right)^2$$

where $\hat{q}_\tau \left(\mathbf{x}_j\right)$ is the sample quantile given $\mathbf{x} = \mathbf{x}_j$. If QR is estimated by MD, the loss function should be the average (non-weighted) squared error and selection criterion is directly derived from MD residuals. In contrast to LAD, however, this MD estimator relies on the ability to estimate $\hat{q}_\tau \left(\mathbf{x}_j\right)$ in a nonparametric first step, which, as noted by Chamberlain (1994), may be feasible only when $\mathbf{x}$ is low dimensional, the sample size is large, and sufficient smoothness of q_τ is assumed and the distribution of the vector of covariates $\mathbf{x}$ have finite support with $P\left(\mathbf{x} = \mathbf{x}_j\right) = a_j$ for $j = 1, \ldots, J$.

Recently, the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) is widely used in a high (ultra-high) dimensional circumstance(see also Fan and Li (2001), Tibshirani (2011)). The quantile regression case is studied in Belloni and Chernozhukov (2011). The LASSO estimator performs the "oracle" property in terms of selecting the correct model, when the regularization parameter is appropriately chosen. That is, when the true parameters have some zero components, they are estimated as 0 with probability tending to 1, and the nonzero components are estimated as nonzero one. This LASSO scheme and "oracle" property are very useful for a regression in high-dimensional sparse models. In such models, the number of regressors is possibly larger than the sample size, but the number of significant regressors (nonzero components) is smaller than the sample size. However, the estimated significant nonzero regressors are all biased because of the existence of regularization term of LASSO. Therefore, the approximation property of LASSO estimation with respect to (weighted) L^2 loss may be poor. The reason is clear; the objective functions of LASSO type quantile regression are under constraint of L^1-regularization, and then, the loss function, weighted L^2 loss, is inappropriate. The appropriate loss function for LASSO estimate may be the weighted L^2 loss with L^1-penalty. However, statistical meaning of this loss is unclear for the purpose of selecting the best approximation by anything in the class of models.

4. FINITE SAMPLE INVESTIGATION

We now investigate the finite properties of the our model selection criterion in a simple simulation experiment. We present two examples.

The first example is the linear regression (Example 1) with $P_n = 16$; that is

$$y_i = \beta_1 x_{i1} + \cdots + \beta_{16} x_{i16} + e_i \quad , i = 1, \ldots 50, 100, 400, 1600,$$

where e_i are independent and identically distributed as $N(0, 1)$, x_{ij} is the ith value of the explanatory variable x_j, $x_{i1} = 1$. For simplicity, we assume that x_{ij} $(j = 1, \ldots, 20, i = 1, \ldots, n)$ are orthonormal variables.

figure 4.1 shows the distribution of the number of explanatory variables that are selected using the QC_p for the case of an median regression model. The graph on the upper left, upper right, lower left, and lower right plots represent the cases in which the sample size is 50, 100, 400, and 1600, respectively. The results suggest that when the true order is a finite number, the distribution of dimensions converges to a certain distribution when the size of n becomes large.

The second example considered is the series approximation to a possibly nonlinear curve (Example 2); that is, we select a model from the following class of models. The setting is the infinite order regression $y_i = \sum_{j=1}^{\infty} \beta_j x_{ji} + e_i$. We set $x_{1i} = 1$ to be the intercept; the remaining x_{ji} are independent and identically distributed $N(0, 1)$. The error e_i is $N(0, 1)$ and independent of x_i. The parameters are determined by the rule $\beta_j = c\sqrt{2} \cdot j^{-1/2-1/2}$.

The sample size is varied between $n = 50, 100, 400, 1600$, and 2400. The number of models H_{p_n} is determined by the rule $H_{p_n} = 3\sqrt[3]{n}$ (so $H_{p_n} = 11, 13, 22, 35, 40$ for the five sample sizes). To evaluate the estimator, we compute the risk (expected weighted squared error). We do this by computing averages across 10,000 simulation draws.

The risk calculations are displayed in figure 4.2. In this panel, risk (expected weighted squared error) is displayed on the Y axis and the sample sizes is displayed on the X axis.In this panel, the average loss of model selected by QC_p achieves the lowest risk as sample size increase. When the sample size is small, QC_p selection is dismal

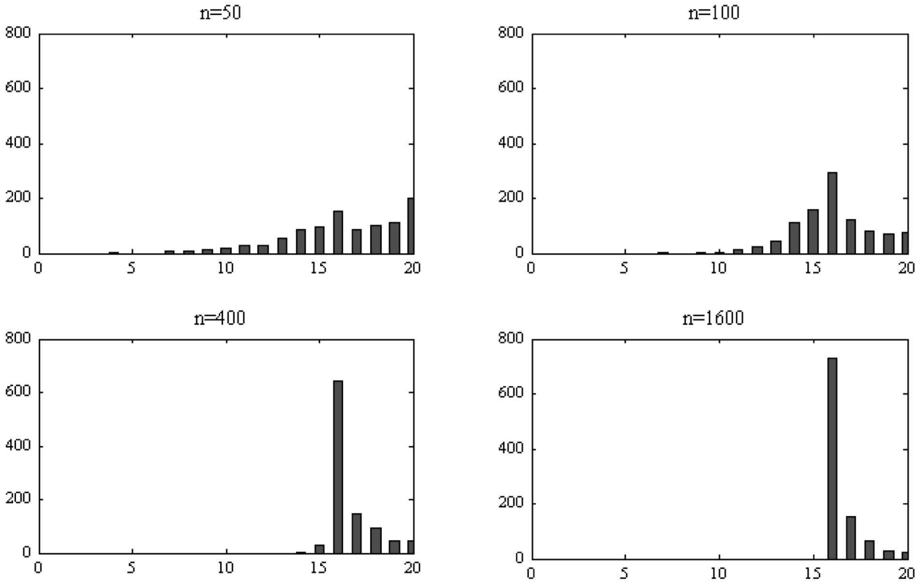

Figure 4.1. Distributions of dimension selected by QC_p. The upper left, upper right, lower left, and lower right plots represent the cases in which the sample size is 50, 100, 400, and 1600, respectively.

performance because of estimation noise of coefficient parameter and approximation noise of criterion. An improvement of small sample performance of this criterion is the future study with extra work.

