Acta Mathematica Sinica, English Series
Apr., 2009, Vol. 25, No. 4, pp. 597–616
Published online: March 25, 2009
DOI: 10.1007/s10114-008-7220-2
Http://www.ActaMath.com

Acta Mathematica Sinica, English Series

Semiparametric Quantile Modelling of Hierarchical Data

Mao Zai TIAN[1)]
School of Statistics, Renmin University of China, Beijing 100872, *P. R. China*
and
Center for Applied Statistics, Renmin University of China, Beijing 100872, *P. R. China*
E-mail: mztian@ruc.edu.cn mztian@263.net

Man Lai TANG
Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong
E-mail: mltang@hkbu.edu.hk

Ping Shing CHAN
Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
E-mail: benchan@cuhk.edu.hk

Abstract The classic hierarchical linear model formulation provides a considerable flexibility for modelling the random effects structure and a powerful tool for analyzing nested data that arise in various areas such as biology, economics and education. However, it assumes the within-group errors to be independently and identically distributed (i.i.d.) and models at all levels to be linear. Most importantly, traditional hierarchical models (just like other ordinary mean regression methods) cannot characterize the entire conditional distribution of a dependent variable given a set of covariates and fail to yield robust estimators. In this article, we relax the aforementioned and normality assumptions, and develop a so-called *Hierarchical Semiparametric Quantile Regression Models* in which the within-group errors could be heteroscedastic and models at some levels are allowed to be nonparametric. We present the ideas with a 2-level model. The level-1 model is specified as a nonparametric model whereas level-2 model is set as a parametric model. Under the proposed semiparametric setting the vector of partial derivatives of the nonparametric function in level-1 becomes the response variable vector in level 2. The proposed method allows us to model the fixed effects in the innermost level (i.e., level 2) as a function of the covariates instead of a constant effect. We outline some mild regularity conditions required for convergence and asymptotic normality for our estimators. We illustrate our methodology with a real hierarchical data set from a laboratory study and some simulation studies.

Keywords hierarchical models, quantile regression, robustness

MR(2000) Subject Classification 62G05, 62G20, 60G42

1 Introduction

In practice, many real data structures are hierarchical in the sense that individual observations are nested within units and super-unit structures that make it likely that the individual observations are not independent. Typical examples include study of patients within physicians

Received April 30, 2007, Accepted May 21, 2008
Research partially supported by the National Natural Science Foundation of China (NSFC) under grant (No. 10871201), the Key Project of Chinese Ministry of Education (No. 108120), National Philosophy and Social Science Foundation Grant (No. 07BTJ002), 2006 New Century Excellent Talents Program (NCET), HKBU261007 and The Chinese University of Hong Kong Faculty of Science Direct Grant 2060333
[1)] Corresponding author

(two levels) and study of patients within physicians within clinics (three levels). Failing to deal with this hierarchy and lack of independence usually leads to biased parameter estimates.

In the past decades, there has been a large volume of literature on hierarchical data analysis. Lindly and Smith [1] and Smith [2] first introduced the term *hierarchical linear model* as part of their seminal contribution on Bayesian estimation of linear models. Since then, hierarchical models bear different titles such as *multilevel models* ([3–4]), *mixed-effects models* and *random-effects models* ([5–6]), *random-coefficient regression models* ([7–8]) and *covariance components models* ([9–10]).

In the classical hierarchical linear model setting, it is usually assumed that one has observations $\{(X_{ij}, W_i, Y_{ij}),\ i = 1, \ldots, n,\ j = 1, \ldots, n_i\}$ of (X, W, Y), where Y_{ij} is the response value of subject j in unit i, X_{ij}'s are known d by 1 vectors of level-1 predictors and W_i's are known d by f matrixes of level-2 predictors. Suppose that response Y and level-1 predictor X satisfy

$$Y_{ij} = X_{ij}^T \beta_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2), \tag{1.1}$$

where $\beta_i = (\beta_{i1}, \ldots, \beta_{id})^T$ is an unknown d by 1 vector of level-1 coefficients and ϵ_{ij}'s are i.i.d. random errors which are normally distributed with mean 0 and variance σ^2. At level-2, the level-1 coefficients become outcome variables:

$$\beta_i = W_i \gamma + U_i, \qquad U_i \sim N(0, \mathbf{T}), \tag{1.2}$$

where γ is an f by 1 vector of fixed effects, and U_i is a d by 1 vector of level-2 random effects which are assumed to be independent of ϵ_{ij}'s and to follow multivariate normal distribution with mean vector 0 and covariance matrix $\mathbf{T}$.

The aforementioned hierarchical linear models take into the consideration of variability between and within units and one can therefore improve estimates based on data from a single unit by appropriate use of data from the remaining units. However, it is well known that the mean regression model is unable to characterize the entire conditional distribution of a dependent variable given a set of independent variables. Koenker & Bassett [11] proposed the *quantile regression model* which has become a comprehensive approach to the use of linear and nonlinear response models for conditional quantile functions and is now regarded as an appealing alternative to the classical mean regression model. Generally speaking, quantile regression method based on minimizing "check function" residual enables us to estimate all conditional quantile functions, just as classical linear regression method based on least squares estimation enables us to estimate conditional mean functions. In practice, linear quantile model is not adequate to describe the underlying relationship between the quantile of the response variable and its covariates. As a result, the nonparametric approach has been gradually introduced to quantile regression analysis (see, e.g., [12–17]).

Combining hierarchical linear modelling with quantile regression, Tian and Chen [18] proposed a class of models called *hierarchical linear quantile regression models*. In that article, a new approach, which is based on the Gauss–Seidel iteration and takes full advantage of quantile regression and hierarchical models, was developed. On the theoretical front, the asymptotic properties of the new method are considered. Simple conditions are required for $n^{1/2}$-convergence and asymptotic normality.

However, the assumption of linearity made in model (1.1) is sometimes not practical. The purpose of this article is to extend the existing hierarchical linear model and local linear quantile regression model to a so-called *Hierarchical Semiparametric Quantile Regression Model*. The new proposed model allows level-1 model to be nonparametric. Under this nonparametric setting, the vector of partial derivatives of the nonparametric function, which is usually called the marginal effects in econometrics, becomes the outcome variable vector of level 2. To study the effect of the covariates on the entire conditional distribution of response, we consider the quantile regression coefficients. Unlike the ordinary hierarchical mean regression model in which the fixed effects are usually assumed to be constant over all values of the covariates, our proposed model allows the quantiles of the fixed effects to be functions of the covariates. We believe that the new proposed model would be appealing to many practitioners in statistics.

This article is organized as follows. In Section 2, the hierarchical semiparametric quantile regression models are developed. The problem of estimation is also discussed. In Section 3, the asymptotic properties are presented. Simple simulation studies are conducted to illustrate the proposed methodology in Section 4. In Section 5, we illustrate our method with a real data set from an animal study.

2 The Models and Estimation

In traditional hierarchical linear models, the level-1 model is assumed to be linear as given in (1.1). In this case, the mean conditional response function is $\mu(x) = E(Y_{ij}|X_{ij} = x) = x^T\beta_i, \quad j = 1, \ldots, n_i$, and its first partial derivative vector is defined as follows:

$$\nabla\mu(x) = (\partial\mu(x)/\partial x_1, \ldots, \partial\mu(x)/\partial x_d)^T = (\beta_{i1}, \ldots, \beta_{id})^T,$$

in which β_{ik} $(k = 1, \ldots, d)$ can be interpreted as a measure of change of the mean response with the i-th covariate being perturbed while keeping other covariates fixed. In fact, $\nabla\mu(x)$ is usually called the *marginal effect* in econometrics (see, for example, [19]). For instance, we note that in the linear model $Y_{ij} = \sum_{k=1}^{d} \gamma_{ik}X_{ijk} + \epsilon_{ij}$, with $E(\epsilon_{ij}) = 0$, the vector $(\gamma_{i1}, \ldots, \gamma_{id})^T$ is the first partial derivative vector of $\mu(x) = E(Y_{ij}|x_1, \ldots, x_d) = \gamma_{i1}x_1 + \cdots + \gamma_{id}x_d$.

However, the assumption of linearity made in model (1.1) is sometimes not practical. In this article, we consider the following semi-parametric model at level 1:

$$Y_{ij} = m(X_{ij}; \beta_i) + \epsilon_{ij}, \qquad i = 1, \ldots, n, \ j = 1, \ldots, n_i, \tag{2.1}$$

where $m(\cdot)$ is an unknown function governing the within-individual behavior and ϵ_{ij}'s are unobservable random error variables with mean $E(\epsilon_{ij}) = 0$ and variance $E(\epsilon_{ij}^2) = \sigma_i^2(X_{ij})$ which allows within-group errors to be heteroscedastic. β_i is a $d \times 1$ random vector, as opposed to observable variable, and is not directly observed but is rather inferred from the higher level variables that are observed and directly measured. The β_i's are also called *hidden variables, model parameters or interim parameters.* Advantages of using β_i here include the reduction of the dimensionality of data and characterization of the effects of higher level variables. A large number of observable variables can be aggregated to represent an underlying concept, making it easier for audience to understand the data.

We now turn to the nonparametric model in (2.1). Assume that $m(\cdot)$ has a 1-st order partial continuous derivative at the point x. For data points X_{ij} in the neighborhood of x we approximate $m(X_{ij};\beta_i)$ via a Taylor expansion by a linear function, i.e.,

$$m(X_{ij};\beta_i) \approx m(x;\beta_i) + (X_{ij}-x)^T \nabla m(x;\beta_i).$$

Hence,

$$m(X_{ij};\beta_i) \approx \widetilde{X}_{ij}^T \theta_i, \tag{2.2}$$

where $\widetilde{X}_{ij} = (1,(X_{ij}-x)^T)^T$, and $\theta_i = (m(x;\beta_i), \nabla m(x;\beta_i)^T)^T = (\theta_{i0},\theta_{i1},\ldots,\theta_{id})^T$. Model (2.2) is based on the idea of local fitting which is indeed particularly useful in nonparametric estimation. In fact, Model (2.2) is not completely new in the literature and has also been considered in existing literature (e.g., [20]).

