Diagnosing Outliers and Visualization of Quantile Regression Models

Wenjing Wang[1], Dianne Cook[2], Earo Wang[2]
[1]Renmin University of China , [2]Monash University

Contents

1 Abstract

Quantile regression model becomes more and more popular and has been widely used in many research areas. Recently, Koenker reviewed the studies of quantile regression in the past 40 years, and pointed out that more diagnostic tools should be provided in addition to the extensive toolbox of estimation. This paper first comprehensively studied the outlier diagnostic for quantile regression and implement the discussed method in R language. The R package `quokar` contains functions for diagnosing outliers in quantile regression and also provide support for easily visualizing the diagnose results. In addition, this paper proposed a general framework to visualize quantile regression model (linear and non-linear) in high dimensional data space using `GGobi`. The integrated plot of quantile regression model and the fitting data set was displayed in `GGobi`, which is a more straight-forward way of examing outliers and also supported other research aspects. This paper discusses the methods used in `quokar` and illustrates the usefulness of the package through examples. At last, the visualization framework and results for quantile regression are provided.

2 Introduction

Quantile regression model has been widely used in many research areas such as economy, finance and social science (most recent researches are Autor, Houseman and Kerr (2017), Mitchell, Dowda and Pate (2017), Gallego-Álvarez and Ortas (2017), Korobilis (2017), Maciejowska, Nowotarski and Weron (2016)). Quantile regression has significant advantages over mean regression mainly on two aspects: (a) observed covariates can describe the whole distribution of response variable which produce comprehensive results; (b) estimators can still maintain optimal properties of in case of heteroscedasticity or heavy tail distribution.

The research scope of quantile regression has been broadened considerably in the past decades. We surveyed some of the most recent developments. Koenker (2004), Geraci and Bottai (2006) conducted quantile regression for longitudinal data. Longitudinal data introduced a large number of "fixed effects" in quantile regression and these "fixed effects" will significantly inflate the variability of estimates of other covariate effects. They proposed using l1 regularization methods as essential computational tools. Parente and Santos Silva (2016) studied properties of the quantile regression estimator when data are sampled from independent and identically distributed clusters. They provided a consistent estimator of the covaiance matrix and showed the regression estimator is consistent and asympototically normal. Researchers also studied the intersection of quantile regression and panel data. Panel data potentially allows the researcher to include fixed effects to control unobserved covariates which extend the original quantile regression model. They presented new model format and fixed effects estimation.

Due to the advantages of quantile regression model held, researches also interested in ebbeding it in other models to enhance model features or conduct better results analysis. Geraci (2014) proposed linear quantile mixed model which dealt with within-subject dependence by embeding subject-specific random intercepts into quantile regression model. Estimation strategies to reduce the computational burden and inefficiency using EM algorithm. Chernozhukov and Hansen (2006) proposed instrumental variable quantile regression to evaluate the impact of endogenous variables or treatments on the entire distribution of outcomes. They modifies the conventional quantile regression and recovers quantile-specific covariate effects in an instrumental variables model.

Except for the recent progresses in model updating, extensive researches have been done in model inferencing. Gutenbrunner, Jureckova, Koenker, and Portnoy (1993) proposed rank-based inference to deal problems of constructing confidence intervals for individual quantile regression parameter estimates. In order to quantify the robustness of inferencing, resampling methods are carefully studied and used (Hahn (1995), Horowitz (1998), Fitzenberger (1998), He and Hu (1999)). Koneker and Machado (1999) used Kolmogorov-Smironov method to measure the goodness of fit for quantile regression. To tackle the "Durbin problem", Koenker and Xiao (2002) developed location shift and location-scale shift test for quantile regression process. There are also studies of quantile regression in bayesian framework (Yu and Moyeed (2001), Yu and Stander (2007), Jkozumi and Kobayashi (2011), Santos and Bolfarine (2016)), which widely extended the research framework.

Based on above numerous of methodology and application studies of quantile regression, varieties of toolboxes conduct model fitting and inference has been developed. Koenker(2017) also pointed out that more work needs to be done to develop better diagnostic tools for quantile regression models. Free software R offers several packages implementing quantile regression, most famous `quantreg` by Roger Koenker, but also `gbm`, `quantregForest`, `qrnn`, `ALDqr` and `bayesQR`. However, few model diagnostic methods were proposed for quantile regression and no toolbox for model diagnostic were implemented in R.

Outlier detection is one aspect of model diagnosing, and it is important in regression analysis because the results of regression can be sensitive to these outliers. Data for regression model may have special points located far away from others either in response variable data or in the space of the predictors. The latter also be called leverage points. In single variabe case, we can easily observe the data based on scatter plot which will help us spot outliers. Difficulty lies in high-dimensional situation, where statistical methods should be used. To deal with this, various methods for detecting outliers have been studied (Atkinson 1994; Barnett and Lewis 1994; Becker and Gather 1999, 2001; Davies and Gather 1993; Gather and Becker 1997; Gnanadesikan and Kettenking 1972; Hadi 1992, 1994; Hawkins 1980; Maronna and Yohai 1995; Penny 1995; Rocke and Woodruff 1996; Rousseeuw and Van Zomeren 1990). Commonly used methods include residuals, leverage value, studentized residuals and jacknife residuals.

