ASYMPTOTICALLY EFFICIENT MEDIAN REGRESSION IN THE PRESENCE OF HETEROSCEDASTICITY OF UNKNOWN FORM

by

QUANSHUI ZHAO†

City University of Hong Kong

ABSTRACT

We consider a linear model with heteroscedasticity of unknown form. Using Stone's (1977) k-nearest neighbors (k-NN) approach, the heteroscedastic relative scale of the disturbances can be estimated by weighted mean of absolute residuals based on preliminary estimation and is shown to be consistent. The weighted least absolute deviation, or median regression estimator with the estimated weights is shown to be asymptotically equivalent to the pseudo-estimator using the true, but unknown, weights. A Bahadur representation and the consequent asymptotic normality of the estimator are established under mild moment conditions. In Monte-Carlo experiments, the proposed method is shown to consistently outperform the generalized least squares (k-NN) estimator of Robinson (1987) as well as the one-step weighted median regression estimator of Newey and Powell (1990).

KEYWORDS: Least median estimation; Linear model; Heteroscedasticity; Bahadur representation; Efficiency; Nearest neighbor estimation.

Economic literature classification code: C14, C21

† I would like to express my appreciation to Roger Koenker and Alain Monfort for their constructive comments on earlier drafts of the paper. This research was partially supported by a strategic grant in the City University of Hong Kong.

Short running head:

Efficient Median Regression

1. INTRODUCTION

In the presence of heteroscedastic disturbances, the median regression estimator, and related quantile regression methods though consistent, are inefficient in much the same way as the familiar least squares estimator is known to be. Koenker and Zhao (1994) examine the large sample properties of regression quantiles, including median regression as the leading special case, for a popular class of linear heteroscedastic models when the conditional scale of the disturbance is a parametric function of the design variables. They provide a Bahadur representation for weighted regression quantiles under mild conditions, and show that a fully efficient estimator can be constructed by weighting. Here we extend these results to study linear regression models with heteroscedasticity of unknown form.

In the sequel, we adopt the problem setting of Robinson (1987), who studies the analogous problem for the least-squares estimator. Suppose (X, Y) is a random vector in $R^d \times R$ Euclidian space, where

$$Y = X'\beta + \sigma(X)u, \tag{1.1}$$

and u is independent of X and has a distribution function F, and $\sigma(X) > 0$ is an unknown Borel function of X. This model structure is representative of many applications. The linear heteroscedastic models discussed in Gutenbrunner and Jurečková (1992) and Koenker and Zhao (1994) constitute one example, the autoregressive conditionally heteroscedastic (ARCH) models proposed by Engle (1982) are another. With these cases, the heteroscedasticity is of known parametric form and estimation of the parameters is relatively easy.

There is a large literature on the parametric specification of $\sigma(X)$. For example, Carroll and Ruppert (1982) discuss the problem when $\sigma(X)$ is a function of the conditional mean of Y. Special functional types such as the power, polynomial and exponential functions of location can be used. For parametric forms of heteroscedasticity, maximum likelihood estimation can be relied upon when the error distribution is known (see e.g. Carroll and Ruppert (1982)). But two step methods are often suggested. For the linear heteroscedastic model studied by Koenker and Zhao (1994) and Gutenbrunner and Jurečková (1992), $\sigma(X)$ is assumed to be a linear function of the regression variables, $\sigma(X) = X'\gamma$, where γ is a d-vector, which is usually fitted by regressing the absolute residuals from initial estimation on X. Weighted regression is then conducted as the second step to achieve a better and more efficient estimator.

However, for heteroscedasticity of unknown functional form we must resort to some form of nonparametric estimation. Given n independent observations $\{(X_i, Y_i), 1 \leq i \leq n\}$, we would like to estimate the parameter β efficiently without prejudging the parametric form of $\sigma(X)$. Under mild regularity conditions, essentially those required to ensure the existence of the matrices specified below, the median regression estimator

$$\check{\beta} = arg \ \min_b \sum_1^n |Y_i - X_i'b| \tag{1.2}$$

is consistent, and asymptotically normal. Gutenbrunner and Jurečková (1992) show that $\sqrt{n}(\check{\beta} - \beta) \rightarrow_{\mathcal{L}} N(0, Q_1^{-1}Q_0Q_1^{-1}/(4f^2(0))$, where $Q_r = E(XX'/\sigma^r(X))$ and $f(u) = F'(u)$ is the density function of u. A Bahadur representation of $\check{\beta}$ can also be obtained using results from Koenker and Zhao (1994).