Similar to the level 2 model in (1.2), we propose the new level 2 model as

$$\theta_i = W_i\gamma(x) + U_i, \tag{2.3}$$

where W_i's are $(d+1)\times f$ matrixes of known level-2 predictors, $\gamma(x)$ is an f by 1 vector of fixed effects function, and U_i is a $d+1$ by 1 vector of level-2 random errors which are independent of ϵ_{ij}'s with mean vector $E(U_i)=0$ and covariance matrix $\mathrm{Cov}(U_i)=\mathbf{T}$.

Remark Obviously, the traditional hierarchical linear models (1.1)–(1.2) are special cases of the (2.1)–(2.3). Also note that in all subsequent discussion in this paper we only consider the special case in which the marginal effect θ_i is independent of the index j. That is to say, the gradient vector θ_i is a constant function of j. Hence θ_i can be written without the subscript j. More background issues in research on this case can be found easily in many literatures. Two examples are presented here for illustration.

Example I A study of the SES-Achievement relationship in J Schools

We consider a study of the relationship between a single student-level predictor variable, the socioeconomic status [SES], and student-level outcome variable, the mathematics achievement, within an entire population of schools. Suppose we have a random sample of J schools from population. We can describe this relationship within any school by a simple two-level hierarchical linear model:

$$\begin{aligned}
&\text{Level I (e.g., students)} && Y_{ij} = \beta_{i0} + \beta_{i1}X_{ij} + r_{ij},\\
&\text{Level II (e.g., schools)} && \beta_{i0} = \gamma_{00} + \gamma_{01}W_i + u_{i0},\\
& && \beta_{i1} = \gamma_{10} + \gamma_{11}W_i + u_{i1},
\end{aligned}$$

where Y_{ij} and X_{ij} denote the mathematics achievement and SES, respectively. In this case, we can call student j is nested within school i. More details can be found in Chapter 2 of [21] (Note that the notations i and j are j and i there). Each parameter in the above model has a substantive meaning. The intercept, β_{i0}, is the expected math achievement of a student whose SES is zero. The slop, β_{i1}, is the expected change in math achievement associated with a unit increase in SES. Usually, β_{i0} and β_{i1} represent respectively the "effectiveness" and

"equitableness" of a school. As for this case, $\theta_i = (\beta_{i0}, \beta_{i1})^T$ which is independent of the subscript j.

Example II The effects of maternal speech on children's vocabulary

Hunttenlocher [22] examined the role of exposure to speech in children's early vocabulary growth. In this study, they characterized vocabulary growth rates for each of 22 children by using data obtained at several time points from 14 to 26 months. A visual examination of individual child vocabulary growth trajectories clearly indicated a nonlinear growth pattern. In fact, all 22 trajectories displayed upward curvature, which motivated them to consider the model at level 1 of the following form

$$Y_{it} = \pi_{i0} + \pi_{i1}(a_{it} - L) + \pi_{i2}(a_{it} - L)^2 + e_{it},$$

where Y_{it} is a measure of the child's vocabulary size at each measurement occasion. L is a specific or a priori centering constant for the level 1 predictors that are powers of a_{it}. Here the center parameter, L, is deliberately set at 12 months, because this is about the time that most children begin to express their first words. The intercept, π_{i0}, represents the status of child i at time L. The linear component, π_{i1}, is the instantaneous growth rate for child i at time L, and π_{i2} captures the curvature or acceleration in each growth trajectory. In this example, it is easy to show that

$$\theta_i = (\theta_{i0}, \theta_{i1})^T = (\pi_{i0} + \pi_{i1}(x - L) + \pi_{i2}(x - L)^2, \quad \pi_{i1}(x - L) + 2\pi_{i2}(x - L))^T.$$

At level 2, we may consider a separate equation for θ_i, that is,

$$\theta_{ik} = \gamma_{0k} + \sum_{q=1}^{Q_k} \gamma_{qk} w_{iq} + u_{ik},$$

where $k = 0, 1$. In this example, θ_i is independent of the subscript t which is the index of level 1 predictors.

In this article, we would like to quantify the relationship between the τ-th quantile of response Y and the covariate vector, (X, W). For this purpose, we assume that the conditional distribution $F(\cdot|\ x, w)$ is an increasing and continuous function at (x, w). Hence, the τ-th quantile regression function of Y given $X = x$ and $W = w$ (denoted as $q_\tau(x, w)$) is defined as

$$q_\tau(x, w) = \inf\{t \in R : F(t|\ x, w) \geq \tau\},$$

for any given $\tau, 0 < \tau < 1$.

The following lemma describes the relationship between the τ-th quantile of response Y_{ij} and the covariate vector, (X_{ij}, W_i).

Lemma 1 *Let* $\{(X_{ij}, W_i, Y_{ij}),\ i = 1, \ldots, n,\ j = 1, \ldots, n_i\}$ *be the observations of* (X, W, Y), *which satisfy* (2.1), (2.2) *and* (2.3), *with* Y_{ij}*'s being the outcome variables,* X_{ij}*'s being the known* $d \times 1$ *vectors for the level-*1 *predictors and* W_i*'s being the known* $(d+1) \times f$ *matrixes for the level-*2 *predictors. Let* $\xi_{ij} = \widetilde{X}_{ij}^T U_i + \epsilon_{ij}$ *with* $\xi_{ij} \sim G_{ij}$ *and* $F(y)$ *be the distribution function of* Y. *Then for all* y *within the interior of the support of* $F(y)$, *the* τ*-th conditional quantile of* Y_{ij} *given* (X_{ij}, W_i) *is*:

$$F_{ij}^{-1}(\tau) = \widetilde{X}_{ij}^T W_i \gamma_\tau(x) + e_{ij}(\tau),$$

where $\gamma_\tau(x)$ depending on τ is a fixed effects function and $e_{ij}(\tau)$ is the τ-th quantile of G_{ij}, respectively.

Proof See Appendix.

Remark If $\epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0,\sigma^2)$, $U_i \sim N(0,T)$ and U_i is independent of ϵ_{ij} for all $i = 1,\ldots,n,\ j = 1,\ldots,n_i$, then $e_{ij}(\tau) = (\widetilde{X}_{ij}^T\mathbf{T}\widetilde{X}_{ij} + \sigma^2)^{1/2}\Phi^{-1}(\tau)$, with $\Phi(\cdot)$ being a standard normal probability distribution function.

Here, we need to estimate $\mathbf{T}$, $\sigma_i^2(x)(i = 1,\ldots,n)$ and $\gamma_\tau(x)$ for any given x. For the estimation of $\gamma_\tau(x)$, we consider the following objective function:

$$R_n(\mathbf{b}_n) \equiv \sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_\tau(Y_{ij} - \widetilde{X}_{ij}^T W_i\mathbf{b}_n - e_{ij}(\tau))K_H(X_{ij} - x), \tag{2.4}$$

where $\rho_\tau(z) = \tau z I_{[0,\infty)}(z) - (1-\tau)zI_{(-\infty,0)}(z)$ and $K_H(\mathbf{z}) = \frac{1}{\det(H)}K(H^{-1}\mathbf{z})$ with $K(\cdot)$ being the Gaussian kernel function (see, e.g., [23–24]). We can then obtain the estimate of $\gamma_\tau(x)$, denoted as $\widehat{\gamma}_\tau(x)$, by minimizing $R_n(\mathbf{b}_n)$ with respect to $\mathbf{b}_n$, i.e.,

$$\widehat{\gamma}_\tau(x) = Arg\min_{\mathbf{b}_n\in R^f} R_n(\mathbf{b}_n). \tag{2.5}$$

Here, $\rho_\tau(\cdot)$ is called the "check function", $I(\cdot)$ is the usual indicator function, $K_H(\cdot)$ is the kernel function that controls the relative influence of level-1 predictors on the estimation of the $\gamma_\tau(x)$ parameter, and H represents the bandwidth matrix. By setting $H = h\mathbf{I}_d$ with $\mathbf{I}_d$ being the $d\times d$ identity matrix, we obtain the kernel function of equal bandwidth, i.e., $K_d(X_{ij}-x) = \frac{1}{h^d}K(\frac{X_{ij1}-x_1}{h},\ldots,\frac{X_{ijd}-x_d}{h})$. For non-equal bandwidth, we set $H = \mathrm{diag}(h_1,\ldots,h_d)$ and the corresponding kernel function is $K_d(X_{ij}-x) = \frac{1}{h_1\ldots h_d}K(\frac{X_{ij1}-x_1}{h_1},\ldots,\frac{X_{ijd}-x_d}{h_d})$.

In general, the advantages of the introduction of the kernel function $K_H(\cdot)$ include: (1) It serves as a smooth weight function with a compact support within the interior of the support of $F(y)$; (2) The estimate $\widehat{\gamma}_\tau(x)$ is less overly influenced either by outliers or the tail behavior of $F(y)$; and (3) In some cases, it reduces the boundary effect on the nonparametric smoothing setting. In this paper, we adopt the Gaussian kernel function and the following automatic bandwidth selection strategy for smoothing conditional quantiles (e.g., [16]):

Step 1 Use ready-made and sophisticated methods to select h_{mean} (see, e.g., [25]).

Step 2 Use $h_\tau = h_{\text{mean}}\{\tau(1-\tau)/\phi(\Phi^{-1}(\tau))^2\}^{1/5}$ to obtain all other h_τ's from h_{mean}, where ϕ and Φ are the standard normal density and distribution functions.

For non-automatic bandwidth selection, one may consult Gooijer and Zerom [26].