In regression context, classic least ordinary square estimation of linear regresssion can be expressed as $\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$, $\hat{Y} = X(X^{'}X)^{-1}X^{'}Y = HY$, where, H is called hat matrix. Residuals can be write as $\hat{\epsilon} = Y - \hat{Y}(1-H)Y = (1-H)\epsilon$. Hence, considering the influence of outliers in vertical direction and leverage points at the same time, we should use studentized residuals, which is $r_i = \frac{\hat{\epsilon}_i}{\sigma^2\sqrt{1-h_i}}$. The larger r_i, the more suspicious the outlier is. Another widely used outlier diagnositc framework is `leave-one-out`. Jackknife residual and Cook's distance (Cook (1977)) are constructed based on this idea. These diagnostic statistics has already become available on widely distributed statistical software packages SAS, SPSS, as well as R.

Due to the different estimation methods and estimator form of quantile regression, outlier diagnosing for quantile regression model should be specially discussed. In addition, quantile regressions can be fitted on every quantile interested, which add difficulties in applying diagnosing methods and displaying results simutaneously. Benites, Lachos, and Vilca (2016) developed case-deletion diagnostics for quantile regression using the asymmetric Laplace distribution. Santos and Bolfarine (2016) discussed Bayesian quantile regression and considered using the posterior distribution of the latent variable for outlier diagnosing. Some simple outlier diagnostic methods for quantile regression can be conducted in statistical software SAS using procedure `QUANTREG`. However, to the authors' knowledge, these methods are still not be impleted in R. To fill the gap, an implementation in R language now is available in rencently developed package `quokar` which provides several outlier diagnostic methods as well as supportive visualization results for quantile regression.

This article aims to introduce R package `quoakr` and display supportive visualizaitons for quantile regression models in high dimension. The remainder of this article is organized as follows: In Section 2, we provide a general introduction to quantile regression model and its robusness property. In Section 3 we give a tour of outlier diagnostic methods for quantile regression used in package `quokar`. In section 4 we will show how to conduct diagnostic methods in package `quokar`. In section 5, we displayed supportive visualizations for quantile regerssion in high-dimension and non-linear situations. The current limitations and future research and development directions are discussed in Section 6.

3 Robustness of Quantile Regression

Koenker and Bassett (1978) first proposed linear model as

$$y_i = x_i^{'}\beta_\tau + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

The τth quantile function of the sample is $Q_y(\tau|x) = x^{'}\beta(\tau)$. Based on the idea of minim izing a sum of asymmetrically weighted absolute residuals, the objective function of quantile regression

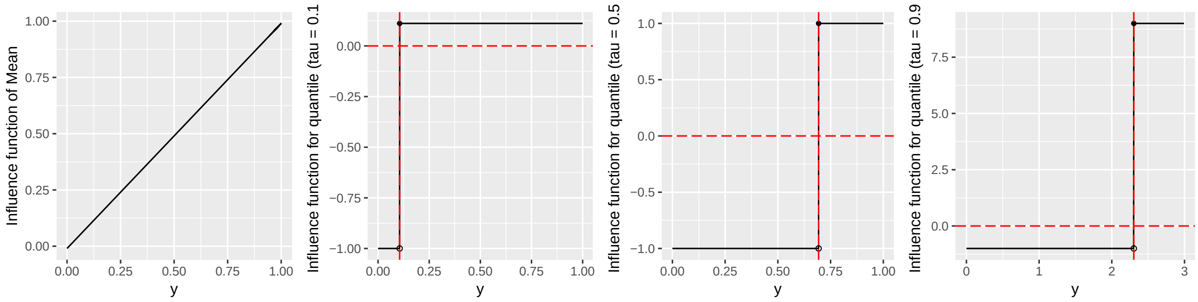

Figure 1: Visualization of influence function for Mean and Quantile. It is obviously that quantile influence functions on quantile 0.1, 0.5 and 0.9 are bounded which indicat that quantile is more robust then Mean. The boundaries of influence function on low and high quantile are asymmetrical.

model is,

$$\min_{\beta_\tau \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^{'}\beta_\tau) \tag{2}$$

where $\rho(.)$ is loss function which was defined as $\rho_\tau(u) = u(\tau - I(u < 0))$. In addition, assuming $Y_1, ..., Y_n$ is a sequence of i.i.d random variables which has distribution function F and continuous density function f. The coefficience vector $\hat{\beta}_\tau$ is asymptotically normal, which can be expressed as,

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \xrightarrow{d} N(0, \tau(1-\tau)D^{-1}\Omega_x D^{-1}) \tag{3}$$

where $D = E(f(\mathbf{X}\beta)\mathbf{X}\mathbf{X}^{'})$ and $\Omega_x = E(\mathbf{X}^{'}\mathbf{X})$.