Given the true values of $\sigma_i = \sigma(X_i)$, consider a weighted estimate of β defined by

$$\tilde{\beta} = arg \ \min_b \sum_1^n \sigma_i^{-1}|Y_i - X_i'b|. \tag{1.3}$$

Under mild conditions it can be shown that $\sqrt{n}(\tilde{\beta} - \beta) \rightarrow_{\mathcal{L}} N(0, Q_2^{-1}/(4f^2(0))$, and the matrix inequality $Q_1^{-1}Q_0Q_1^{-1} \geq Q_2^{-1}$ implies that the weighted estimator has asymptotically smaller covariance matrix than the unweighted one and therefore is more accurate. The asymptotic variance of $\tilde{\beta}$ is shown to be the minimum among the weighted median regression estimators by Newey and Powell (1990).

To estimate the scale function $\sigma(X)$, we apply Stone's (1977) k nearest neighbors (k-NN) smoothing method to the absolute residuals obtained from preliminary estimation of the model (1.1) and then estimate $\sigma(X)$ as a weighted mean of absolute residuals. Suppose W_{ij} $(i, j = 1, 2, ..., n)$ is the k-NN weight function defined by Stone (1977, p599) or Robinson (1987, p878) satisfying

$$\sum_{j=1}^n W_{ij} = 1, \quad W_{ii} = 0, \ W_{ij} \geq 0, \ i = 1, ..., n. \tag{1.4}$$

The estimate of the conditional scale of the heteroscedastic disturbance at X_i is defined as

$$\hat{\sigma}_i = \hat{\sigma}(X_i) = \sum_{j=1}^n W_{ij}|\hat{\epsilon}_j|, \tag{1.5}$$

where $\hat{\epsilon}_j = Y_j - X_j'\check{\beta}$, $j = 1, 2, ..., n$. $\hat{\sigma}_i$ is a consistent estimator of σ_i.

Following Koenker and Zhao (1994), the weighted median regression estimator is defined as

$$\hat{\beta} = arg \min_b \sum_1^n \hat{\sigma}_i^{-1} |Y_i - X_i' b|. \tag{1.6}$$

As with weighted least squares procedures, one may wish to iterate the procedure to refine the weights and the estimator. Our main theorem establishing the asymptotic efficiency of the resulting weighted estimator is given in Section 2, with a brief sketch of its proof. The full proof and related results are collected in Appendices A and B. Results from a Monte Carlo experiment are reported in Section 3.

2. MAIN RESULTS

In this section, we examine the regression model defined in (1.1) under the following assumptions:

C1. The random variable $\{u_i\}$, independent of the design variables X, have distribution function F, with $F^{-1}(1/2) = 0$ and density function f continuous at the origin with $f(0) > 0$ and $\| f \|_\infty < \infty$, and $E|u| = 1$ and $E|u|^p < \infty$ for some $p > 2$.

C2. $P(\sigma(X) \geq \delta) = 1$, for some $\delta > 0$.

C3. $E \| X \|^p < \infty$, $E|\sigma(X)|^p < \infty$ for some $p > 2$.

C4. $E(XX'/\sigma^2(X))$ is positive definite.

C5. $\max_{i,j} W_{ij} = \mathcal{O}(k^{-1})$, as $k \to \infty$.

C6. $k/n \to 0$, $nk^{-p/2} \to 0$ for $p > 2$.

Here the monent conditions assumed are substantially weaker than the analogous conditions employed by Robinson (1987) for the related least squares estimation problem. We basically require p-th finite moments ($p > 2$) for X, the scale function $\sigma(X)$ and u. In contrast, at least finite fourth moments are required for $\sigma(X)$ and u for the least squares case. The moment condition (C1) on u is required for consistent weights estimation using the mean absolute values of the residuals. It may be removed if some trimming scheme could be used to screen out extreme residuals.

THEOREM *Under conditions C1-C6, the weighted estimator defined in (1.6) has Bahadur representation*

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{Q_2^{-1}}{f(0)} n^{-1/2} \sum_1^n \sigma_i^{-1} X_i \psi(u_i) + o_p(1), \tag{2.1}$$

where $\psi(u) = 0.5 - I(u < 0)$ *and* $Q_2 = E(XX'/\sigma^2(X))$.

Corollary. *Under the conditions of the theorem, the estimator has asymptotic normal distribution*

$$\sqrt{n}(\hat{\beta} - \beta) \to_{\mathcal{L}} N\left(0, Q_2^{-1}/(4f^2(0))\right).$$

The variance obtained in the Corollary is the minimum that can be achieved over the class of weighted least absolute deviations estimators as explained by Newey and Powell (1990). The asymptotic properties obtained here can be used to construct improved tests for the regression parameters and their contrasts either using direct Wald type tests, the rank tests of Gutenbrunner et al (1994) or the resampling ideas like those of Parzen, Wei and Ying (1994).