The 1-st order condition for the optimal problem in (2.5) is given by $\nabla R_n(\mathbf{b}_n)|_{\mathbf{b}_n=\widehat{\gamma}_\tau(x)} = 0$ with

$$\begin{aligned}\nabla R_n(\mathbf{b}_n) &= -\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij} - \widetilde{X}_{ij}^T W_i\mathbf{b}_n + e_{ij}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij} - \widetilde{X}_{ij}^T W_i\{\mathbf{b}_n - \gamma_\tau(x)\} - F_{ij}^{-1}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij},\end{aligned}$$

where $\psi_\tau(z) = \tau - I(z<0)$. To solve the minimization problem in (2.5), we need to calculate the estimate of the 100τ-th percentile $e_{ij}(\tau)$ of G_{ij}. Usually, estimating quantiles is computationally

more difficult than estimating mean and variance. Here, we compute the point estimator of $e_{ij}(\tau)$ using order statistics (see [27]). In our case, the sample τ quantile $\widetilde{e}_{ij}(\tau)$ is given by

$$\widetilde{e}_{ij}(\tau) = \big(Y_{ij} - \widetilde{X}_{ij}^T W_i \widehat{\gamma}_\tau(X_{ij})\big)_{([n_i\tau])}, \tag{2.6}$$

where $[a]$ represents the largest integer that is less than or equal to the real number a, $(Y_{ij} - \widetilde{X}_{ij}^T W_i \widehat{\gamma}_\tau(X_{ij}))_{([n_i\tau])}$ denotes the $[n_i\tau]$-th smallest value of $\{Y_{i1} - \widetilde{X}_{i1}^T W_i \widehat{\gamma}_\tau(X_{i1}), \ldots, Y_{in_i} - \widetilde{X}_{in_i}^T W_i \widehat{\gamma}_\tau(X_{in_i})\}$, and $\widehat{\gamma}_\tau(\cdot)$ is any reasonable estimator of $\gamma(\cdot)$. In practice, the estimator $\widehat{\gamma}_\tau(x)$ in (2.5) and $\widehat{e}_{ij}(\tau)$ can be obtained by the following Guass–Seidel-type of iteration procedure:
Step 1 Initial estimation: $\widehat{e}_{ij}^{(0)}(\tau) = (Y_{ij} - \widetilde{X}_{ij}^T W_i \widehat{\theta}_i^{(0)})_{([n_i\tau])}$, where $\widehat{\theta}_i^{(0)}$ is defined in (2.7) below.
Step 2 Iteration:

$$\widehat{\gamma}_\tau^{(l+1)}(x) = \operatorname{Arg}\min_{\mathbf{b}\in R^f} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_\tau(Y_{ij} - \widetilde{X}_{ij}^T W_i \mathbf{b}_n - \widehat{e}_{ij}^{(l)}(\tau)) K_H(X_{ij} - x),$$

$$\widehat{e}_{ij}^{(l+1)}(\tau) = \operatorname{Arg}\min_{a\in R} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_\tau(Y_{ij} - \widetilde{X}_{ij}^T W_i \widehat{\gamma}_\tau^{(l+1)}(X_{ij}) - a),$$

where $l = 0, 1, \ldots$.
Step 3. Repeat Step 2 until the convergence criterion is satisfied (e.g., until the largest change in the value of any parameters is sufficiently small).

Here, we use the regression mean $E(\theta_i)$ as the initial value of θ_i. Specifically, the coefficient vector θ_i can be estimated by employing some appropriate smoothing technique and the intuitive estimator is its mean value:

$$\widehat{\theta}_i^{(0)} = (\widehat{\theta}_{i0}^{(0)}, \widehat{\theta}_{i1}^{(0)}, \ldots, \widehat{\theta}_{id}^{(0)})^T \equiv E(m(x), \nabla m(X)^T)^T, \tag{2.7}$$

where $\widehat{\theta}_{i0}^{(0)}$, the initial state of the response, is the expected value of the response with all marginal effects of covariates being zero, and $\widehat{\theta}_{ik}^{(0)}$ $(k = 1, \ldots, d)$ represents the average change in the response with the k-th covariate being perturbed while the other covariates being held fixed. Equivalently, we want to estimate

$$\int (m(x), \nabla m(t)^T)^T f(t) dt.$$

Obviously, the direct plug-in estimator can be adopted as an initial estimator:

$$n_i^{-1} \sum_{j=1}^{n_i} (\widehat{m}(x), \nabla \widehat{m}(X_{ij})^T)^T, \tag{2.8}$$

where $\widehat{m}(x)$ is any reasonable nonparametric estimator of $m(x)$. (see [19]).

For the estimation of $\mathbf{T}$ and $\sigma_i^2(x)$, we adopt the following E–M type algorithm (see [2, 9]):
M-Step *Maximization.* The maximization estimators for $\mathbf{T}$ and $\sigma_i^2(x)$ are straightforward:

$$\widehat{\mathbf{T}} = \frac{1}{n}\sum_{i=1}^{n} U_i U_i^T,$$

and

$$\widehat{\sigma}_i^2(x) = \frac{1}{n_i}\sum_{j=1}^{n_i} (Y_{ij} - \widetilde{X}_{ij}^T W_i \widehat{\gamma}_\tau(x) - \widetilde{X}_{ij}^T U_i)^2.$$

E-Step *Expectation.* The conditional expectations of $\widehat{\sigma}_i^2(x)$ and $\widehat{\mathbf{T}}$ (given the data Y and other parameters) can be shown to be

$$E(\widehat{\mathbf{T}}|Y,\gamma_\tau(x),\sigma_i^2(x),\mathbf{T}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(U_i^* U_i^{*T} + \mathbf{T}_i^*)$$

and

$$E(\widehat{\sigma}_i^2(x)|Y,\gamma_\tau(x),\sigma_i^2(x),\mathbf{T}) = \frac{1}{n_i}\sum_{j=1}^{n_i}(Y_{ij} - \widetilde{X}_{ij}^T W_i\gamma_\tau(x) - \widetilde{X}_{ij}^T U_i^*)^2 + tr(\widetilde{X}_{ij}\mathbf{T}_i^*\widetilde{X}_{ij}^T),$$

where $\mathbf{T}_i^* = \sigma_i^2(x)(\widetilde{X}_{ij}\widetilde{X}_{ij}^T + \sigma_i^2(x)\mathbf{T}^{-1})^{-1}$ and $U_i^* = \sigma_i^{-2}(x)\mathbf{T}_i^*\widetilde{X}_i^T(Y_{ij} - \widetilde{X}_{ij}^T W_i\gamma_\tau(x))$.

The E–M type algorithm continues until the absolute changes of all parameter values are less than a pre-specified sufficiently small quantity. Since the data are assumed to be heteroscedastic at level-1, i.e., $E(\epsilon_{ij}^2) = \sigma_i^2(X_{ij})$ with a smooth function $\sigma_i^2(\cdot)$, we can consider the class of local variance estimates at x as an initial estimator for $\sigma_i^2(x)$. That is, $\widetilde{\sigma}_i^2(x) = (\sum_{j=j_1}^{j_2}\omega_j Y_{ij})^2$, where $j_1 = -[m/2]$, $j_2 = [m/2 - 1/4]$, $\sum_{j=j_1}^{j_2}\omega_j = 0$ and $\sum_{j=j_1}^{j_2}\omega_j^2 = 1$, $m \geq 2$ is a fixed integer (see, e.g., [28]).

3 Asymptotic Results

In this section, we investigate the asymptotic behavior of $\sqrt{n\cdot\det(H)}\ (\widehat{\gamma}_\tau(x) - \gamma_\tau(x))$ in the hierarchical nonparametric quantile regression models with heteroscedastic errors at different levels. The asymptotic properties of the order statistics in (2.6) will also be studied. Throughout this section, we make the following mild assumptions.

Assumption 1 *Let $\{Y_{ij}\}$ $(i = 1,\ldots,n;\ j = 1,\ldots,n_i)$ be a sequence of independent random variables with distribution function $F_{ij}(\cdot)$ which is absolutely continuous, and has finite, positive, and absolutely continuous density $F_{ij}'(\cdot)$ for all z such that $0 < F_{ij}(z) < 1$. Moreover, its second derivative $F_{ij}''(\cdot)$ is bounded in a neighborhood of $F_{ij}^{-1}(\tau)$, $\tau \in (0,1)$.*

Assumption 2 *Let $e_k = (0,\ldots,0,1,0,\ldots,0)^T$ with the k-th element being 1.* $\max_{1\leq i\leq n}\sum_{k=1}^{f}|\widetilde{X}_{ij}^T W_i e_k| = O(1/n^4)$, *as $n\to\infty$.*

Assumption 3 $\frac{1}{n}\sum_{i=1}^{n}|\widetilde{X}_{ij}^T W_i e_k|^4 = O(1)$, *as $n\to\infty$, $k = 1,\ldots,f$.*

Assumption 4 *The d-dimensional kernel function $K(\cdot)$ is a bounded density function with a compact support $\mathbf{C}^d$ within the interior of the support of $f(x)$ such that $\int K(u)du = 1$, $\int uK(u)du = 0_d$, $\int uu^T K(u)du > 0_{d\times d}$.*

Assumption 5 *The bandwidth matrix H of the kernel function satisfies* $\det(H)\to 0$ *and* $n\cdot\det(H)\to\infty$, *as $n\to\infty$.*

Remark Note that the Assumption 1 covers the special case in which $n_i = 1$ and $Y_{ij} = Y_i$ while Assumptions 2 and 3 are regular assumptions borrowed from [18, 29–31]. Consistency is essentially ensured by Assumptions 1–5.

Let

$$H(\mathbf{b}_n) = -\frac{1}{\sqrt{n\cdot\det(H)}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij} - (n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n - F_{ij}^{-1}(\tau)) \cdot K_H(X_{ij} - x)W_i^T\widetilde{X}_{ij}.$$

Below, we first state some lemmas that are necessary for Theorem 1. All proofs will be provided in Appendix.