APPENDIX

A. PROOFS

A.1. Proof of Lemmas

A.1.1. Proof of Lemma 13

Proof. First, observe the identity

$$\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i - \mathbf{x}_i'\left(h\right)\hat{\boldsymbol{\beta}}(h) - m\right)^2$$

$$+\frac{1}{n^2}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i - \boldsymbol{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)$$

$$=L_n(h)+\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(e_i - m\right)^2 + 2\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(e_i - m\right)\left(q_i - \mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)$$

$$+\frac{1}{n}\left[-2\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i - \mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^* - m\right)\mathbf{x}_i'\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right) - \boldsymbol{\beta}^*\left(h\right)\right)\right.$$

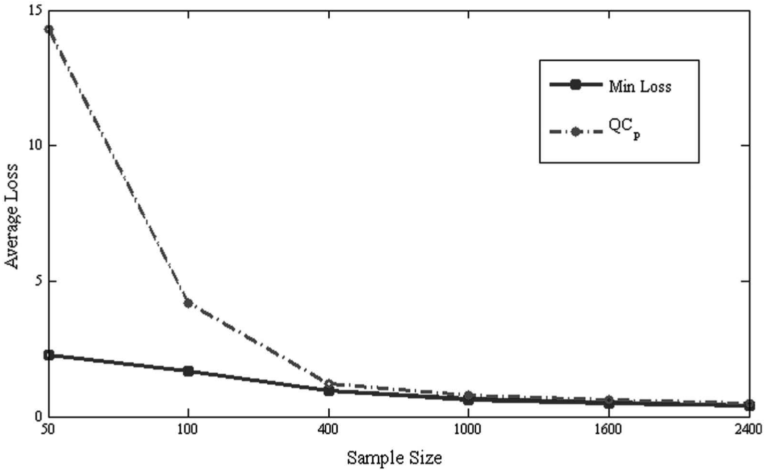

Figure 4.2. The solid and dotted lines correspond to minimum loss and QC_p selection respectively.

$$+\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\boldsymbol{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)$$
$$\left.+2\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(q_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right].$$

And take the expectation

$$\mathbb{E}\left[L_n\left(h\right)\right]+\mathbb{E}\left[w\left(\mathbf{x}_i\left(h\right)\right)\sigma^2\right]$$
$$=\frac{1}{n}\mathbb{E}\left[-2\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*-m\right)\mathbf{x}_i'\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right.$$
$$+\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\boldsymbol{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)$$
$$\left.+2\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(q_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right].$$

And using the Bahadur representation (2.8), the fourth term bracket on the righthand side is as follows:

First term in the bracket

$$\frac{1}{n}\mathbb{E}\left[-2\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*-m\right)\mathbf{x}_i'\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right.$$
$$\left.+\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\boldsymbol{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)\right]$$

is as follows: using Bahadur representation, we have

$$
\begin{aligned}
&\frac{1}{n}\mathbb{E}\left[-2\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*-m\right)\mathbf{x}_i'\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right.\\
&\left.+\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\boldsymbol{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)\right]\\
=&\frac{1}{n}\mathbb{E}\left[-2\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*-m\right)\right.\\
&\times\left(\frac{1}{2}\mathbf{x}_i'\left(h\right)J_h^{-1}\frac{1}{n}\sum_{j=1}^{n}\left\{\varphi\left(y_j-\mathbf{x}_j'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_j\left(h\right)\right\}+RE\left(h\right)\right)\\
&\left.+\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\boldsymbol{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)\right]
\end{aligned}
$$

and calculate

$$
\begin{aligned}
&\frac{1}{n}\mathbb{E}\left[-2\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)\frac{1}{2}J_h^{-1}\right.\\
&\left.\times\frac{1}{n}\left(\varphi\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_i\left(h\right)+\sum_{j\neq i}^{n}\varphi\left(y_j-\mathbf{x}_j'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_j\left(h\right)+RE\left(h\right)\right)\right]\\
&+\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\boldsymbol{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)\right]
\end{aligned}
$$

because the $\mathbf{x}_i$ and $\mathbf{x}_{j\neq i}$ are independent, we have

$$
\begin{aligned}
&-\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\varphi\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_i\left(h\right)\right]\\
&+\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\boldsymbol{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)\right]\\
&-\mathbb{E}\left[\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)\right]\\
&\quad\times\mathbb{E}\left[J_h^{-1}\frac{1}{n}\left(\sum_{j\neq i}^{n}\varphi\left(y_j-\mathbf{x}_j'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_j\left(h\right)\right)\right]\\
&+\mathbb{E}\left[-\frac{1}{n}2\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)RE\left(h\right)\right].
\end{aligned}
$$

By assumption, $\mathbb{E}\left[J_h^{-1}\frac{1}{n}\left(\sum_{j\neq i}^{n}\varphi\left(y_j-\mathbf{x}_j'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_j\left(h\right)\right)\right]=0$. Then, the first term is

$$-\mathbb{E}\left[\frac{1}{n}2\sum_{i=1}^{n}\left(y_i-\mathbf{x}_i'(h)\boldsymbol{\beta}(h)^*\right)\mathbf{x}_i'(h)\,RE(h)\right]. \tag{A.1}$$

Second term in the bracket , $\frac{1}{n}\mathbb{E}\left[2\sum_{i=1}^{n}w(\mathbf{x}_i(h))\left(q_i-\mathbf{x}_i'(h)\boldsymbol{\beta}(h)^*\right)\mathbf{x}_i'(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right]$, is obtained similar way. Calculate the same way and we have

$$\begin{aligned}&\frac{1}{n}\mathbb{E}\left[2\sum_{i=1}^{n}w(\mathbf{x}_i(h))\left(q_i-\mathbf{x}_i'(h)\boldsymbol{\beta}(h)^*\right)\mathbf{x}_i'(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right]\\&=\frac{1}{n}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}w(\mathbf{x}_i(h))\left(q_i-\mathbf{x}_i'(h)\boldsymbol{\beta}(h)^*\right)\mathbf{x}_i'(h)\,J_h^{-1}\varphi\left(y_i-\mathbf{x}_i'(h)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i(h)\right.\\&\left.\quad+2\sum_{i=1}^{n}w(\mathbf{x}_i(h))\left(q_i-\mathbf{x}_i'(h)\boldsymbol{\beta}(h)^*\right)\mathbf{x}_i'(h)\,RE(h)\right]\end{aligned} \tag{A.2}$$

So the second term in the bracket is ;

$$\begin{aligned}&\tfrac{1}{n}\mathbb{E}\left[\tfrac{1}{n}\sum_{i=1}^{n}w^2(\mathbf{x}_i(h))\left(q_i-\mathbf{x}_i'(h)\boldsymbol{\beta}(h)^*\right)^2\mathbf{x}_i'(h)\,J_h^{-1}\mathbf{x}_i(h)\right]\\&+\tfrac{1}{n}\mathbb{E}\left[2\sum_{i=1}^{n}w(\mathbf{x}_i(h))\left(q_i-\mathbf{x}_i'(h)\boldsymbol{\beta}(h)^*\right)\mathbf{x}_i'(h)\,RE(h)\right]\end{aligned}$$

Finally the remainder (A.1) plus (A.2) is

$$\begin{aligned}&\frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}w^2(\mathbf{x}_i(h))\left(q_i-\mathbf{x}_i'(h)\boldsymbol{\beta}(h)^*\right)^2\mathbf{x}_i'(h)\,J_h^{-1}\mathbf{x}_i(h)\right]\\&-\frac{1}{n}\mathbb{E}\left[2\sum_{i=1}^{n}w(\mathbf{x}_i(h))(e_i-m)\,\mathbf{x}_i'(h)\,RE(h)\right].\end{aligned} \tag{A.3}$$

And since

$$\begin{aligned}&\frac{1}{n}\mathbb{E}\left[2\sum_{i=1}^{n}w(\mathbf{x}_i(h))(e_i-m)\,\mathbf{x}_i'(h)\,RE(h)\right]\\&\leq\frac{1}{n}\mathbb{E}\left[2\left|\left\langle\frac{1}{\sqrt{n}}W(h)\,X'(h)(\mathbf{e}_n-m),\ \sqrt{n}RE_n(h)\right\rangle\right|\right]\\&\underset{\text{Cauchy-Schwarz}}{\leq}\frac{1}{n}2\mathbb{E}\left[n\left\|RE_n(h)\right\|^2\right]^{\frac{1}{2}}\mathbb{E}\left[n^{-1}\left\|X'(h)(\mathbf{e}_n-m)\right\|_{W^2(h)}^2\right]^{\frac{1}{2}}\end{aligned}$$