It may be noted that because (1.1) takes the form of a location-scale model, for which the disturbance is a scale shift of an i.i.d. innovation sequence the k-NN weighted mean of the absolute residuals gives a consistent estimate of the optimal scale $\sigma(X)$ of the disturbance term. In the general case, the optimal weight is the conditional density function of the disturbance at the median, $f(0|X) = f_u(0|X)/\sigma(X)$, as discussed in Newey and Powell (1990). Without the assumption of scale shift, $f_u(0|X)$ would not be a constant. The optimal weights, therefore, would have to be estimated in such a way that they provide a local estimate of the conditional density of the response at the median.

Newey and Powell (1990) investigate a class of censored regression models, of which the model (1.1) is included. The estimator constructed by Newey and Powell (1990) is, in fact, a one-step Huber estimator (see Bickel (1975)), which is substantially different from our estimator defined in (1.6). The one-step method is based on the idea that the efficient estimator of weighted LAD regression is the solution of the asymptotic first order condition

$$n^{-1/2} \sum_1^n S(x_i, \tilde{\beta}) = n^{-1/2} \sum_1^n w_i x_i \psi(y_i - x_i'\tilde{\beta}) = o_p(1)$$

where $w_i = f(0|x_i)$ is the optimal weight. Given an initial $\sqrt{n}$-consisten estimator $\check{\beta}$, following Bickel (1975), a one-step Huber estimator in the form

$$\check{\beta}^* = \check{\beta} + \left[\sum_1^n S(x_i, \check{\beta}) S(x_i, \check{\beta})'/n\right] \sum_1^n S(x_i, \check{\beta})/n$$

would be asymptotically efficient under suitable regularity conditions. However, the score function S contains unknown weights, which have to be estimated. Newey and Powell (1990) propose to estimate $g(x) = \| x \| \; f(0|x)$ by Stone's (1977) k-NN method, combined with kernel density estimation

$$\hat{g}(x, \beta) = \sum_1^n W_{ni}(x) \| x_i \| \; 1(-h_n \leq y_i - x_i - x_i'\beta \leq 0)/h_n,$$

where $\{h_n\}$ is a sequence of "window width" that decrease to zero as sample size increases to infinity, and $W_{ni}(x)$ is the k-NN weights. To assure the efficiency of the one-step estimator, Shick's (1986) sample splitting device is adopted. Newey and Powell's (1990) method can be described by the following steps:

(1) Split the sample into two subsets indexed by $I_1 = \{1, 2, ..., n_1\}$ and $I_2 = \{n_1 + 1, ..., n\}$ with $n_1 = [n/2]$. For each sub-sample, compute the LAD estimate $\tilde{\beta}_j$ and then $\hat{g}_j(x, \beta)$ (j=1,2).

(2) Define the score S as

$$\hat{S}(x_i) = \hat{g}_j(x_i, \tilde{\beta}_j)(x_i / \| x_i \|)\psi(y_i - x_i'\tilde{\beta}_j), \; if \; i \notin I_j, \; j = 1, 2$$

The efficient estimator is thus

$$\hat{\beta} = (\tilde{\beta}_1 + \tilde{\beta}_2)/2 + \left[\sum_1^n \hat{S}(x_i)\hat{S}(x_i)'\right]^{-1} \sum_1^n \hat{S}(x_i)/n.$$

From the definition, it is clear that the one-step estimator is a linear approximation of the optimal solution of the first order equation by using an initial estimate. While the estimator (1.6) in this paper is the actual solution of the first order equation. The assumption of scale shift of disturbance is stronger, but the resulted estimator is much simpler and it may be expected to deliver better finite sample performance in the location scale model. Monte Carlo simulation results in Section 3 validate this conjecture.

In the following, we outline the proofs and collect the details in the appendices. Consider the function

$$V(\Delta, h) = n^{-1/2} \sum_{1}^{n} h_i^{-1} X_i \psi(u_i - n^{-1/2} \sigma_i^{-1/2} X_i' \Delta), \tag{2.2}$$

where $h = (h_1, h_2, ..., h_n) > 0$. It is easy to see that the derivative of the objective function used in the median regression is linear in V, as

$$\sum_{1}^{n} \hat{\sigma}_i^{-1} X_i \psi(Y_i - X_i' b) = n^{1/2} V(\Delta, \hat{\sigma}),$$

where $\Delta = \sqrt{n}(b - \beta)$. For $\hat{\Delta} = \sqrt{n}(\hat{\beta} - \beta)$, it follows from Ruppert and Carroll (1980, Lemma A.2, p836) that

$$\| V(\hat{\Delta}, \hat{\sigma}) \| = o_p(1). \tag{2.3}$$

The following lemma is crucial.