Lemma 2 *Under Assumptions* 1–5, *we have*

$$E(H(\mathbf{b}_n)) = \left\{(n\det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} F'_{ij}(F_{ij}^{-1}(\tau))W_i^T\widetilde{X}_{ij}K_H(X_{ij}-x)\widetilde{X}_{ij}^T W_i\right\}\mathbf{b}_n + o(1),$$

for any sequence of random vectors such that $\|\mathbf{b}_n\|_\infty = O_P(1)$.

Lemma 3 *Suppose the second order derivative of* F_{ij} *is bounded in a neighborhood of* $F_{ij}^{-1}(\tau)$, $\tau \in (0,1)$. *We have*

$$\sup_{\|\mathbf{b}_n\|\le C}\|H(\mathbf{b}_n) - H(\mathbf{0})\|_\infty = EH(\mathbf{b}_n) + O_P(1), \qquad n\to\infty,$$

for some constant C *and any sequence of random vectors such that* $\|\mathbf{b}_n\|_\infty = O_P(1)$.

The following lemma is an immediate result from Lemmas 2 and 3.

Lemma 4 *Under Assumptions* 1–5, *we have*

$$\sup_{\|\mathbf{b}_n\|\le C}\|H(\mathbf{b}_n) - H(\mathbf{0}) - \Delta_n\mathbf{b}_n\|_\infty = o_P(1), \qquad as \quad n\to\infty,$$

for some fixed constant C.

Lemma 5 *Under Assumptions* 1–3, *we have*

$$\sqrt{n\cdot\det(H)}\Big(\widehat{\gamma}_\tau(x) - \gamma_\tau(x)\Big) = O_P(1), \qquad as \quad \det(H)\to 0,\ n\det(H)\to\infty.$$

Theorem 1 *Under Assumptions* 1–5, *we have the following Bahadur representation*:

$$\begin{aligned}&\sqrt{n\cdot\det(H)}(\widehat{\gamma}_\tau(x) - \gamma_\tau(x))\\ &\quad= \frac{1}{\sqrt{n\cdot\det(H)}}\Delta_n^{-1}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij} - F_{ij}^{-1}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\right\} + o_P(1),\end{aligned}$$

where $\Delta_n = (n\ \det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}F'_{ij}(F_{ij}^{-1}(\tau))W_i^T\widetilde{X}_{ij}K_H(X_{ij}-x)\widetilde{X}_{ij}^T W_i$.

The following theorem states the asymptotic distribution of $\widehat{\gamma}_\tau(x)$.

Theorem 2 *Under Assumptions* 1–5, *and assuming that* Δ_n *and* Θ_n *tend to positive definite matrixes,* $\sqrt{n\cdot\det(H)}(\widehat{\gamma}_\tau(x)-\gamma_\tau(x))$ *converges in distribution to a multivariate Gaussion vector. Specifically,*

$$\sqrt{n\cdot\det(H)}(\widehat{\gamma}_\tau(x) - \gamma_\tau(x)) \xrightarrow{D} N(0, \tau(1-\tau)\Delta_n^{-1}\Theta_n\Delta_n^{-1}),$$

where $\Theta_n = (n\det(H))^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}W_i^T\widetilde{X}_{ij}K_H^2(X_{ij}-x)\widetilde{X}_{ij}^T W_i$.

The assumption of Theorem 2 is similar to that of A.4 in [31] and that of Appendix of [30]. Luckily, there exits a large literature on estimating $\{F'_{ij}(F_{ij}^{-1}(\tau))\}^{-1}$. Our proposal follows that of [31].

The following theorem states the asymptotic distribution of $\widetilde{e}_{ij}(\tau)$.

Theorem 3 *Under Assumption* 1 *and* (2.6), *we have*

$$\sqrt{n}\left(\widetilde{e}_{ij}(\tau) - e_{ij}(\tau) + \frac{\tau(1-\tau)G''_{ij}(e_{ij}(\tau))}{2(n_i+2)G'^3_{ij}(e_{ij}(\tau))}\right) \xrightarrow{\mathcal{D}} N(0, \tau(1-\tau)/G'^2_{ij}(e_{ij}(\tau))),$$

where $G_{ij}(e_{ij}(\tau)) = F_{ij}(\widetilde{X}_{ij}^T W_i \gamma_\tau(x) + e_{ij}(\tau))$.

4 Simulation Studies

In this section, we conduct two simulation studies to evaluate the proposed method. For the sake of simplicity, we only consider models with one explanatory variable. The first study, which is based on hierarchical linear models with errors having multivariate Cauchy distributions, shows that our proposed method is robust to non-normality/outliers. The second study investigates the marginal effects on hierarchical nonparametric quantile regression.

4.1 Hierarchical Linear Models with Errors Having Multivariate Cauchy Distributions

Assume that the model at level 1 model is linear with Cauchy errors, i.e., $Y_{ij} = \beta_{i0} + \beta_{i1} X_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \sim$ Cauchy (mode $= 0$, scale $= 1$), and that the model at level 2 is defined as in (1.2), i.e., $\beta_{i0} = \gamma_0 + u_{i0}$, and $\beta_{i1} = \gamma_1 + u_{i1}$, where

$$\begin{pmatrix} u_{i0} \\ \\ u_{i1} \end{pmatrix} \sim \text{Cauchy} \left(\text{mode} = \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}, \Sigma = \frac{1}{9} \begin{pmatrix} 1 & -0.25 \\ \\ -0.25 & 1 \end{pmatrix} \right).$$

Here, all comparisons are based on the average of 200 replications.

We will compare the results from the ordinary mean regression and those from our hierarchical nonparametric quantile regression at median (i.e., $\tau = 0.5$). The sample sizes are set as $n = 11,\ n_1 = 290,\ n_2 = 270,\ n_3 = 260,\ n_4 = 290,\ n_5 = 290,\ n_6 = 290,\ n_7 = 270,\ n_8 = 310,\ n_9 = 250,\ n_{10} = 290,\ n_{11} = 270$. Let $X_{ij} \sim U(0,1)$, $\gamma_0 = 60$, and $\gamma_1 = 12$. The Gaussian kernel function and the automatic bandwidth selection strategy for smoothing conditional quantiles are employed here. Details are referred to those presented in Section 2. In Figure 1, we plot the estimated marginal effects of γ_1 only. It should be noted that the scales in Figure 1 are different for the two plots. In fact, the differences between the maximum and the true value (which is 12) of the marginal effect γ_1 for the two plots in Figure 1 are substantial. That is, $\max(|\widehat{\gamma}_1^* - 12|) = 451.75$ and $\max(|\widehat{\gamma}_1^{**} - 12|) = 1.13$ where $\widehat{\gamma}_1^*$ and $\widehat{\gamma}_1^{**}$ are the estimators for the marginal effects γ_1 based on the ordinary mean regression method and our method at median, respectively. Obviously, our estimator is robust to outliers.

4.2 Hierarchical Nonparametric Quantile Regression Models with Heteroscedastic Errors

Suppose that the heteroscedastic level-1 model is defined as

$$Y_{ij} = m(X_{ij}) + (\sin(2\pi X_{ij}) + 1)^{1/2} \epsilon_{ij}, \tag{4.1}$$

where $m(x) = \phi_{0i} + \phi_{1i} \sin(2\pi\phi_{2i} x) + \phi_{3i} \cos(2\pi\phi_{2i} x),\ \epsilon_{ij} \sim N(0,1)$, and that the level-2 model is defined as

$$\theta_i(x) = (m(x), m'(x))^T = A_i \varphi(x), \tag{4.2}$$

where

$$A_i = \begin{pmatrix} \phi_{0i} & \phi_{1i} & \phi_{3i} \\ 0 & -2\pi\phi_{2i}\phi_{3i} & 2\pi\phi_{1i}\phi_{2i} \end{pmatrix},$$

$$\varphi(x) = \begin{pmatrix} 1 \\ \sin(2\pi\phi_{2i}x) \\ \cos(2\pi\phi_{2i}x) \end{pmatrix}.$$

One of the primary interests is to estimate the fixed effects across the population. For this reason, we let (4.2) to be $\theta_i(x) = \theta(x) + U_i$, where the fixed-effect function is given by

$$\begin{aligned} \theta(x) &= A\varphi(x) \\ &= \begin{pmatrix} \phi_0 & \phi_1 & \phi_3 \\ 0 & -2\pi\phi_2\phi_3 & 2\pi\phi_1\phi_2 \end{pmatrix} \begin{pmatrix} 1 \\ \sin(2\pi\phi_2 x) \\ \cos(2\pi\phi_2 x) \end{pmatrix} \end{aligned}$$

and $U_i = \left(\begin{smallmatrix} u_{0i} \\ u_{1i} \end{smallmatrix}\right) \sim N(\mu, \mathbf{\Sigma})$ with $\mu = (0,0)^T$ and $\mathbf{\Sigma} = \left(\begin{smallmatrix} 4 & 1 \\ 1 & 2 \end{smallmatrix}\right)$.