, by assumption (3.5) and (2.9), we have

$$\mathbb{E}\left[n\left\|RE_n(h)\right\|^2\right]^{\frac{1}{2}}\to 0$$

$$\mathbb{E}\left[\frac{1}{n}\left\|X'(h)(\mathbf{e}_n - m)\right\|^2_{W^2(h)}\right]^{\frac{1}{2}} < \infty .$$

Thus the remainder (A.3) has $O_p\left(n^{-1}\right)$ rate. □

A.1.2. Proof of Lemma 14

Proof. Since $\hat{e}_K = y_i - \mathbf{x}_i'(K)\hat{\boldsymbol{\beta}}(K) = q_i - \mathbf{x}_i'(K)\boldsymbol{\beta}^*(K) - \mathbf{x}_i'(K)\left(\hat{\boldsymbol{\beta}}(K) - \boldsymbol{\beta}^*(K)\right)$ $+e_i$, we see that

$$\hat{m}_K = \frac{1}{n}\sum_{i=1}^{n}\left(q_i - \mathbf{x}_i'(K)\boldsymbol{\beta}^*(K)\right) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'(K)\left(\hat{\boldsymbol{\beta}}(K) - \boldsymbol{\beta}^*(K)\right) + \frac{1}{n}\sum_{i=1}^{n}e_i \tag{A.4}$$

We examine the terms on the right side of A.4.

$$\begin{aligned} &P\left\{\left|q(\mathbf{x}) - \mathbf{x}_i'(K)\boldsymbol{\beta}^*(K)\right| > \delta\right\} \\ &\underset{\text{Markov}}{\leq} \frac{1}{\delta}\mathbb{E}\left|\sum_{j=1}^{\infty}x_{ji}\beta_j - \sum_{j=1}^{K}x_{ji}\beta_j^*\right| \end{aligned} \tag{A.5}$$

First, by an application of Section 2.3. of Angrist et al. (2006), (A.5)

$$\begin{aligned} &\sum_{j=1}^{\infty}x_{ji}\beta_j - \sum_{j=1}^{K}x_{ji}\beta_j^* \\ &= \sum_{j=1}^{\infty}x_{ji}\beta_j - \sum_{j=1}^{K}x_{ji}\beta_j \\ &\quad - \mathbb{E}\left[\tilde{w}(\mathbf{x})\mathbf{x}(K)\mathbf{x}'(K)\right]^{-1}\mathbb{E}\left[\tilde{w}(\mathbf{x})\mathbf{x}(K)\left(\sum_{j=1}^{\infty}x_{ji}\beta_j - \sum_{j=1}^{K}x_{ji}\beta_j\right)\right] \\ &= \sum_{j=K}^{\infty}x_{ji}\beta_j - \mathbb{E}\left[\tilde{w}(\mathbf{x})\mathbf{x}(K)\mathbf{x}'(K)\right]^{-1}\mathbb{E}\left[\tilde{w}(\mathbf{x})\mathbf{x}(K)\left(\sum_{j=K}^{\infty}x_{ji}\beta_j\right)\right] \end{aligned}$$

where $\tilde{w}(\mathbf{x})$ is weight function of $\mathbf{x}$ only. Since $K \to \infty$ as $n \to \infty$ and the integrability of q implies $\mathbb{E}\left|\sum_{j=K}^{\infty}x_{ji}\beta_j\right| \to 0$ as $K \to \infty$ (A.5) vanish.

Second,

$$\begin{aligned} &p\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'(K)\left(\hat{\boldsymbol{\beta}}(K) - \boldsymbol{\beta}^*(K)\right)\right| > \delta\right) \\ &\leq \frac{\mathbb{E}\left[\left(\hat{\boldsymbol{\beta}}(K) - \boldsymbol{\beta}^*(K)\right)' X'(h)X(h)\left(\hat{\boldsymbol{\beta}}(K) - \boldsymbol{\beta}^*(K)\right)\right]}{\delta^2 n^2} \\ &\to 0. \end{aligned}$$

Third, by law of large number, $\frac{1}{n}\sum_{i=1}^{n} e_i \underset{p}{\rightarrow} m$, and we conclude that $\hat{m}_K \underset{p}{\rightarrow} m$. □

A.1.3. Proof of Lemma 15

Proof. Note that

$$\begin{aligned}\hat{m} - m &= \frac{1}{n}\sum_i y_i - \mathbf{x}_i'(K)\hat{\boldsymbol{\beta}}(K) - m \\ &= \frac{1}{n}\sum_i \left(q_i - \mathbf{x}_i'(K)\hat{\boldsymbol{\beta}}(K)\right) + \frac{1}{n}\sum_i e_i - m \\ &= \frac{1}{n}\sum_i \left(q_i - \mathbf{x}_i'(K)\boldsymbol{\beta}^*(K)\right) + \mathbf{x}_i'(K)\left(\boldsymbol{\beta}^*(K) - \hat{\boldsymbol{\beta}}(K)\right) + \frac{1}{n}\sum_i e_i - m.\end{aligned}$$

If we can show the following

$$\left|\frac{1}{n}\sum_{j=1}^{n}\left(q_j - \mathbf{x}_j'(K)\boldsymbol{\beta}^*(K)\right)\sum_{i=1}^{n} w(\mathbf{x}_i(h))\left(q_i - \mathbf{x}_i'\boldsymbol{\beta}^*(h)\right)\right| \xrightarrow{p} 0 \tag{A.6}$$

$$\left|\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j'(K)\left(\boldsymbol{\beta}^*(K) - \hat{\boldsymbol{\beta}}(K)\right)\sum_{i=1}^{n} w(\mathbf{x}_i(h))\left(q_i - \mathbf{x}_i'\boldsymbol{\beta}^*(h)\right)\right| \xrightarrow{p} 0 \tag{A.7}$$

$$\left|\left(\frac{1}{n}\sum_{j=1}^{n} e_j - m\right)\sum_{i=1}^{n} w(\mathbf{x}_i(h))\left(q_i - \mathbf{x}_i'\boldsymbol{\beta}^*(h)\right)\right| \xrightarrow{p} 0 \tag{A.8}$$

the prove is established. (A.7)(A.8) is similar to the optimality proof of (A.9), so we only show (A.8).