LEMMA 2.1 *Under the conditions C1-C6, for any* $0 < L < \infty$,

$$\sup_{\|\Delta\| \leq L} \| V(\Delta, \hat{\sigma}) - V(\Delta, \sigma) \| = o_p(1).$$

This lemma says that the difference between $V(\Delta, \hat{\sigma})$ and $V(\Delta, \sigma)$ is very small, in probability term. While applying Ruppert and Carroll's (1980, p836) Lemma A.3 to $V(\Delta, \sigma)$, one has $\sup_{\|\Delta\| \leq L} \| V(\Delta, \sigma) - V(0, \sigma) + f(0) Q_2 \Delta \| = o_p(1)$. Therefore $V(\Delta, \hat{\sigma})$ can be represented by a linear function of Δ as

$$\sup_{\|\Delta\| \leq L} \| V(\Delta, \hat{\sigma}) - V(0, \sigma) + f(0) Q_2 \Delta \| = o_p(1). \tag{2.4}$$

As mentioned earlier, $V(\Delta, \hat{\sigma})$ is the derivative of the convex objective function of the weighted median regression, the derivative of which on any direction Δ, $-\Delta' V(\lambda \Delta, \hat{\sigma})$, is monotone increasing in λ for $-\infty < \lambda < \infty$. It follows that $\Delta' V(\lambda \Delta, \hat{\sigma}) \leq \Delta' V(\Delta, \hat{\sigma})$ for any $\lambda \geq 1$. Applying Koenker and Zhao's (1996, p809) Lemma A.4 combined with (2.3) and (2.4), we have $\hat{\Delta} = \mathcal{O}_p(1)$ and the representation (2.1). The corollary follows directly from the central limit theorem.

3. Monte Carlo Results

In this section we report on a Monte Carlo experiment designed to examine the performance of several different weighted estimation methods for a linear heteroscedastic model proposed in Carroll and Ruppert (1982).

Consider the linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \sigma(X_{1i}, X_{2i})\ u_i,\ \ i = 1, 2, ..., N, \tag{3.1}$$

where the regressors X_1 and X_2 are generated from

$$X_1 = U + .2V,\ \ X_2 = .2U + V,\ \ U \sim N(5, 9),\ \ V \sim U(0, 4).$$

The scale function σ is of the form

$$\sigma(X_1, X_2) = \alpha_1 exp(\alpha_2 |\tau(X_1, X_2)|),\ \ \tau(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{3.2}$$

We consider two error distributions for u_i, the standard normal and Student's t(3), with the latter representing disturbances of heavy tails.

Following Carroll and Ruppert (1982), the parameters in the model are chosen as $(\beta_0, \beta_1, \beta_2) = (10, -4, 2)$, and $(\alpha_1, \alpha_2) = (10, .1)$. We generate samples of size $n = 100$ and $n = 300$ from model (3.1-2), with 500 replications for each parameter specification. Mean square errors (MSE) of the estimators for $\beta_0, \beta_1, \beta_2$ are evaluated using the methods follows. The simulation was run on a Sun Ultra using Splus. Four estimators were compared:

(1) Robinson's (1987) generalized least squares method (GLS). To estimate the weightings, roughly 40% of the observations are used in k-NN estimation.

(2) Least squares (LS) and weighted least squares with true weights (WLSTW) methods. The true weights are known from the data generating process. We use the performance of LS and WLSTW to evaluate the effectiveness of the generalized least squares, to gauge the efficiency gain by weighting and the cost of using estimated weights.

(3) Newey and Powell's (1990) one-step method (NP's one-step). To estimate the g function specified by Newey and Powell (1990, p302), the band width h_n is chosen as half of the estimated standard deviation of the residuals from preliminary estimation.

Table 1. The Relative MSE Ratios to the MSE of L_1 Method for Linear Models with heteroscedastic disturbances

Methods (n=100)	N(0,1) Errors β_0	β_1	β_2	t(3) Errors β_0	β_1	β_2
LS	9.95	7.99	5.98	18.36	15.26	15.00
GLS	2.30	1.97	1.46	6.49	5.08	5.07
WLSTW	0.42	0.41	0.42	0.95	0.88	0.95
NP's one-step	1.02	0.98	1.51	1.48	1.01	2.50
L1	1.00	1.00	1.00	1.00	1.00	1.00
New WL1	0.75	0.72	0.82	0.80	0.72	0.82
WL1TW	0.61	0.59	0.69	0.67	0.61	0.71
MSE of L1	59.29	5.31	3.61	76.20	7.13	4.49

Methods (n=300)	N(0,1) Errors β_0	β_1	β_2	t(3) Errors β_0	β_1	β_2
LS	10.23	8.07	7.33	25.12	20.11	17.71
GLS	1.97	1.70	1.35	8.16	7.59	7.15
WLSTW	0.45	0.43	0.46	1.06	1.08	1.06
NP's one-step	0.93	0.82	1.33	1.19	0.87	2.03
L1	1.00	1.00	1.00	1.00	1.00	1.00
New WL1	0.81	0.75	0.81	0.75	0.70	0.78
WL1TW	0.69	0.65	0.74	0.62	0.59	0.71
MSE of L1	17.21	1.72	1.05	21.35	2.08	1.38

(4) Finally, the reweighted median regression estimator (1.6) (New WL1) proposed in this article. We use the LAD as the preliminary estimator. Similar to the LS case, the LAD (L1) and weighted LAD using true weightings (WL1TW) are provided for comparison purposes.