Here, we consider five different values for $\tau = 0.05, 0.25, 0.5, 0.75$ and 0.95. The sample sizes are $n = 11,\ n_1 = 29,\ n_2 = 27,\ n_3 = 26,\ n_4 = 29,\ n_5 = 29,\ n_6 = 29,\ n_7 = 27,\ n_8 = 31,\ n_9 = 25,\ n_{10} = 29,\ n_{11} = 27$. Let $X_{ij} \sim U(0,1)$, and $\phi_0 = 12,\ \phi_1 = -3,\ \phi_2 = 1$, and $\phi_3 = -1$ for $i = 1, \ldots, 11,\quad j = 1, \ldots, n_i$. We report only the case for $\tau = 0.05$ in Figure 2 and omit other cases at different quantiles. Plot (a) in Figure 2 is the scatter plot for the heteroscedastic model (4.1). The solid line in Plot (b) indicates the true fixed effect function $\gamma^{(1)}_{0.5}(x)$ at quantile 0.5, and the corresponding estimator $\widehat{\gamma}^{(1)}_{0.5}(x)$ is shown as the dotted line. Visually, the estimate of the fixed effect based on our method is very close to the true one. For nonparametric setting, instead of looking at the estimate at one particular point x, it might be worthy of calculating a measure of the goodness of fit for the whole estimator. For this purpose, we consider the *Integrated Absolute Errors* (IAE) defined as $\int_x |\gamma^{(1)}_\tau(x) - \widehat{\gamma}^{(1)}_\tau(x)| f(x)dx$. We compute the mean IAE based on 500 replications and the results are $1.02, 0.65, 0.68, 0.59,$ and 1.02 for $\tau = 0.05, 0.25,\ 0.5, 0.75$ and 0.95, respectively.

5 Real data example

In this section, we will illustrate the application of our proposed model with the pixel intensity dataset which is available from the software S-PLUS. Specially, the mean pixel intensity of the right and left lymphnodes in the axillary region obtained from Computed Tomography (CT) scans of 10 dogs were recorded over a period of 14 days after application of a contrast. The variables of interest include (a) *Pixel* (i.e.,mean pixel intensity of lymphnodes in the CT scan); (b) *Day* (i.e.,number of days since contrast administration); (c) *Dog* (i.e., a factor giving the unique identifier for each dog); and (d) *Side* (i.e., a factor indicating the side on which the measurement was made). The experimenters injected each of the ten dogs with a dye contrast and recorded the mean pixel intensities from CT scans of the right and left lymph nodes in the axillary region on several occasions up to 21 days post injection. We show the pixel intensity on CT scans over time for lymph nodes on the left and the right sides of the 10 dogs in Figure 3. In this experiment, the left and right sides are expected to be different. In subsequent analysis, *Dog* and *Side* are considered to be nested classification variables.

First, we would like to demonstrate the inadequacy of the linearity assumption at level-1. For simplicity, we just consider the special case in which level 1 involves only the covariate *Day* and level 2 represents the random effects for *Dog*. Specifically, denote the pixel density on the j-th side of the i-th dog at the k-th occasion as $Y_{ijk},\ i = 1, \ldots, 10;\ j = 1, 2; k = 1, \ldots, n_{ij}$, and the time of the k-th scan on the i-th dog as X_{ik}. Hence, we have the following hierarchical linear model:

$$\begin{aligned} &\text{Level-1} \qquad Y_{ijk} = \beta_{0k} + \beta_{1k} X_{ik} + \epsilon_{ijk}, \\ &\text{Level-2} \qquad \beta_{0k} = \gamma_{00} + u_{0k}, \quad \text{and} \quad \beta_{1k} = \gamma_{10} + u_{1k}, \end{aligned} \tag{5.1}$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$, $E((u_{0k}, u_{1k})^T) = (0, 0)^T$, $\mathrm{Var}((u_{0k}, u_{1k})^T) = T_{2\times 2}$ and $\mathrm{Cov}(u_{0k}, \epsilon_{ijk}) = \mathrm{Cov}(u_{1k}, \epsilon_{ijk}) = 0$. It is noteworthy that the number of observations on side j of dog i (i.e., n_{ij}) does not depend on j but i. Hence, the Y_{ijk} and ϵ_{ijk} at level-1 can be simplified as Y_{ik} and ϵ_{ik}, respectively. We can fit (5.1) by the **lme** function from the **nlme** library of **S**. We have $\widehat{\sigma} = 14.53$, $\widehat{\gamma} = (1093.22, -0.15)^T$, $\widehat{T} = \left(\begin{smallmatrix} 31.49 & -0.79 \\ -0.79 & 1.07 \end{smallmatrix}\right)$. In this case, the variable *Day* is not statistically significant (with p-value being 0.76).

In Figure 3, we observe that the pixel intensities generally increase and then decrease over time, and generally reach their peaks at around day 10. Besides, there is considerable variability among dogs in these plots. This motivates us to consider the hierarchical nonparametric quantile model in (2.1). That is, level-1 model in (5.1) is now represented as $Y_{ik} = m(X_{ik}) + \epsilon_{ik}$. Figure 4 displays $\widehat{\gamma}^{(1)}(x)$ based on the ordinary mean regression and $\widehat{\gamma}^{(1)}_{\tau}(x)$ at five different quantiles (i.e., $\tau = 0.05, 0.25, 0.5, 0.75, 0.95$) based on our proposed hierarchical nonparametric quantile regression. We first note that the fixed effects of the time factor (i.e., *Day*) monotonically decrease at lower quantiles (e.g., $\tau = 0.05, 0.25$) of the response *pixel*. Inversely, the fixed effects of the time factor monotonically increase at higher quantiles (e.g., $\tau = 0.75, 0.95$). Only in the case of median (i.e., $\tau = 0.5$), the fixed-effect function remains to be a constant (i.e., $\widehat{\gamma}^{(1)}_{0.5}(x) = 2.28$). Although it is generally believed that the results based on the ordinary mean regression should be close to those based on the quantile regression at median, we observe that the difference between $\widehat{\gamma}^{(1)}_{0.5}(x)$ and $\widehat{\gamma}^{(1)}(x)$ is $|\widehat{\gamma}^{(1)}_{0.5}(x) - \widehat{\gamma}^{(1)}(x)| = |2.28 - 0.15| = 2.43$. This substantial difference can be easily explained by the incorrect application of the linear model at level 1 of (5.1). To see this, we fit the data with a quadratic model at level 1 with respect to the variable *Day*, and get $\widehat{\gamma}^{(1)}(x) = 3.38$ which makes the difference less substantial. In fact, we can calculate the model comparison criterion AIC (i.e., the Akaike Information Criterion) and BIC (i.e., the Bayesian Information Criterion) for the linear model (5.1) and the quadratic model at level 1. Under the linear model, the AIC and BIC are respectively equal to 840.41 and 860.56. Under the quadratic model, the AIC and BIC dramatically increase to 889.43 and 901.13 respectively. Obviously, both AIC and BIC support the quadratic model at level 1. Nonetheless, neither models could correctly describe the underlying behavior of pixel intensity over time. In this regard, our proposed hierarchical nonparametric quantile regression model provides practitioners a powerful tool for exploring complicated underlying relationships.

6 Conclusion

In this article, we consider a hierarchical nonparametric quantile regression model for multi-level

data. Unlike traditional hierarchical quantile models, our model allows (1) nonparametric model at level 1; (2) heteroscedastic within-group errors; and (3) general non-normal random errors (e.g., with finite mean, variance, and covariance). Our simulations clearly demonstrate that the proposed model is robust to non-normal errors (e.g. Cauchy errors). We also show in the pixel intensity example that our method can be a powerful tool for exploring complicated underlying relationships between various quantile functions and covariates. However, none of the above are satisfactory in traditional hierarchical quantile modelling. Although our presentation is confined to 2-level data, similar results can be extended to multi-level data. Also, we can extend the proposed model to models in which the model at level 2 is also nonparametric.

Appendix

Proof of Lemma 1

$$\begin{aligned}\tau &= P\left(Y_{ij} \le F_{ij}^{-1}(\tau)|X_{ij}, W_i\right)\\ &\approx P\left(\widetilde{X}_{ij}^T W_i\gamma_\tau(x) + \widetilde{X}_{ij}^T U_i + \epsilon_{ij} \le F_{ij}^{-1}(\tau)\right)\\ &= P\left(\widetilde{X}_{ij}^T U_i + \epsilon_{ij} \le F_{ij}^{-1}(\tau) - \widetilde{X}_{ij}^T W_i\gamma_\tau(x)\right)\\ &= P\left(\xi_{ij} \le F_{ij}^{-1}(\tau) - \widetilde{X}_{ij}^T W_i\gamma_\tau(x)\right).\end{aligned}$$

Therefore, $F_{ij}^{-1}(\tau) - \widetilde{X}_{ij}^T W_i\gamma_\tau(x) \approx e_{ij}(\tau)$, which completes the proof of the lemma.

Proof of Lemma 2 Notice that

$$\begin{aligned}&E\{H(\mathbf{b}_n)\}\\ &= -\frac{1}{\sqrt{n\cdot\det(H)}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_{-\infty}^{\infty}\psi_\tau(z-(n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n - F_{ij}^{-1}(\tau))K_H(X_{ij}-x)\\ &\qquad\qquad\cdot W_i^T\widetilde{X}_{ij}dF_{ij}(z)\\ &= -\frac{1}{\sqrt{n\cdot\det(H)}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_{-\infty}^{\infty}\psi_\tau(v)K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}dF_{ij}(v+(n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n\\ &\qquad\qquad + F_{ij}^{-1}(\tau))\\ &= -\frac{1}{\sqrt{n\cdot\det(H)}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\int_{-\infty}^{0}(\tau-1)dF_{ij}(v+(n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n + F_{ij}^{-1}(\tau))\right.\\ &\quad\left.+\int_{0}^{\infty}\tau dF_{ij}(v+(n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n + F_{ij}^{-1}(\tau))\right)K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\\ &= -\frac{1}{\sqrt{n\cdot\det(H)}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}[(\tau-1)F_{ij}((n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n + F_{ij}^{-1}(\tau))\\ &\quad+\tau\{1-F_{ij}((n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n + F_{ij}^{-1}(\tau))\}]K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\\ &= -\frac{1}{\sqrt{n\cdot\det(H)}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{\tau - F_{ij}((n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n + F_{ij}^{-1}(\tau))\}K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\\ &= -\frac{1}{\sqrt{n\cdot\det(H)}}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{\tau - F_{ij}(F_{ij}^{-1}(\tau)) - F_{ij}'(F_{ij}^{-1}(\tau))(n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n\right.\\ &\quad\left.-\frac{1}{2}\mathbf{b}_n^T W_i^T\widetilde{X}_{ij}\widetilde{X}_{ij}^T W_i\mathbf{b}_n\cdot F_{ij}''(F_{ij}^{-1}(\tau)+\theta(n\cdot\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n)\right\}K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\end{aligned}$$