Let $\frac{1}{n}\sum_{j\neq i}^{n}\left(q_j - \mathbf{x}_j'(K)\boldsymbol{\beta}^*(K)\right) = m'$. We have

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{n}\left(q_j - \mathbf{x}_j'(K)\boldsymbol{\beta}^*(K)\right)\sum_{i=1}^{n} w(\mathbf{x}_i(h))\left(q_i - \mathbf{x}_i'\boldsymbol{\beta}^*(h)\right) \\ &= \left(m' + \frac{1}{n}\left(q_i - \mathbf{x}_i'(K)\boldsymbol{\beta}^*(K)\right)\right)\sum_{i=1}^{n} w(\mathbf{x}_i(h))\left(q_i - \mathbf{x}_i'\boldsymbol{\beta}^*(h)\right)\end{aligned}$$

By Chebysev's inequality, we have

$$p\left\{\sup_{h\in\mathcal{H}} \frac{\left|\sum_{i=1}^{n}\left(m' - \mathbb{E}m'\right) w(\mathbf{x}_i(h))\left(q_i - \mathbf{x}_i'(h)\boldsymbol{\beta}^*(h)\right)\right|}{nR_n^2(h)} > \delta\right\}$$

$$\underset{\text{Chebysev}}{\le} \sum_{h\in H} \frac{\mathbb{E}\left[\left|\sum_{i=1}^{n} \left(m' - \mathbb{E}m'\right) w\left(\mathbf{x}_i\left(h\right)\right)\left(q_i - \mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\right|^2\right]}{\left(nR_n^2\left(h\right)\right)^2}$$

By Whittle's inequality, for some constant C we have

$$\begin{aligned}
&\sum_{h\in H} \frac{\mathbb{E}\left[\left|\left(m' - \mathbb{E}m'\right)\sum_{i=1}^{n} w\left(\mathbf{x}\left(h\right)\right)\left(q - \mathbf{x}'\left(h\right)\boldsymbol{\beta}^*(h)\right)\right|\right]}{nR_n^2\left(h\right)}\\
\underset{\text{Whittle}}{\le}\ & C\cdot\mathbb{E}\left|m'\right|\sum_{h\in H}\frac{\mathbb{E}\left[\sum_{i=1}^{n}\left(w\left(\mathbf{x}\left(h\right)\right)\left(q - \mathbf{x}'\left(h\right)\boldsymbol{\beta}^*(h)\right)\right)^2\right]}{\left(nR_n^2\left(h\right)\right)^2}\\
\le & C\cdot\mathbb{E}\left|m'\right|\\
\to & 0
\end{aligned}$$

and for some constant C', we have

$$\begin{aligned}
&\mathbb{E}m'\cdot\left(\frac{\mathbb{E}\left[\left|\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\left(q_i - \mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\right|\right]}{nR_n^2\left(h\right)}\right)\\
\le & C'\cdot\mathbb{E}m'\\
\to & 0
\end{aligned}$$

□

A.2. *Proof of Asymptotic Optimality of QC_p*

A.2.1. *Proof Strategy*

First, observe the identity

$$\begin{aligned}
&\frac{1}{n}\left\|\mathbf{y}_n - X_n\left(h\right)\hat{\boldsymbol{\beta}}\left(h\right) - m\right\|_{W(h)}^2\\
&+\frac{1}{n^2}\sum_{i=1}^{n} w_\tau\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i - \mathbf{x}_i'\left(h\right)\hat{\boldsymbol{\beta}}(h)\right)\mathbf{x}_i'\left(h\right)Q_h^{-1}\mathbf{x}_i\left(h\right)\\
=&L_n\left(h\right) + \frac{1}{n}\left\|\mathbf{e}_n - m\right\|_{W(h)}^2 + 2\frac{1}{n}\left\langle\left(\mathbf{e}_n - m\right), W\left(h\right)\boldsymbol{\Delta}_n\left(h\right)\right\rangle\\
&+\frac{1}{n}\left[-2\left\langle W\left(h\right)\left(\mathbf{y}_n - X_n\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right),\ X_n\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right) - \boldsymbol{\beta}^*\left(h\right)\right)\right\rangle\right.\\
&+2m\left\langle\mathbf{w}_n\left(h\right),\ X_n\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right) - \boldsymbol{\beta}^*\left(h\right)\right)\right\rangle
\end{aligned}$$

$$+\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*(h)\right)\cdot\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)$$

$$+2\left\langle W\left(h\right)\boldsymbol{\Delta}_n\left(h\right),\ X_n\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right\rangle\Big]$$

$\frac{1}{n}\sum_{i=1}^{n} e_i^2$ is independent of h. And using the Bahadur representation (2.8), the fourth term bracket on the righthand side is as follows:

First term in the bracket ,$2\left\langle W\left(h\right)\left(\mathbf{y}_n-X_n\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right),\ X_n\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right\rangle$, is as follows

$$2\left\langle W\left(h\right)\left(\mathbf{y}_n-X_n\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right),\ X_n\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right\rangle$$

$$=2\left\langle W(h)\left(\mathbf{y}_n-X_n(h)\boldsymbol{\beta}^*(h)\right),\ X_n(h)\left(\frac{1}{2}J_h^{-1}\frac{1}{n}\sum_{j=1}^{n}\varphi\left(y_j-\mathbf{x}_j'(h)\boldsymbol{\beta}^*(h)\right)\mathbf{x}_j(h)+RE(h)\right)\right\rangle$$

$$=\sum_{i=1}^{n} w_\tau\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\cdot$$

$$\cdot\frac{1}{n}\left(\varphi\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_i\left(h\right)+\sum_{j\neq i}^{n}\varphi\left(y_j-\mathbf{x}_j'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_j\left(h\right)+RE\left(h\right)\right)$$

$$=\Bigg[\frac{1}{n}\sum_{i=1}^{n} w_\tau\left(\mathbf{x}_i\left(h\right)\right)\rho_\tau\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\mathbf{x}_i\left(h\right)$$

$$\frac{1}{n}\sum_{1\leq i<j\leq n}^{n} w_\tau\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)J_h^{-1}\varphi\left(y_j-\mathbf{x}_j'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_j\left(h\right)$$

$$+\sum_{i=1}^{n} w_\tau\left(\mathbf{x}_i\left(h\right)\right)\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\mathbf{x}_i'\left(h\right)RE\left(h\right)\Bigg]$$

Fourth term in the bracket ,$2\left\langle W\left(h\right)\boldsymbol{\Delta}_n\left(h\right),\ X_n\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right\rangle$,is as follows;

$$2\left\langle W\left(h\right)\boldsymbol{\Delta}_n\left(h\right),\ X_n\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right\rangle$$

$$=2\left\langle W\left(h\right)\boldsymbol{\Delta}_n\left(h\right),\ X_n\left(h\right)\left(\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2}J_h^{-1}\varphi\left(y_j-\mathbf{x}_j'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_j\left(h\right)+RE\left(h\right)\right)\right\rangle$$

$$=\sum_{i=1}^{n} w_\tau\left(\mathbf{x}_i\left(h\right)\right)\Delta_i\left(h\right)\mathbf{x}_i'\left(h\right)\frac{1}{n}\left(J_h^{-1}\varphi\left(y_i-\mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_i\left(h\right)+RE\left(h\right)\right)$$

$$+\left(\sum_{i=1}^{n} w_\tau\left(\mathbf{x}_i\left(h\right)\right)\Delta_i\left(h\right)\mathbf{x}_i'\left(h\right)\cdot\frac{1}{n}\sum_{j\neq i}^{n} J_h^{-1}\varphi\left(y_j-\mathbf{x}_j'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_j\left(h\right)\right)$$

$$= \left[\frac{1}{n} \sum_{i=1}^{n} w_\tau \left(\mathbf{x}_i \left(h \right) \right) \Delta_i \left(h \right) \mathbf{x}_i' \left(h \right) J_h^{-1} \varphi \left(y_i - \mathbf{x}_i' \left(h \right) \boldsymbol{\beta}^* \left(h \right) \right) \mathbf{x}_i \left(h \right) \right.$$
$$+ \frac{1}{n} \sum_{1 \le i < j \le n}^{n} w_\tau \left(\mathbf{x}_i \left(h \right) \right) \Delta_i \left(h \right) \mathbf{x}_i' \left(h \right) J_h^{-1} \varphi \left(y_j - \mathbf{x}_j' \left(h \right) \boldsymbol{\beta}^* \left(h \right) \right) \mathbf{x}_j \left(h \right)$$
$$\left. + 2 \frac{1}{n} \sum_{i=1}^{n} w_\tau \left(\mathbf{x}_i \left(h \right) \right) \Delta_i \left(h \right) \mathbf{x}_i' \left(h \right) RE \left(h \right) \right]$$