The mean square errors (MSE) are often used as the criterion for evaluating estimator performances. To make the table more readable, we only report the MSEs of LAD estimator and the MSEs of other estimators are reported as a ratio to the MSE of LAD. Clearly, the smaller the MSE ratio, the better the method performs.

Table 1 reports the relative MSE ratios of various methods for two error distributions, the standard normal and Student's t(3). We can see that the least squares method performs the worst when heteroscedasticity exists. While the GLS makes substantial improvement, but still underperforms the benchmark method, LAD, in all the simulated cases including the case when the errors are conditional Gaussian.

Newey and Powell's (1990) one-step method does not perform well for small sample size cases. When $n = 100$, its efficiency is even lower than the LAD, even though the initial estimator is LAD. This shows that the one-step method does not make any improvement when sample size is small. The performance improves when sample size is increased to $n = 300$, but still does not outperform LAD for all the parameters.

The proposed weighted median regression method is quite successful in all the simulated cases. On average the MSEs of all the parameters are 20% smaller than those of LAD and are quite close to the MSEs when the true weights are used. Table 1 gives a clear illustration which confirms the analysis in Section 2.

APPENDIX A: PROOF OF LEMMA 2.1

To prove Lemma 2.1, we introduce the following notation. Define the n-vectors $\tilde{\sigma}$ and $\bar{\sigma}$ with attributes

$$\tilde{\sigma}_i = \sum_1^n W_{ij}|\epsilon_j|, \quad \bar{\sigma}_i = \sum_1^n W_{ij}\sigma_j, \quad 1 \le i \le n,$$

where $\epsilon_j = \sigma_j u_j = Y_j - X_j'\beta$. Since

$$\| V(\Delta,\hat{\sigma}) - V(\Delta,\sigma) \| \le \| V(\Delta,\hat{\sigma}) - V(\Delta,\tilde{\sigma}) \| + \\ \| V(\Delta,\tilde{\sigma}) - V(\Delta,\bar{\sigma}) \| + \| V(\Delta,\bar{\sigma}) - V(\Delta,\sigma), \|$$

Lemma 2.1 can be decomposed into three parts, which will be proved one by one.

LEMMA A.1 $\sup_{||\Delta|| \le L} \| V(\Delta,\hat{\sigma}) - V(\Delta,\tilde{\sigma}) \| = o_p(1)$ *for* $0 < L < \infty$.

LEMMA A.2 $\sup_{||\Delta|| \le L} \| V(\Delta,\tilde{\sigma}) - V(\Delta,\bar{\sigma}) \| = o_p(1)$ *for* $0 < L < \infty$.

LEMMA A.3 $\sup_{||\Delta|| \le L} \| V(\Delta,\bar{\sigma}) - V(\Delta,\sigma) \| = o_p(1)$ *for* $0 < L < \infty$.

PROOF OF LEMMA A.1: Since $\psi(u)$ is bounded by 0.5,

$$\begin{aligned}
&\sup_{\|\Delta\|\leq L} \| V(\Delta,\hat{\sigma}) - V(\Delta,\tilde{\sigma}) \| \\
&\leq (\min_i \hat{\sigma}_i \cdot \min_i \tilde{\sigma}_i)^{-1} n^{-1/2} \sum_1^n |\hat{\sigma}_i - \tilde{\sigma}_i| \ \| X_i \| \\
&\leq \mathcal{O}_p(1) \left[\sum_1^n (\hat{\sigma}_i - \tilde{\sigma}_i)^2 n^{-1} \sum_1^n \| X_i \|^2 \right]^{1/2} = o_p(1),
\end{aligned}$$

by Lemma B.2 and condition C4. □

PROOF OF LEMMA A.2: For any numbers $x > 0$, $y > 0$,

$$x^{-1} - y^{-1} = (y-x)y^{-2} + (y-x)^2 x^{-1} y^{-2}.$$

Hence

$$\frac{1}{\tilde{\sigma}_i} - \frac{1}{\bar{\sigma}_i} = \frac{\bar{\sigma}_i - \tilde{\sigma}_i}{\bar{\sigma}_i^2} + \frac{(\bar{\sigma}_i - \tilde{\sigma}_i)^2}{\tilde{\sigma}_i \bar{\sigma}_i^2}.$$