$$
\begin{aligned}
&= \left\{ \frac{1}{n\cdot \det(H)} \sum_{i=1}^{n}\sum_{j=1}^{n_i} F'_{ij}(F_{ij}^{-1}(\tau)) W_i^T \widetilde{X}_{ij} K_H(X_{ij}-x) \widetilde{X}_{ij}^T W_i \right\} \mathbf{b}_n \\
&\quad + \frac{1}{\sqrt{(n\cdot\det(H))^3}} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \left\{ F''_{ij}(F_{ij}^{-1}(\tau) + \theta(n\cdot\det(H))^{-1/2} \widetilde{X}_{ij}^T W_i \mathbf{b}_n) \right. \\
&\quad \left. \cdot \frac{1}{2} \mathbf{b}_n^T W_i^T \widetilde{X}_{ij} \widetilde{X}_{ij}^T W_i \mathbf{b}_n \right\} K_H(X_{ij}-x) W_i^T \widetilde{X}_{ij}, \\
&= \Delta_n \mathbf{b}_n + \nu_n, \quad \text{where } \theta \in [0,1].
\end{aligned}
$$

Now, it remains to prove $\sup_{\|\mathbf{b}_n\|_\infty \le C} \|\nu_n\|_\infty = o(1)$ for some fixed $C > 0$. By Assumption 1, $|F''_{ij}(\cdot)| = O(1)$. By Assumptions 2, 3 and 7, we easily get $\sup_{\|\mathbf{b}_n\|\le C} \|\nu_n\|_\infty \le O((n\det(H))^{-1/2})\ \|\mathbf{b}_n\|_\infty^2$ for some fixed $C > 0$ as $n \to \infty$.

Proof of Lemma 3 We write $U \le V$ for $U, V \in R^f$, if $U'e_j \le V'e_j,\ j = 1, \ldots, f$, where $e_j = (0, \ldots, 1, \ldots, 0)'$. Denote $H^*(\mathbf{b}_n) = H(\mathbf{b}_n) - EH(\mathbf{b}_n) - H(\mathbf{0})$. It follows from the proof of Lemma 2 that $EH(\mathbf{0}) = 0$. For any $U, V \in R^f$ with $\|U\|_\infty = O(1), \|V\|_\infty = O(1)$, and $U \le V$, we have

$$
\begin{aligned}
H^*(U) - H^*(V) = &\left[- (n\det(H))^{-1/2} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \big\{ \psi_\tau(Y_{ij} - (n\det(H))^{-1/2} \widetilde{X}_{ij}^T W_i U - F_{ij}^{-1}(\tau)) \right. \\
&\left. - \psi_\tau(Y_{ij} - (n\det(H))^{-1/2} \widetilde{X}_{ij}^T W_i V - F_{ij}^{-1}(\tau)) \big\} K_H(X_{ij}-x) W_i^T \widetilde{X}_{ij} \right] \\
&- \left[(n\det H)^{-1/2} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \big\{ F_{ij}((n\det(H))^{-1/2} \widetilde{X}_{ij}^T W_i U + F_{ij}^{-1}(\tau)) \right. \\
&\left. - F_{ij}((n\det(H))^{-1/2} \widetilde{X}_{ij}^T W_i V + F_{ij}^{-1}(\tau)) \big\} K_H(X_{ij}-x) W_i^T \widetilde{X}_{ij} \right],
\end{aligned}
$$

in which the last equation follows from the fifth equality in the proof of Lemma 2. Under Assumptions 2 and 3, we could easily verify that $E\|H^*(U) - H^*(V)\|_\infty^4 \le O((n\det(H)^{-1}))\|U - V\|^8$. By Schwarz's inequality: $P(E_1 \cap E_2) \le P^{1/2}(E_1)P^{1/2}(E_2)$ for any events E_1 and E_2. By the refined Markov Inequality/Chebyshev Inequality: If $g : [0,\infty) \to [0,\infty)$ is a strictly increasing and nonnegative function, then $P\{|X| \ge \epsilon\} \le E[g(|X|)]/g(\epsilon)$ for any random variable X and $\epsilon > 0$, therefore, for $U, V, W \in R^f$ and $g(t) = t^4$,

$$
\begin{aligned}
&P\{\|H^*(W) - H^*(V)\|_\infty \ge \lambda,\ \|H^*(V) - H^*(U)\|_\infty \ge \lambda\} \\
&\quad \le P^{1/2}\{\|H^*(W) - H^*(V)\|_\infty \ge \lambda\} \cdot P^{1/2}\{\|H^*(V) - H^*(U)\|_\infty \ge \lambda\} \\
&\quad \le \big\{\lambda^{-4} E\|H^*(W) - H^*(V)\|_\infty^4\big\}^{1/2} \cdot \big\{\lambda^{-4} E\|H^*(V) - H^*(U)\|_\infty^4\big\}^{1/2} \\
&\quad \le \lambda^{-2} O((n\det(H))^{-1/2}) \|W - V\|_\infty^4 \cdot \lambda^{-2} O((n\det H))^{-1/2}) \|V - U\|_\infty^4 \\
&\quad \le \lambda^{-4} O(n\det(H)^{-1}) \|W - U\|_\infty^4.
\end{aligned}
$$

The result comes immediately from an extension of Theorem 12.1 in [32] to the vector arguments (see, Lemma 3.1 in [29]).

Proof of Lemma 5 We adopt the approach from [33]. Let $\widehat{\gamma}_\tau(x)$ be a solution of the minimization problem (2.4). It follows from Lemma A.2 of [33] that

$$
\frac{1}{\sqrt{n\cdot\det(H)}} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \psi_\tau(Y_{ij} - \widetilde{X}_{ij}^T W_i \widehat{\gamma}_\tau(x) - e_{ij}(\tau)) K_H(X_{ij}-x) W_i^T \widetilde{X}_{ij} = O_p(n^{-1/2}). \quad \text{(A1)}
$$

Furthermore, the partial first-order conditions for (2.4) can be shown to be

$$(n\det(H))^{-1/2}\nabla(R_n)(\mathbf{b}_n) = (n\det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-\widetilde{X}_{ij}^T W_i\mathbf{b}_n - e_{ij}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij} = o_p(1). \quad \text{(A2)}$$

Hence,

$$(n\det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-\widetilde{X}_{ij}^T W_i\{\mathbf{b}_n-\gamma_\tau(x)\}-F_{ij}^{-1}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij} = O_p(1), \quad \text{(A3)}$$

which is an immediate result from Lemma 1.

It is noteworthy that these partial first order conditions may not be zero exactly. Hence, we would like to relax condition (A2). By Lemma 5.2 of [34], we know that, for any $\epsilon>0$, there exist $\delta>0,\eta>0$ and $N\in\mathcal{N}$, s.t.

$$P\Bigg\{\inf_{\|\mathbf{\Delta}\|_\infty\geq\delta}(n\det(H))^{-1/2}\Bigg\|\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-(n\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{\Delta}-F_{ij}^{-1}(\tau)) \cdot K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\Bigg\|_\infty<\eta\Bigg\}<\epsilon. \quad \text{(A4)}$$

Let $A=\{(n\det(H))^{1/2}\|\widehat{\gamma}_\tau(x)-\gamma_\tau(x)\|_\infty\geq\delta\}$ and $B=\{\|(n\det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-\widetilde{X}_{ij}^T W_i\widehat{\gamma}_\tau(x)-e_{ij}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\|_\infty<\eta\}$. By the inequality $P(A)\leq P(AB)+P(\bar{B})$ and Lemma 1, for $n>N$, we have

$$\begin{aligned}
&P\{|(n\det(H))^{1/2}\|\widehat{\gamma}_\tau(x)-\gamma_\tau(x)\|_\infty\geq\delta\}\\
&\quad\leq P\Bigg\{(n\det(H))^{1/2}\|\widehat{\gamma}_\tau(x)-\gamma_\tau(x)\|_\infty\geq\delta,\|(n\det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau\\
&\qquad\cdot(Y_{ij}-(n\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\cdot(n\det(H))^{1/2}(\widehat{\gamma}_\tau(x)-\gamma_\tau(x))-F_{ij}^{-1}(\tau))\\
&\qquad\cdot K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\|_\infty<\eta\Bigg\}\\
&\qquad+P\Bigg\{\|(n\det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-\widetilde{X}_{ij}^T W_i\widehat{\gamma}_\tau(x)-e_{ij}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\|_\infty\\
&\qquad\geq\eta\Bigg\}\leq 2\epsilon,
\end{aligned}$$

which follows from (A1), (A3) and (A4). We complete the proof.

Proof of Theorem 1 According to Lemma 4, we have, for any $\|\mathbf{b}_n\|_\infty=O_p(1)$,

$$\begin{aligned}
&\Bigg\{-(n\det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-(n\det(H))^{-1/2}\widetilde{X}_{ij}^T W_i\mathbf{b}_n-F_{ij}^{-1}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\Bigg\}\\
&\quad-\Bigg\{-(n\det(H))^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-F_{ij}^{-1}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\Bigg\}-\Delta_n\mathbf{b}_n=o_p(1). \quad \text{(A5)}
\end{aligned}$$

Here, we take $\mathbf{b}_n=\sqrt{n\cdot\det(H)}(\widehat{\gamma}_\tau(x)-\gamma_\tau(x))$, and note that the first term on the left of (A5) is $o_p(1)$,which follows from (A2). It then follows from Lemma 5 that

$$\sqrt{n\cdot\det(H)}(\widehat{\gamma}_\tau(x)-\gamma_\tau(x))$$
$$=\frac{1}{\sqrt{n\cdot\det(H)}}\Delta_n^{-1}\bigg\{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-F_{ij}^{-1}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\bigg\}+o_p(1), \tag{A6}$$

which completes the proof.