So the bracket term on the righthand side is

$$\left[\frac{1}{n} \sum_{i=1}^{n} w_\tau \left(\mathbf{x}_i \left(h \right) \right) \Delta_i \left(h \right) \varphi \left(y_i - \mathbf{x}_i' \left(h \right) \boldsymbol{\beta}^* \left(h \right) \right) \mathbf{x}_i' \left(h \right) J_h^{-1} \mathbf{x}_i \left(h \right) \right.$$
$$- \frac{1}{n} \sum_{1 \le i < j \le n}^{n} w_\tau \left(\mathbf{x}_i \left(h \right) \right) \left(e_i - m \right) \mathbf{x}_i' \left(h \right) \cdot J_h^{-1} \varphi \left(y_j - \mathbf{x}_j' \left(h \right) \boldsymbol{\beta}^* \left(h \right) \right) \mathbf{x}_j \left(h \right)$$
$$\left. - 2 \sum_{i=1}^{n} \left(e_i - m \right) \mathbf{x}_i' \left(h \right) RE \left(h \right) \right]$$

Finally the $\hat{h}$ also achieves

$$\min_{h \in H_{pn}} \left[L_n \left(h \right) \right.$$
$$- 2 \frac{1}{n} \sum_{i=1}^{n} w \left(\mathbf{x}_i \left(h \right) \right) \left(e_i - m \right) \Delta_i \left(h \right)$$
$$+ \frac{1}{n^2} \sum_{i=1}^{n} w \left(\mathbf{x}_i \left(h \right) \right) \Delta_i \left(h \right) \varphi \left(y_i - \mathbf{x}_i' \left(h \right) \boldsymbol{\beta}^* \left(h \right) \right) \mathbf{x}_i' \left(h \right) J_h^{-1} \mathbf{x}_i \left(h \right)$$
$$- \frac{1}{n^2} \sum_{1 \le i < j \le n}^{n} w \left(\mathbf{x}_i \left(h \right) \right) \left(e_i - m \right) \mathbf{x}_i' \left(h \right) \cdot J_h^{-1} \varphi \left(y_j - \mathbf{x}_j' \left(h \right) \boldsymbol{\beta}^* \left(h \right) \right) \mathbf{x}_j \left(h \right)$$
$$\left. - 2 \frac{1}{n} \sum_{i=1}^{n} w \left(\mathbf{x}_i \left(h \right) \right) \left(e_i - m \right) \mathbf{x}_i' \left(h \right) RE \left(h \right) \right]$$

If we can show that the second term are negligible compared with $L_n \left(h \right)$ uniformly for any $h \in H_n$, then the asymptotic optimality property (3.3) is established for $\hat{h}$. If we can show that

$$\sup_{h \in \mathcal{H}} \frac{\left| \left\langle \mathbf{e}_n - m, \ W \left(h \right) \boldsymbol{\Delta}_n \left(h \right) \right\rangle \right|}{n R_n^2 \left(h \right)} \to 0 \tag{A.9}$$

$$\sup_{h\in\mathcal{H}} \frac{\left| \sum_{i=1}^{n} w\left(\mathbf{x}_i\left(h\right)\right) \Delta_i\left(h\right) \varphi\left(y_i - \mathbf{x}_i'\left(h\right) \boldsymbol{\beta}^*\left(h\right)\right) \mathbf{x}_i'\left(h\right) J_h^{-1} \mathbf{x}_i\left(h\right) \right|}{n^2 R_n^2\left(h\right)} \to 0 \tag{A.10}$$

$$\sup_{h\in\mathcal{H}} \frac{\left|\left\langle \mathbf{e}_n - m, , \ W\left(h\right) X\left(h\right) RE\left(h\right)\right\rangle\right|}{nR_n^2\left(h\right)} \to 0 \tag{A.11}$$

$$\sup_{h\in\mathcal{H}} \frac{\left| \sum_{1\le i<j\le n} w\left(\mathbf{x}_i\left(h\right)\right)\left(e_i - m\right) \mathbf{x}_i'\left(h\right) \cdot J_h^{-1} \varphi\left(y_j - \mathbf{x}_j^{\top}\left(h\right) \boldsymbol{\beta}^*\left(h\right)\right) \mathbf{x}_j\left(h\right) \right|}{n^2 R_n^2\left(h\right)} \to 0 \tag{A.12}$$

$$\sup_{h\in\mathcal{H}} \left| \frac{L_n^2\left(h\right)}{R_n^2\left(h\right)} - 1 \right| \to 0 \tag{A.13}$$

the asymptotic optimality property is established.

A.2.2. Proof of (*A.9*)

We shall prove (A.9) first. Pick any $\delta > 0$, by Chebyshev's inequality we have

$$P\left\{ \sup_{h\in H_{pn}} \frac{\frac{1}{n}\left|\left\langle \mathbf{e}_n - m, \ W\left(h\right)\Delta\left(h\right)\right\rangle\right|}{R_n^2\left(h\right)} > \delta \right\} \le \sum_{h\in H_{pn}} \frac{n^{-2}\mathbb{E}\left[\left\langle \mathbf{e}_n - m, \ W\left(h\right)\Delta\left(h\right)\right\rangle^2\right]}{\delta^2\left(R_n^2\left(h\right)\right)^2}$$

which, by martingale inequality of Dharmadhikari, Fabian and Jogdeo (68 annals), is no greater than

$$C\delta^{-2} \sum_{h\in H_{pn}} n^{-2} n^{\frac{2}{2}-1} \mathbb{E}\left[\sum_{i=1}^{n} \left\{ w\left(\mathbf{x}_i\left(h\right)\right)\left(q_i - \mathbf{x}_i'\left(h\right) \boldsymbol{\beta}\left(h\right)^*\right)\right\}^2 \right] \left(R_n^2\left(h\right)\right)^{-2}$$

for some constant $C > 0$. The last expression does not exceed $C'\delta^{-1}\sum_{h\in H_{pn}}\left(nR_n^2\left(h\right)\right)^{-1}$ for some constant C' , which tends to 0 by $(3.6)^{3,4}$.