Using this formula, $V(\Delta,\tilde{\sigma}) - V(\Delta,\bar{\sigma})$ can be represented as the summation of the following two components:

$$\begin{aligned}
I_1 &= n^{-1/2} \sum_1^n \frac{\bar{\sigma}_i - \tilde{\sigma}_i}{\bar{\sigma}_i^2} X_i \psi(u_i - n^{-1/2}\sigma_i^{-1} X_i' \Delta), \\
I_2 &= n^{-1/2} \sum_1^n \frac{(\bar{\sigma}_i - \tilde{\sigma}_i)^2}{\tilde{\sigma}_i \bar{\sigma}_i^2} X_i \psi(u_i - n^{-1/2}\sigma_i^{-1} X_i' \Delta).
\end{aligned}$$

For the second term, using Jensen's inequality and the boundedness of ψ, we have

$$\begin{aligned}
\sup_{\|\Delta\|\leq L} \| I_2 \|^2 &\leq n^{-1} \sum_1^n \| X_i \|^2 \ \sum_1^n \left[\frac{(\bar{\sigma}_i - \tilde{\sigma}_i)^2}{\tilde{\sigma}_i \bar{\sigma}_i^2} \right]^2 \\
&\leq \mathcal{O}_p(1) (\min_i \tilde{\sigma}_i \min_i \bar{\sigma}_i^2)^{-2} \max_i |\tilde{\sigma}_i - \bar{\sigma}_i|^{(4-p)} \sum_1^n |\tilde{\sigma}_i - \bar{\sigma}_i|^p = o_p(1),
\end{aligned}$$

by Lemma B.3 and B.4. Here we used the fact that $p > 2$, but less than 4.

The term I_1, using the inequality

$$|\psi(u_i - n^{-1/2}\sigma_i^{-1} X_i' \Delta) - \psi(u_i)| \leq 1(|u_i| \leq n^{-1/2} \| X_i \| L/\delta)$$

for $\| \Delta \| \leq L$, can be decomposed into two parts, $\sup_{\|\Delta\|\leq L} \| I_1 \| \leq \| I_3 \| + I_4$, where

$$\begin{aligned}
I_3 &= n^{-1/2} \sum_1^n \frac{\bar{\sigma}_i - \tilde{\sigma}_i}{\bar{\sigma}_i^2} X_i \psi(u_i), \\
I_4 &= n^{-1/2} \sum_1^n \frac{|\bar{\sigma}_i - \tilde{\sigma}_i|}{\bar{\sigma}_i^2} \| X_i \| \ \eta_j, \quad \eta_j = 1(|u_i| \leq n^{-1/2} \| X_i \| L/\delta).
\end{aligned}$$

The term I_3 is $o_p(1)$, because it has mean zero and variance $o(1)$ shown as follows

$$E \parallel I_3 \parallel^2 = n^{-1} E \left[\sum_1^n \frac{(\tilde{\sigma}_i - \bar{\sigma}_i)^2}{\bar{\sigma}_i^4} \parallel X_i \parallel^2 \psi^2(u_i) \right]$$
$$+ n^{-1} E \left[\sum_{i \neq j} \frac{(\tilde{\sigma}_i - \bar{\sigma}_i)(\tilde{\sigma}_j - \bar{\sigma}_j)}{\bar{\sigma}_i^2 \bar{\sigma}_j^2} X_i' X_j \psi(u_i) \psi(u_j) \right]$$
$$\leq \delta^{-4} E \parallel X_1 \parallel^2 (\tilde{\sigma}_1 - \bar{\sigma}_1)^2 + \mathcal{O}(nk^{-2}) = o(1),$$

due to the fact that $nk^{-2} \to 0$ and

$$E((\tilde{\sigma}_1 - \bar{\sigma}_1)^2 | X) = \sum_j W_{ij}^2 \sigma_j^2 E(|u_1| - 1)^2 \leq \mathcal{O}(k^{-2}) \sum_{j \neq 1} \sigma_j^2,$$

$$E \parallel X_1 \parallel^2 (\tilde{\sigma}_j - \bar{\sigma}_j)^2 \leq \mathcal{O}(nk^{-2}) = o(1).$$