Proof of Theorem 2 After some tedious algebras, we have $\mathrm{Var}(\psi_\tau(Y_{ij}-F_{ij}^{-1}(\tau)))=\tau(1-\tau)$, which follows from the proof of Lemma 2. Denote $\Gamma_n=(n\cdot\det(H))^{1/2}\{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\psi_\tau(Y_{ij}-F_{ij}^{-1}(\tau))K_H(X_{ij}-x)W_i^T\widetilde{X}_{ij}\}$. It can easily verified that

$$\mathrm{Var}(\Gamma_n)=\frac{1}{n\det(H)}\tau(1-\tau)\bigg\{\sum_{i=1}^{n}\sum_{j=1}^{n_i}W_i^T\widetilde{X}_{ij}K_H^2(X_{ij}-x)\widetilde{X}_{ij}^TW_i^T\bigg\}=\tau(1-\tau)\Theta_n. \tag{A7}$$

Finally, it is easy to see from Assumptions 4–5 that $\sqrt{n\cdot\det(H)}(\widehat{\gamma}_\tau(x)-\gamma_\tau(x))$ satisfies the standard multivariate Lindeberg condition and is therefore asymptotically a Gaussian vector. The covariance matrix follows from (A6) and (A7).

Proof of Theorem 3 The proof is basically a modification and extension of proofs given in, for example, [35], [36] and [38]. Here, we simply provide the main idea and skip all technical details, which are available from the aforementioned reference. Let $\zeta_{ij}=Y_{ij}-\widetilde{X}_{ij}^TW_i\gamma(x)$. All we need to do is checking whether the three conditions shown in [35] are satisfied.

1 The process $\{\zeta_{ij}\}(i=1,\ldots,n,\ j=1,\ldots,n_i)$ satisfies the so-called ϕ-mixing condition. Roughly speaking, we say $\zeta_{i1},\ldots,\zeta_{in_i}$ is ϕ-mixing if ζ_{ij} and $\zeta_{i(j+k)}$ become independent as k becomes large (see, [37]). In particular, we have already known that $\zeta_{ij},j=1,\ldots,n_i$ are independent, which is the special case.

2 Clearly, $G_{ij}(t)=F_{ij}(\widetilde{X}_{ij}^TW_i\gamma(x)+t)$. Hence, the cumulative function, $G(t)$, is absolutely continuous because F(t) is absolutely continuous under Assumption 1.

3 Obviously, $G'(t)$ is finite, positive, and absolutely continuous for all $t=G^{-1}(\tau)$ and $0<\tau<1$ since $F'(t)$ satisfies the conditions given in Assumption 1. Similarly to the proofs given in, for example, [38], we have $E(\widehat{e}_{ij}(\tau))=e_{ij}-\frac{\tau(1-\tau)F''(\widetilde{X}_{ij}^TW_i\widehat{\gamma}_\tau(x)+e_{ij})}{2(n_i+2)F'^3(\widetilde{X}_{ij}^TW_i\widehat{\gamma}_\tau(x)+e_{ij})}+O(1/n_i)$ and $\mathrm{Var}(\widehat{e}_{ij}(\tau))=\frac{\tau(1-\tau)}{(n_i+2)F'^2(\widetilde{X}_{ij}^TW_i\widehat{\gamma}_\tau(x)+e_{ij})}+O(1/n_i)$. Hence, we complete the proof.

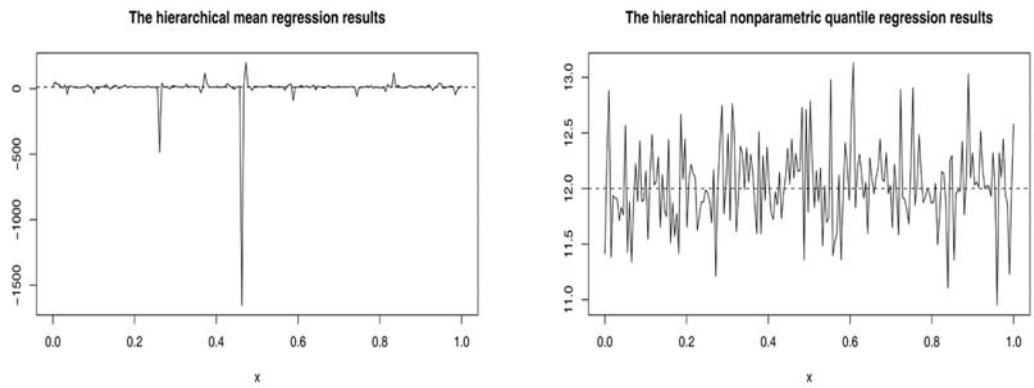

Figure 1 The horizontal dotted line indicates the true marginal effect. The solid line on the left plot is the fitted marginal effect function based on the ordinary mean regression and the solid line on the right plot is the fitted median ($\tau = 0.5$) marginal effect function based on our proposed hierarchical nonparametric quantile regression

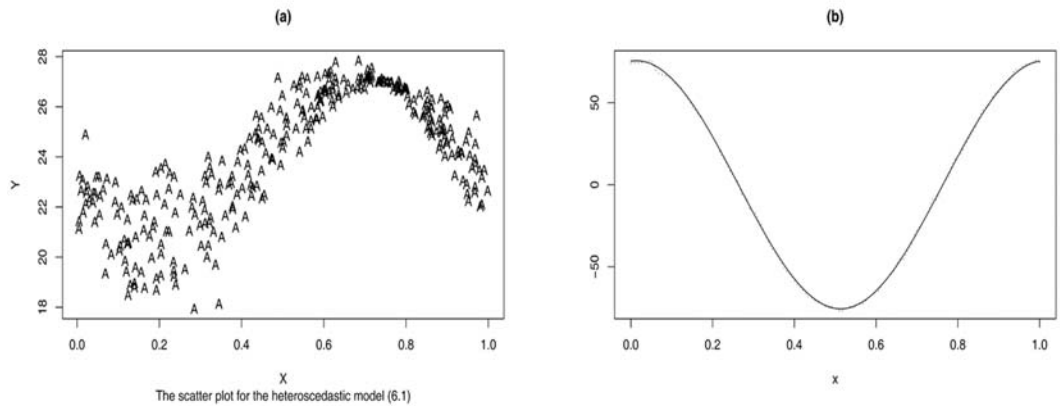

Figure 2 The scatter plot for heteroscedastic model (4.1) is shown in (a); The solid line in (b) indicates the the real fixed effect function at quantile 0.5; The dotted line in (b) is the estimates of the real fixed effect function at quantile 0.5

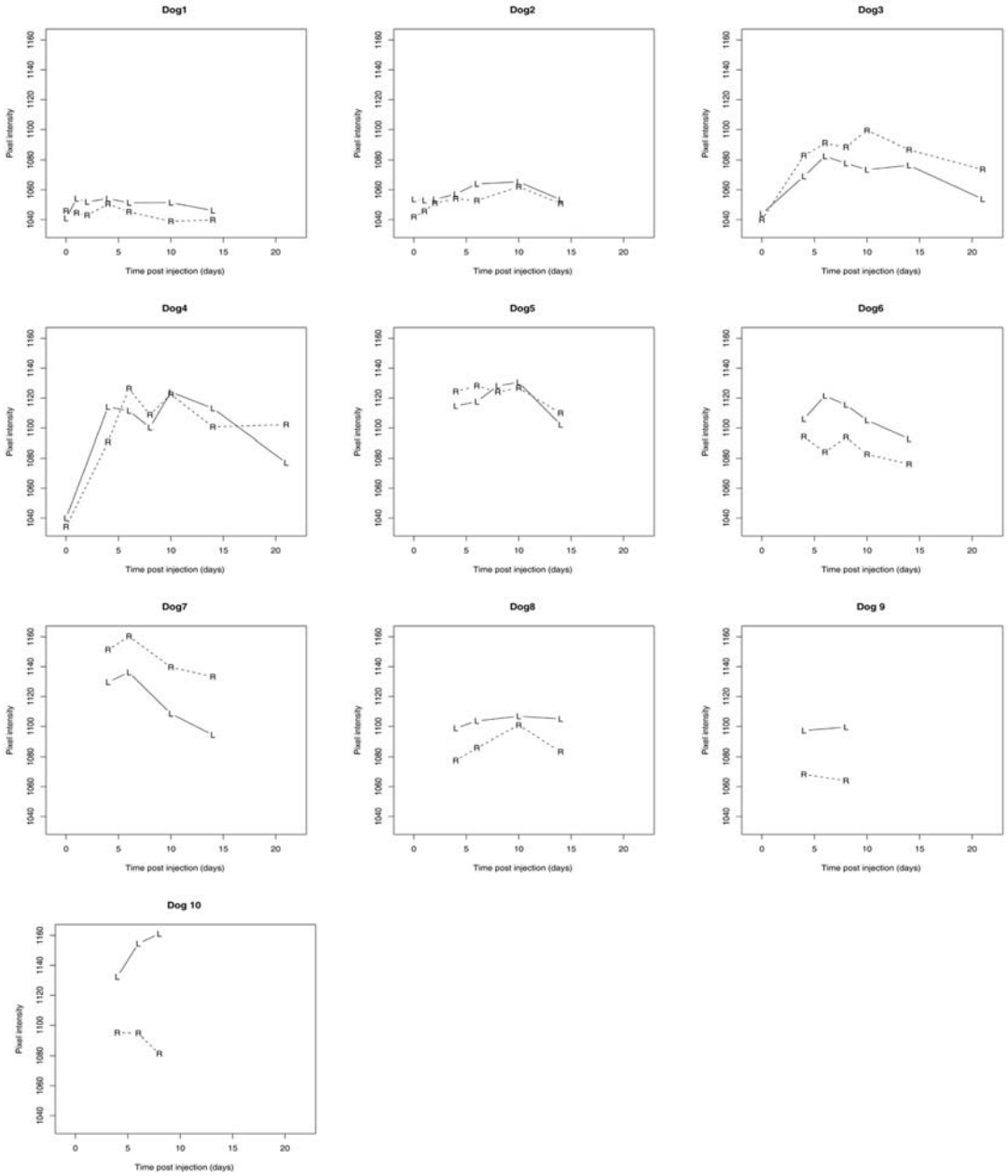

Figure 3 Pixel intensity on CT scans over time for lymph nodes on the left indicated by the dashed line with the letter "L" and on the right indicated by the dotted line with the letter "R"