A.2.3. Proof of (*A.10*)

Equation (A.10) can be established similar manner, denote

$$h\left(\mathbf{x}_i\right) = w\left(\mathbf{x}_i(h)\right)\left(q_i - \mathbf{x}_i'(h)\boldsymbol{\beta}^*\left(h\right) - m\right)\left(\mathbf{x}_i'\left(h\right) J_h^{-1} \operatorname{sgn}\left(y_i - \mathbf{x}_i'\left(h\right)\boldsymbol{\beta}^*\left(h\right)\right)\mathbf{x}_i\left(h\right)\right)$$

noting that , by assumption (3.4)(3.7),

$$\mathbb{E}\left[\left|h\left(\mathbf{x}\right)\right|^2\right] < \infty.$$

Given any $\delta > 0$, by Chebyshev's inequality we have

$$P\left\{\sup_{h\in\mathcal{H}}\frac{\frac{1}{n}\left|\frac{1}{n}\sum_{i=1}^{n}h\left(\mathbf{x}_i\right)\right|}{R_n^2\left(h\right)}>\delta\right\}\underset{\text{Chebyshev}}{\leq}\sum_{h\in\mathcal{H}}\frac{n^{-2}\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}h\left(\mathbf{x}_i\right)\right\}^2\right]}{\delta^2\left(R_n^2\left(h\right)\right)^2}.$$

It is well known inequality that

$$\sum_{h\in\mathcal{H}}\frac{n^{-2}\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}h\left(\mathbf{x}_i\right)\right\}^2\right]}{\delta^2\left(R_n^2\left(h\right)\right)^2}\leq\sum_{h\in\mathcal{H}}\frac{n^{-2}n\sum_{i=1}^{n}\frac{1}{n^2}\mathbb{E}\left[\left\{h\left(\mathbf{x}_i\right)\right\}^2\right]}{\delta^2\left(R_n^2\left(h\right)\right)^2}.$$

For some constant $C > 0$, the last expression does not exceed $C\delta^{-2}\sum_{h\in H_{pn}}\left(nR_n\left(h\right)\right)^{-2}$, which tends to 0.

A.2.4. Proof of (A.11)

Equation (A.11) can be established by the following way. First, by Markov's inequality, we have

$$P\left\{\sup_{h\in H_{pn}}\frac{\frac{1}{n}\left|\langle\mathbf{e}_n-m,,\ W\left(h\right)X\left(h\right)RE\left(h\right)\rangle\right|}{R_n^2\left(h\right)}>\delta\right\}$$
$$\underset{\text{Markov}}{\leq}\sum_{h\in H_{pn}}\frac{n^{-1}\mathbb{E}\left[\left|\left\langle\frac{1}{\sqrt{n}}W\left(h\right)X'\left(h\right)\left(\mathbf{e}_n-m\right),\ \sqrt{n}RE_n\left(h\right)\right\rangle\right|\right]}{\delta\left(R_n^2\left(h\right)\right)}.$$

Now since Cauchy-Schwarz inequality,

$$\sum_{h\in H_{pn}}\frac{\mathbb{E}\left[\left|\left\langle n^{-1/2}W\left(h\right)X'\left(h\right)\left(\mathbf{e}_n-m\right),\ n^{1/2}RE_n\left(h\right)\right\rangle\right|\right]}{\delta\left(nR_n^2\left(h\right)\right)}$$

[3] Since

$$w\left(\mathbf{x}_i\left(h\right)\right)=\frac{1}{2}\int_0^1 f_e\left(u\cdot\Delta_i\left(h\right)\mid\mathbf{x}\right)du$$
$$\sup_X\max_h w\left(\mathbf{x}_i\left(h\right)\right)<\infty$$

[4] Note that

$$\mathbb{E}\left[\sum_{i=1}^{n}\left\{w\left(\mathbf{x}_i\left(h\right)\right)\left(q_i-\mathbf{x}_i^\top\left(h\right)\boldsymbol{\beta}\left(h\right)^*\right)\right\}^2\right]=\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}\left[\varphi_\tau\left(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}^*\right)\mid X\right]^2\right]$$
$$\underset{\text{Jensen}}{\leq}\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}\left[\varphi_\tau\left(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}^*\right)^2\mid X\right]\right]$$
$$=n$$

$$\underset{\text{Cauchy-Schwarz}}{\leq} \sum_{h\in H_{pn}} \frac{\mathbb{E}\left[n\left\|RE_n(h)\right\|^2\right]^{\frac{1}{2}}\mathbb{E}\left[\frac{1}{n}\left\|W(h)X'(h)(\mathbf{e}-m)\right\|^2\right]^{\frac{1}{2}}}{\delta\left(nR_n^2(h)\right)}$$

and ,by assumption (3.5) and (2.9), we have

$$\mathbb{E}\left[n\left\|(RE_n(h))\right\|^2\right]^{\frac{1}{2}} \to 0$$

$$\mathbb{E}\left[\frac{1}{n}(\mathbf{e}_n-m)'X'(h)W^2(h)X(h)(\mathbf{e}_n-m)\right]^{\frac{1}{2}} < \infty.$$

Thus (A.11) is proved.

A.2.5. Proof of (A.12)

Equation (A.12) can be established as an application of U-statistics, denote

$$h\left(\mathbf{x}_i,\mathbf{x}_j\right) = \left\{w\left(\mathbf{x}_j(h)\right)\left(e_j-m\right)\mathbf{x}_j'(h)\right\}\left(J_h^{-1}\varphi\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}^*\right)\mathbf{x}_i(h)\right)$$

$$h^*\left(\mathbf{x}_i,\mathbf{x}_j\right) = \frac{1}{2}\left[h\left(\mathbf{x}_i,\mathbf{x}_j\right)+h\left(\mathbf{x}_j,\mathbf{x}_i\right)\right]$$

The corresponding U-statistic for (A.12) is obtained by

$$U_n = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} h^*\left(\mathbf{x}_i,\mathbf{x}_j\right).$$

Exact formula for the second moment of U_n may be stated as follows(Serfling 80). Writing

$$\begin{aligned} h_1^*\left(\mathbf{x}_i,\mathbf{x}_j\right) =&\mathbb{E}\left[\frac{1}{2}\left[h\left(\mathbf{x}_i,\mathbf{x}_j\right)+h\left(\mathbf{x}_j,\mathbf{x}_i\right)\right] \mid \mathbf{x}_i\right] \\ =&\frac{1}{2}\mathbb{E}\left[\left\{\left(e_j-m\right)\mathbf{x}_j'\right\}\left(J_h^{-1}\varphi\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}^*\right)\mathbf{x}_i\right) \mid \mathbf{x}_i\right] \\ &+\frac{1}{2}\mathbb{E}\left[\left\{\left(e_i-m\right)\mathbf{x}_i'\right\}\left(J_h^{-1}\varphi\left(y_j-\mathbf{x}_j'\boldsymbol{\beta}^*\right)\mathbf{x}_j\right) \mid \mathbf{x}_i\right] \\ =&\frac{1}{2}J_h^{-1}\varphi\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}^*\right)\mathbf{x}_i\mathbb{E}\left[\left(e_j-m\right)\mathbf{x}_j' \mid \mathbf{x}_i\right] \\ &+\frac{1}{2}\left\{\left(e_i-m\right)\mathbf{x}_i'\right\}\mathbb{E}\left[J_h^{-1}\varphi\left(y_j-\mathbf{x}_j'\boldsymbol{\beta}^*\right)\mathbf{x}_j \mid \mathbf{x}_i\right] \\ =&0 \end{aligned}$$

we have

$$\mathrm{Var}\left[U_n\right] = \frac{2}{n(n-1)}\left\{2(n-2)\mathbb{E}\left[h_1^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]+\mathbb{E}\left[h^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]\right\}.$$