For I_4, since η_j is bounded by 1, using Hölder's inequality twice, we have

$$I_4 \leq \left[\sum_1^n |\tilde{\sigma}_i - \bar{\sigma}_i|^p \right]^{\frac{1}{p}} \left[n^{-1} \sum_1^n \parallel X_i \parallel^p \right]^{\frac{1}{p}} \left[n^{-1/2} \sum_1^n \eta_i \right]^{\frac{1-2}{p}},$$

where the first term is $o_p(1)$ by Lemma B.3 and the last two terms are $\mathcal{O}_p(1)$ by the assumption C3 and the fact $E(n^{-1/2} \sum_1^n \eta_i) \leq (2L/\delta) \parallel f \parallel E \parallel X_1 \parallel < \infty$. Therefore $I_4 = o_p(1)$ and the lemma is proved. □

PROOF OF LEMMA A.3: Let $T(\Delta) = V(\Delta, \bar{\sigma}) - V(\Delta, \sigma)$. Then

$$T(\Delta) = n^{-1/2} \sum_1^n (\bar{\sigma}_i^{-1} - \sigma_i^{-1}) X_i \psi(u_i - n^{-1/2} \sigma_i^{-1} X_i' \Delta).$$

As both $\bar{\sigma}_i$ and σ_i are bounded from below by $\delta > 0$, applying Koenker and Zhao's (1996) Lemma A.2 to $T(\Delta)$, we obtain $\sup_{\|\Delta\| \leq L} \parallel T(\Delta) - T(0) \parallel = o_p(1)$ by employing the fact that $n^{-1} \sum_1^n (\bar{\sigma}_i^{-1} - \sigma_i^{-1}) \sigma_i^{-1} X_i X_i' \to^P G = 0$. This is true because X_1 has finite p-th moment, $E|\bar{\sigma}_1 - \sigma_1| = o(1)$ by Lemma B.1, and

$$E \left[n^{-1} \sum_1^n |\bar{\sigma}_i^{-1} - \sigma_i^{-1}| \sigma_i^{-1} \parallel X_i \parallel^2 \right] \leq \delta^{-1} E |\bar{\sigma}_1^{-1} - \sigma_1^{-1}| \parallel X_1 \parallel^2$$
$$\leq [E \parallel X_1 \parallel^p]^{2/p} \left[E |\bar{\sigma}_1^{-1} - \sigma_1^{-1}|^{p/(p-2)} \right]^{1-2/p} = \mathcal{O}(1)(E|\bar{\sigma}_1 - \sigma_1|)^{1-2/p} = o(1).$$

Finally $E \parallel T(0) \parallel^2 = E \parallel X_1 \parallel^2 |\bar{\sigma}_1^{-1} - \sigma_1^{-1}|^2 \leq \mathcal{O}(1)(E|\bar{\sigma}_1 - \sigma 1|)^{1-2/p} = o(1)$ implies $\parallel T(0) \parallel = o_p(1)$ and the lemma is proved. □

APPENDIX B

Results regarding the estimates of the conditional scale $\sigma(X)$ of the disturbances are collected in this section. The first lemma is from Robinson (1987, Lemma 1, p887). This is a fundamental result for the consistency of k-NN estimation (see Stone (1977)).

LEMMA B.1 *Let $g(\cdot)$ be any Borel function such that $E|g(X)|^p < \infty$ for $p \geq 1$. Then for the k-NN weights $\{W_{ij}\}$,*

$$E\sum_i W_{1j}|g(X_i) - g(X_1)|^p \to 0, as \quad n \to \infty.$$

LEMMA B.2 $\sum_i (\hat{\sigma}_i - \tilde{\sigma}_i)^2 = o_p(1)$.

PROOF: Using the identity

$$|x-y| - |x| = 2(-y)\psi(x) + 2(y-x)(I(y > x > 0) - I(y < x < 0)),$$

which is valid for $x \neq 0$, and the inequality $(a+b)^2 \leq 2(a^2+b^2)$. We have

$$\begin{aligned}(\hat{\sigma}_i - \tilde{\sigma}_i)^2 &= \left[\sum_j W_{ij}(|\epsilon_j - n^{-1/2}X_j'\Delta_n| - |\epsilon_j|)\right]^2 \\ &\leq 8 \parallel \Delta_n \parallel^2 \left[\parallel n^{-1/2}\sum_j W_{ij}X_j\psi(u_j) \parallel^2 + n^{-1}\sum_j W_{ij} \parallel X_j \parallel^2 \eta_j^*\right],\end{aligned}$$

where we used Jensen's inequality for the second term, and

$$\eta_j^* = I(|u_j| \leq n^{-1/2} \parallel X_j \parallel \Delta_n/\delta), \quad \Delta_n = \sqrt{n}(\check{\beta} - \beta).$$

Since u_i has median zero, $E\psi(u_j) = 0$. Hence

$$\begin{aligned}&E\sum_i \parallel n^{-1/2}\sum_j W_{ij}X_j\psi(u_j) \parallel^2 \\ &\leq E\sum_j W_{1j}^2 \parallel X_j \parallel^2 = \mathcal{O}(k^{-1})E\sum_j W_{1j} \parallel X_j \parallel^2 \to 0,\end{aligned}$$

by noting condition C5 and the fact that $E\sum_j W_{1j} \parallel X_j \parallel^2 \to E \parallel X_1 \parallel^2$ assured by Lemma B.1. Thus $\sum_i \parallel n^{-1/2}\sum_j W_{ij}X_j\psi(u_j) \parallel^2 = o_p(1)$.