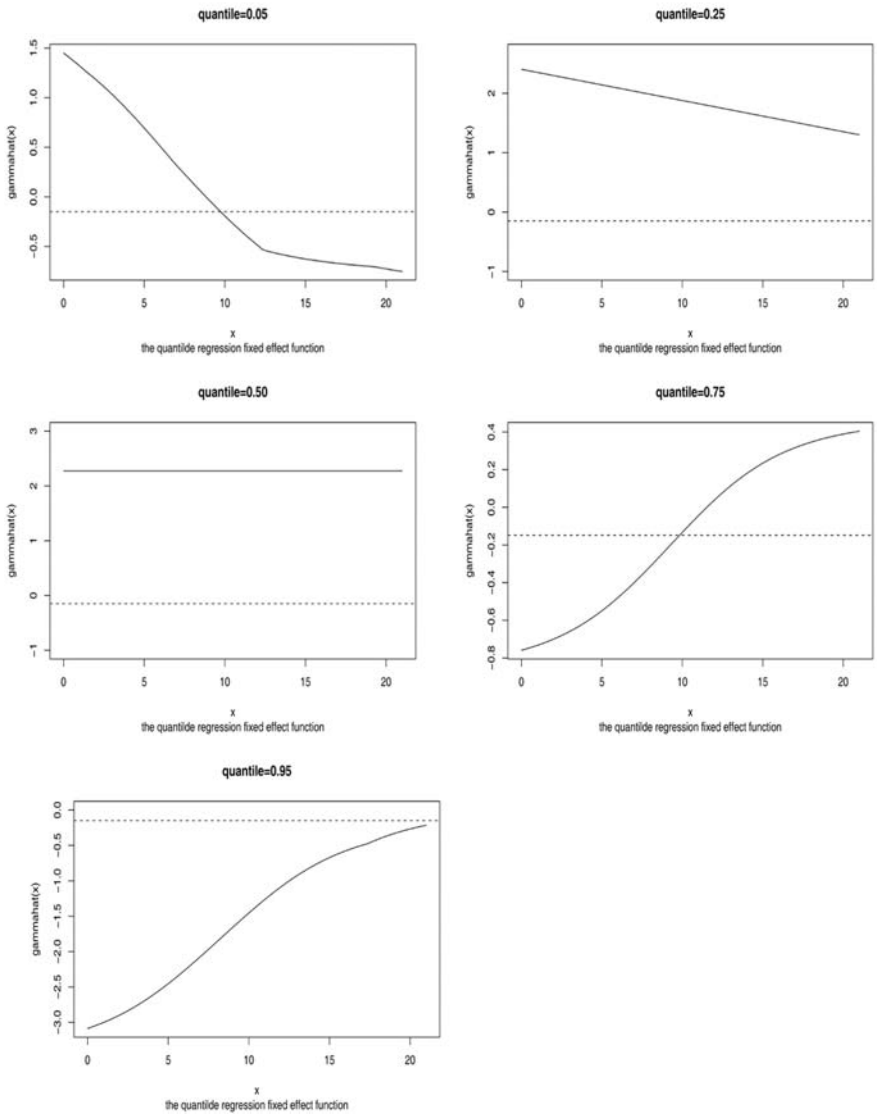

Figure 4 The solid lines represent the estimated fixed-effect functions for τ ranging from 5% to 95%, respectively. The horizontal dotted line indicates the mean regression results based on the hierarchical linear model

References

[1] Lindley, D. V., Smith, A. F. M.: Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34,** 1–41 (1972)

[2] Smith, A. F. M.: A general Bayesian linear model. *Journal of the Royal Statistical Society, Series B*, **35,** 67–75 (1973)

[3] Goldstein, H.: Multilevel statistical models, (2nd ed), New York, John Wiley, 1995

[4] Mason, W. M., Wong, G. M., Entwistle, B.: Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.) Sociological methodology, San francisco, Jossey–Bass, 1983, 72–103

[5] Elston, R. C., Grizzle, J. E.: Estimation of time response curves and their confidence bands. *Biometrics,* **18,** 148–159 (1962)

[6] Singer, J. D.: Using SAS PROC MIXED to fit multilevel models, hierarchical models and individual growth models. *Journal of Educational and Behavioral Statistics,* **23**, 323–355 (1998)

[7] Rosenberg, B.: Linear regression with randomly dispersed parameters. *Biometrika,* **60,** 61–75 (1973)
[8] Longford, N.: Random coefficient models, Oxford, Clarendon, 1993
[9] Dempster, A. P., Rubin, D. B., Tsutakawa, R. K.: Estimation in covariance components models. *Journal of the American Statistical Association,* **76**, 341–353 (1981)
[10] Longford, N.: A fast scoring algorithm for maximum likelihood estimation in unbalanced models with nested random effects. *Biometrika,* **74**, 817–827 (1987)
[11] Koenker, R., Bassett, G.: Regression quantiles. *Econometrica,* **46**, 33–50 (1978)
[12] Bhattacharya, P. K., Gangopadhyay, A. K.: Kernel and nearest-neighbor estimation of a conditional quantile. *Ann. Statist,* **18**, 1400–1415 (1990)
[13] Chaudhuri, P.: Nonparametric estimates of Regression Quantiles and their local Bahadur Representation. *The Annals of Statistics,* **2**, 760–777 (1991)
[14] Fan, J., Hu, T. C., Truong, Y. K.: Robust nonparametric function estimation. *Scandinavian Journal of Statistics,* **21,** 433–446 (1994)
[15] Koenker, R., Ng, P., Portnoy, S.: Quantile smooth splines. *Biometrika,* **81,** 673–680 (1994)
[16] Yu, K., Jones, M. C.: Local linear quantile regression. *Journal of the American statistical Association,* **93,** 228–238 (1998)
[17] De Gooijer, J. G., Gannoun, A., Zerom, D.: Mean square error properties of the kernel-based multistage median predictor for time series. *Statistics & Probability letters,* **56**, 51–56 (2002)
[18] Tian, M. Z., Chen, G. M.: Hierarchical Linear Regression Models for Conditional Quantiles. *Science in China Series A, Mathematics,* **49,** 11–16 (2006)
[19] Chaudhuri, P., Doksum, K., Samarov: On average derivative quantile regression. *The Annals of Statistics,* **25,** 715–744 (1997)
[20] Fan, J., Farmen, M.: Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society, Series B*, **60**, 591–608 (1998)
[21] Bryk, A. S., Rausendenbush, S. W.: Hierarchical Linear Models, SAGE Publications, Inc., 1992
[22] Huttenlocher, J. E., Haight, W., Bryk, A. S., Seltzer, M.: Early vocabulary growth: relation to language input and gender. *Developmental Psychology,* **27**, 236–249 (1991)
[23] Silverman, B. W.: Density Estimation for Statistics and Data Analysis, Vol. 26 of Monographs on Statistics and Applied Probability, Chapman and Hall, London, 1986
[24] Scott, D. W.: Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, New York, Chichester, 1992
[25] Ruppert, D., Sheather, S. J., Wand, M. P.: An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90,** 1257–1270 (1995)
[26] Gooijer, Zerom: On additive conditional quantiles with high-dimensional covariates. *Journal of the American Statistical Association,* **98,** 135–146 (2003)
[27] Hogg, R. V., Craig, A. T.: Introduction to Mathematical Statistics, 5-th ed. New York, Macmillan, 1995
[28] Müller, H. G., Stadtmüller, U.: Estimation of herteroscadasticity in regression analysis. *The Annals of Statistics,* **15**, 610–625 (1987)
[29] Jurečkovă, J., Sen, P. K.: On adaptive scale-equivariant M-estimators in linear models. *Statistics & Decisions, Supplement Issue*, **1**, 31–46 (1984)
[30] Hendricks, W., Koenker, R.: Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American statistical Association*, **87,** 58–68 (1992)
[31] Koenker, R., Machado, J. A. F.: Goodness of fit and related inference processes for quantile regression. *Journal of the American statistical Association*, **94**, 1296–1310 (1999)
[32] Billingsley, P.: Convergence of Probability Measures, New York: John Wiley & Sons, Inc., 1968
[33] Ruppert, D., Carroll, R. J.: Trimmed least-squares estimation in the linear model. *Journal of the American Statistical Association,* **75**, 828–838 (1980)
[34] Jurečkovă, J.: Regression quantiles and trimmed least squares estimator under a general design. *Kybernetika*, **20,** 345–356 (1984)
[35] Sen, P. K.: On the Bahadur Representation of sample quantiles for sequences of ϕ-mixing random variables. *Journal of Multivariate analysis,* **2**, 77–95 (1972)
[36] Chen, E. J., Kelton, W. D.: Simulation-based estimation of quantiles, Proceedings of the 1999 Winter Simulation Conference, ed. P.A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 428-434. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, 1999
[37] Billingsley, P.: Convergence of Probability Measures, 2nd ed. New York, John Wiley & Sons, Inc., 1999
[38] David, H. A.: Order Statistics, 2nd ed. New York: Wiley, 1981