We have, by assumption (3.4)(3.5)(3.7),

$$\mathbb{E}\left[h_1^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right] = 0$$

$$\mathbb{E}\left[h^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]\le\mathbb{E}\left[e_j^2\right]\mathbb{E}\left[\left\{\mathbf{x}_j'\mathbb{E}\left[f\left(\Delta_x\mid X\right)\mathbf{xx}'\right]^{-1}\mathbf{x}_i\right\}^2\right]$$
$$<\infty.$$

Then we can prove the equation (A.12). By Chebyshev's inequality we have

$$P\left(\sup_{h\in\mathcal{H}}\frac{\frac{1}{n}\left|2\frac{1}{n}\sum_{1\le i<j\le n}h\left(\mathbf{x}_i,\mathbf{x}_j\right)\right|}{R_n^2\left(h\right)}>\delta\right)$$
$$\underset{\text{Chebyshev}}{\le}\sum_{h\in\mathcal{H}}\frac{n^{-2}\left(n-1\right)^2\mathbb{E}\left[\left\{\frac{2}{n\left(n-1\right)}\sum_{1\le i<j\le n}h\left(\mathbf{x}_i,\mathbf{x}_j\right)\right\}^2\right]}{\delta^2\left(R_n^2\left(h\right)\right)^2}$$

which, by variance of U-statistic, is equivalent to

$$\sum_{h\in\mathcal{H}}\frac{n^{-2}\left(n-1\right)^2}{\delta^2R_n^2\left(h\right)}\mathrm{Var}\left[U\right]$$
$$=\sum_{h\in\mathcal{H}}\frac{n^{-2}\left(n-1\right)^2}{\delta^2R_n^2\left(h\right)}\frac{2}{n\left(n-1\right)}\left\{2\left(n-2\right)\mathbb{E}\left[h_1^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]+\mathbb{E}\left[h^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]\right\}$$
$$=\sum_{h\in\mathcal{H}}\frac{2n^{-3}\left(n-1\right)}{\delta^2R_n^2\left(h\right)}\mathbb{E}\left[h^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]$$

The first term does not exceed $\sum_{h\in\mathcal{H}}4\mu^2\left(\delta^2nR_n\left(h\right)\right)^{-1}$, which tends to 0 by (3.6).

A.2.6. Proof of (A.13)

(A.13) is as follow as

$$\sup_{h\in\mathcal{H}}\left|\frac{L_n^2\left(h\right)}{R_n^2\left(h\right)}-1\right|=\sup_{h\in\mathcal{H}}\left|\frac{L_n^2\left(h\right)-R_n^2\left(h\right)}{R_n^2\left(h\right)}\right|$$
$$=\sup_{h\in\mathcal{H}}\left|\frac{L_n^2\left(h\right)-R_n^2\left(h\right)}{\mathbb{E}\left[\left\|\mathbf{\Delta}_n\left(h\right)\right\|_{W(h)}^2\right]+\mathbb{E}\left[\left\|X\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right\|_{W(h)}^2\right]}\right|$$
$$\times\left|\frac{\mathbb{E}\left[\left\|\mathbf{\Delta}_n\left(h\right)\right\|_{W(h)}^2\right]+\mathbb{E}\left[\left\|X\left(h\right)\left(\hat{\boldsymbol{\beta}}\left(h\right)-\boldsymbol{\beta}^*\left(h\right)\right)\right\|_{W(h)}^2\right]}{R_n^2\left(h\right)}\right|$$

It is clear that (A.13) will follow from the following three statements:

$$\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\frac{\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}-\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]}{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}\right|\to 0 \qquad \text{(A.14)}$$

$$\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\frac{\begin{array}{c}\left\langle W(h)\,\boldsymbol{\Delta}_n(h),\ X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\rangle\\ -\ \mathbb{E}\left[\left\langle W(h)\,\boldsymbol{\Delta}_n(h),\ X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\rangle\right]\end{array}}{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}\right|\to 0 \qquad \text{(A.15)}$$

$$\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\frac{\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}-\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}\right|\to 0 \qquad \text{(A.16)}$$

A.2.7. Proof of (A.15)

(A.15) is established similar manner as (A.10), since

$$\begin{aligned}&\left\langle W(h)\Delta(h),\,X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\rangle-\mathbb{E}\left[\left\langle W(h)\,\Delta(h),\ X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\rangle\right]\\ =&\sum_i\left\{w_i(h)\,\Delta_i(h)\,\mathbf{x}_i'\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)-\mathbb{E}\left[w_i(h)\,\Delta_i(h)\,\mathbf{x}_i'(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right]\right\}\\ =&\frac{1}{n}\sum_{1\le i<j\le n}^{n}w_i(h)\,\Delta_i(h)\,\mathbf{x}_i'(h)\,J_h^{-1}\varphi\left(y_j-\mathbf{x}_j'(h)\,\boldsymbol{\beta}^*(h)\right)\mathbf{x}_j(h) \qquad \text{(A.17)}\\ &+\sum_i\left\{\Delta_i(h)\,\mathbf{x}_i'(h)\,RE(h)-\mathbb{E}\left[\Delta_i(h)\,\mathbf{x}_i'(h)\,RE(h)\right]\right\}\end{aligned}$$

The first term on the right hand side of (A.17) is established similar manner as (A.12). Denote

$$h\left(\mathbf{x}_i,\mathbf{x}_j\right)=\left\{w_i(h)\,\Delta_i(h)\,\mathbf{x}_i'(h)\right\}\left(J_h^{-1}\varphi\left(y_j-\mathbf{x}_j'(h)\,\boldsymbol{\beta}^*(h)\right)\mathbf{x}_j(h)\right)$$

$$h^*\left(\mathbf{x}_i,\mathbf{x}_j\right)=\frac{1}{2}\left[h\left(\mathbf{x}_i,\mathbf{x}_j\right)+h\left(\mathbf{x}_j,\mathbf{x}_i\right)\right]$$

Exact formula for the second moment of U_n may be stated as follows(Serfling 80). Writing

$$\begin{aligned}h_1^*\left(\mathbf{x}_i,\mathbf{x}_j\right)=&\mathbb{E}\left[\frac{1}{2}\left[h\left(\mathbf{x}_i,\mathbf{x}_j\right)+h\left(\mathbf{x}_j,\mathbf{x}_i\right)\right]\mid\mathbf{x}_i\right]\\ =&\frac{1}{2}w_i(h)\,\Delta_i(h)\,\mathbf{x}_i'(h)\,\mathbb{E}\left[J_h^{-1}\varphi\left(y_j-\mathbf{x}_j'(h)\,\boldsymbol{\beta}^*(h)\right)\mathbf{x}_j(h)\mid\mathbf{x}_i\right]\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\mathbb{E}\left[w_j(h)\,\Delta_j(h)\,\mathbf{x}_j'(h)\mid\mathbf{x}_i\right]J_h^{-1}\varphi\left(y_i-\mathbf{x}_i'(h)\,\boldsymbol{\beta}^*(h)\right)\mathbf{x}_i(h)\\
=&0\\
\mathbb{E}\left[h_1^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]=&0
\end{aligned}
$$

we have

$$
\operatorname{Var}\left[U_n\right]=\frac{2}{n\,(n-1)}\left\{2\,(n-2)\,\mathbb{E}\left[h_1^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]+\mathbb{E}\left[h^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]\right\}.
$$