For the second term, using Hölder's inequality iteratively

$$E\left[n^{-1}\sum_{i,j} W_{ij} \parallel X_j \parallel^2 \eta_j^*\right] \leq \left[E\sum_j W_{1j} \parallel X_j \parallel^p\right]^{\frac{2}{p}} \left[E\sum_j W_{1j}\eta_j^*\right]^{\frac{1-2}{p}},$$

where $E\sum_j W_{1j} \parallel X_j \parallel^p$ converges to $E \parallel X_1 \parallel^p$ by Lemma B.1, while

$$E\sum_j W_{1j}\eta_j^* \leq E\sum_j W_{ij}\eta_j + P(\parallel \Delta_n \parallel \geq L) = \mathcal{O}(n^{-1/2}L) + P(\parallel \Delta_n \parallel \geq L),$$

for any $0 < L < \infty$. Set a large L, say $L = k^{1/2}$, as $n \to \infty$, $P(|\Delta_n| \geq L) \to 0$ and $n^{-1/2}L \to 0$. Therefore $n^{-1}\sum_{i,j} W_{ij} \parallel X_j \parallel^2 \eta_j^* = o_p(1)$. The lemma is proved by noting $\parallel \Delta_n \parallel = \mathcal{O}_p(1)$. □

Finally, the following two lemmas are parallel to Robinson's (1987) Lemmas 7, 9, 10-14. Their proofs are omitted.

LEMMA B.3 $\sum_i |\tilde{\sigma}_i - \bar{\sigma}_i|^p = \mathcal{O}_p(nk^{-p/2})$ and $(\max_i |\tilde{\sigma}_i - \bar{\sigma}_i|)^p = \mathcal{O}_p(nk^{-p/2})$.

LEMMA B.4 $P(\min_i \sigma_i < \delta) = 0$, $P(\min_i \bar{\sigma}_i < \delta) = 0$, and

$$(\min_i \hat{\sigma}_i)^{-1} = \mathcal{O}_p(1), \quad (\min_i \tilde{\sigma}_i)^{-1} = \mathcal{O}_p(1).$$

REFERENCES

Bickel, P.J. (1975): "One-step Huber estimates in the linear models," *Journal of the American Statistical Association*, 70, 428-434.

Carroll, R.J. and D. Ruppert (1982): "A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model," *Journal of the American Statistical Association*, 77, 878-882.

Engle, R.F. (1982): "Autoregressive conditional heteroscedasticity with estimates of the variance of the United Kingdom inflation," *Econometrica*, 50, 987-1007.

Gutenbrunner, G. and J. Jurečková (1992): "Regression rank scores and regression quantiles," *Annals of Statistics*, 20, 305-330.

Gutenbrunner, G., J. Jurečková , R. Koenker and S. Portnoy (1994): "Tests of linear hypotheses based on regression rank scores," *Journal of Nonparametrics*, 2, 307-331.

Jurečková , J. (1977): "Asymptotic relations of M-estimates and R-estimates in linear regression model," *Annals of Statistics*, 5, 464-472.

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

Koenker, R. and Q. Zhao (1994): "L-estimation for linear heteroscedastic models," *Journal of Nonparametric Statistics*, 3, 223-235.

—— (1996): "Conditional quantile estimation and inference for ARCH models," *Econometric Theory*, 12, 793-813.

Newey, W.K. and J.L. Powell (1990): "Efficient estimation of linear and type I censored regression models under conditional quantile restrictions," *Econometric Theory*, 6, 295-317.

Parzen, M.I., L.J. Wei and Z. Ying (1994): "A resampling method based on pivotal estimating functions," *Biometrika*, 81, 341-350.

Robinson, P.M. (1987): "Asymptotic efficient estimation in the presence of heteroscedasticity of unknown form," *Econometrica*, 55, 875-891.

Ruppert, D. and R.J. Carroll (1980): "Trimmed least square estimation in the linear models," *Journal of the American Statistical Association*, 75, 828-838.

Shick, A. (1986), "On asymptotically efficient estimation in semiparametric models", *Annals of Statistics*, **14**, 1139-1151.

Stone, C.J. (1977): "Consistent nonparametric regression," *Annals of Statistics*, 5, 595-645.