By Chebyshev's inequality,

$$
P\left(\sup_{h\in\mathcal{H}}\frac{\frac{1}{n}\left|2\frac{1}{n}\sum_{1\leq i<j\leq n}h\left(\mathbf{x}_i,\mathbf{x}_j\right)\right|}{R_n^2(h)}>\delta\right)\leq\sum_{h\in\mathcal{H}}\frac{\frac{2n^{-2}(n-1)^2}{n\,(n-1)}\mathbb{E}\left[\left[\left\{\sum_{1\leq i<j\leq n}h\left(\mathbf{x}_i,\mathbf{x}_j\right)\right\}^2\right]\right]}{\delta^2\left(R_n^2(h)\right)^2}
$$

which, by variance of U-statistic, is equivalent to

$$
\begin{aligned}
&\sum_{h\in\mathcal{H}}\frac{n^{-2}(n-1)^2}{\delta^2\left(R_n^2(h)\right)^2}\operatorname{Var}[U]\\
=&\sum_{h\in\mathcal{H}}\frac{n^{-2}(n-1)^2}{\delta^2\left(R_n^2(h)\right)^2}\frac{2}{n\,(n-1)}\left\{2\,(n-2)\,\mathbb{E}\left[h_1^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]+\mathbb{E}\left[h^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]\right\}\\
=&\sum_{h\in\mathcal{H}}\frac{2n^{-3}(n-1)}{\delta^2R_n^2(h)}\mathbb{E}\left[h^*\left(\mathbf{x}_i,\mathbf{x}_j\right)^2\right]
\end{aligned}
$$

The first term does not exceed $\sum_{h\in\mathcal{H}}4\left(\delta^2nR_n(h)\right)^{-1}$, which tends to 0 by (3.6).

The second term on the right hand side of (A.17) is

$$
\begin{aligned}
&P\left\{\sup_{h\in\mathcal{H}}\frac{1}{n}\frac{\left|\left\langle W(h)\,\boldsymbol{\Delta}_n(h)\,,\ X(h)\,RE(h)\right\rangle\right|}{\mathbb{E}\left[\left\|\boldsymbol{\Delta}_n(h)\right\|_{W(h)}^2\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|_{W(h)}^2\right]}>\delta\right\}\\
\underset{\text{Markov}}{\leq}&\sum_{h\in H_n}\frac{\mathbb{E}\left[\left|\left\langle W(h)\,\boldsymbol{\Delta}_n(h)\,,\ X(h)\,RE(h)\right\rangle\right|\right]}{\delta\left(\mathbb{E}\left[\left\|\boldsymbol{\Delta}_n(h)\right\|_{W(h)}^2\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|_{W(h)}^2\right]\right)}\\
\underset{\text{Cauchy-Schwarz}}{\leq}&\sum_{h\in H_n}\frac{\mathbb{E}\left[\left\|W(h)\,\boldsymbol{\Delta}_n(h)\right\|^2\right]^{1/2}\mathbb{E}\left[\left\|X(h)\,RE(h)\right\|^2\right]^{1/2}}{\delta\left(\mathbb{E}\left[\left\|\boldsymbol{\Delta}_n(h)\right\|_{W(h)}^2\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|_{W(h)}^2\right]\right)}
\end{aligned}
$$

$$\leq \sum_{h\in H_n} C\frac{1}{\delta}\left(\frac{\mathbb{E}\left[\|X(h)\,RE(h)\|^2\right]}{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}\right)^{1/2}$$

$$\to 0$$

A.2.8. Proof of (A.14)(A.16)

(A.14)(A.16) are by application of Whittle(60)

$$P\left\{\sup_{h\in\mathcal{H}}\frac{1}{n}\left|\frac{\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}-\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]}{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}\right|>\delta\right\}$$

$$\underset{\text{Chebyshev}}{\leq}\sum_{h\in H_n}\frac{1}{n^2}\frac{\mathbb{E}\left[\left\{\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}-\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]\right\}^2\right]}{\left(\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right)^2\delta^2}$$

$$\underset{\text{Whittle}}{\leq} C\sum_{h\in H_n}\frac{1}{n^2}\frac{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W^2(h)}\right]}{\left(\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right)^2\delta^2}$$

$$\leq C'\sum_{h\in H_n}\frac{1}{n}\frac{1}{\left(\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right)\delta^2}$$

$$\to 0$$

and

$$P\left\{\sup_{h\in\mathcal{H}}\frac{1}{n}\frac{\left|\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|-\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right|}{\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]}>\delta\right\}$$

$$\underset{\text{Chebyshev}}{\leq}\sum_{h\in H_n}\frac{1}{n^2}\frac{\mathbb{E}\left[\left\{\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}-\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right\}^2\right]}{\left(\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right)^2\delta^2}$$

$$\underset{\text{Whittle}}{\leq} C\sum_{h\in H_n}\frac{1}{n^2}\frac{\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W^2(h)}\right]}{\left(\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|^2_{W(h)}\right]+\mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h)-\boldsymbol{\beta}^*(h)\right)\right\|^2_{W(h)}\right]\right)^2\delta^2}$$

$$\le C' \sum_{h \in H_n} \frac{1}{\left(\mathbb{E}\left[\|\boldsymbol{\Delta}_n(h)\|_{W(h)}^2\right] + \mathbb{E}\left[\left\|X(h)\left(\hat{\boldsymbol{\beta}}(h) - \boldsymbol{\beta}^*(h)\right)\right\|_{W(h)}^2\right]\right)\delta^2}$$
$$\to 0$$

REFERENCES

Akaike, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle," *Second International Symposium on Information Theory*, 267–281.

Andrews, D. W. (1999): "Consistent Moment Selection Procedures for Generalized Method of Moment Estimaton," *Econometrica*, 67, 543–564.

Angrist, J., V. Chernozhukov, and I. Fernandez-Val (2006): "Quantile Regression under Misspecification, with an Application to the US Wage Structure," *Econometrica*, 74, 539–563.

Belloni, A. and V. Chernozhukov (2011): "1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, 39, 82–130.

Burman, P. and D. Nolan (1995): "A general Akaike-type criterion for model selection in robust regression," *Biometrika*, 82, 877–886.

Chamberlain, G. (1994): "Quantile Regression, Censoring, and the structure of Wages," *Advances in Econometrics, Sixth World Congress*, 1, 171–209.

Fan, J. and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, 96, 1348–1360.

Fox, M. and H. Rubin (1964): "Admissibility of Quantile Estimation of a Single Location Parameter," *The Annals of Mathematical Statistics*, 35, 1019–1030.

Giacomini, R. and I. Komunjer (2005): "Evaluation and Combination of Conditional Quantile Forcasts," *Journal of Business and Economic Statistics*, 23, 416–431.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, 50, 1029–1054.

Li, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

Mallows, C. (1973): "Some Comments on C_p," *Technometrics*, 15, 661–675.

Rissanen, J. (1986): "Stochastic Complexity and modeling," *The Annals of Statistics*, 14, 1080–1100.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

——— (2011): "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 273–282.

White, H. (1980): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149–170.