Quantile is more robust than mean when extreme values exist in the dataset interested. This property applies equally in regression context. Onyedikachi (2015) discussed the robustness of quantile and quantile regression using influence function.

Set T as a functional of F, the influence function is the directional derivative of $T(F)$ at F, and it measures the effect of a small perturbation in F on $T(F)$. For Mean, the influence function is

$$IF(y; T; F) = y - T(F) \tag{4}$$

For the τth quantile points, influence function can be expressed as,

$$IF(y; T; F) = \begin{cases} \dfrac{\tau}{f(F^{-1}(\tau))}; & y > F^{-1}(\tau) \\ \dfrac{(\tau - 1)}{f(F^{-1}(\tau))}; & y \leq F^{-1}(\tau) \end{cases} \tag{5}$$

where f is the density function of F. Comparing (**??**) and (5), the latter obviously has boundary when y is changing. To explain the characteristic of the boundaries on different quantile, we provide visualization results with an example. Data are generated from distribution function $F(x) = 1 - e^{-\lambda x} \quad x > 0$, and the density function and inverse distribution function are $f(x) = e^{-x}$, $Q(\tau) = -ln(1-p)$ respectively.

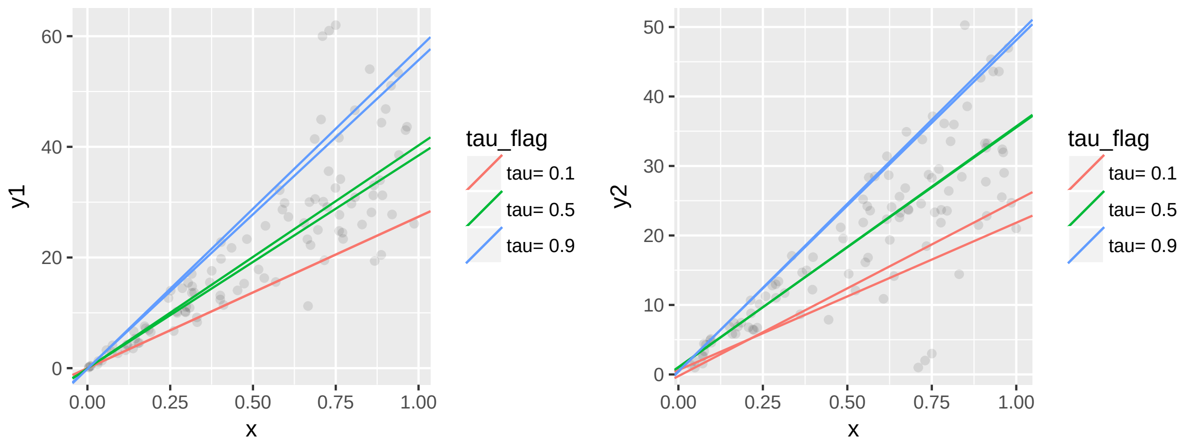

Figure 2: Fitting quantile regression model on quantile 0.1, 0.5 and 0.9 using simulated datasets with and without outliers. The outliers located at the top-left of the original dataset. Results show that outliers pull up the slope of the 0.9 and 0.1 regression line. When outliers located at the bottom-right of the original dataset, results show that outliers pull down the slope of the 0.1 regression line.

For quantile regression, suppose F represent the joint distribution of the pairs (x, y), the influce function is

$$IF((y,x), \hat{\beta}_{F(\tau)}, F) = Q^{-1} x sgn(y - x^{'} \hat{\beta}_F(\tau)) \tag{6}$$

where

$$dF = dG(x) f(y|x) dy \tag{7}$$

$$Q = \int x x^{'} f(X^{'} \hat{\beta}_F(\tau)) dG(x) \tag{8}$$

Equation (6) implies that quantile regression estimates will not be affected by changes in value of dependent variable as long as the relative positions of the observation points to the fitted plane are maintained.

We generate 100 sample observation and 3 outliers to see the relation between outlier location and the change of coefficients. The outliers are located at top-left and bottom-right of the original data. Figure 2 show that the former pulled up the regression lines on quantile 0.9 and 0.5, and the latter pulled down them.

We also conduct simulations to study the robustness of quantile regression by generating 100 data with 5 condaminated points considered as outliers. In each experiment, we changed the y axis value of the outliers. The results show that when outliers moving down in y direction for 10 unit, outliers pull down the slope on every quantile (by comparing the result of rq(y1~x) and rq(y2~x)). However, keeping moving down the outliers does no change to regression slopes. This reflect the theory of bounded influence function.

To observed the change of coefficients in multi-variable model, we fit quantile regression model $y = x1 + x2 + \epsilon$. The results show that coefficients changes slowly when keep moving down the outliers in y-direction.

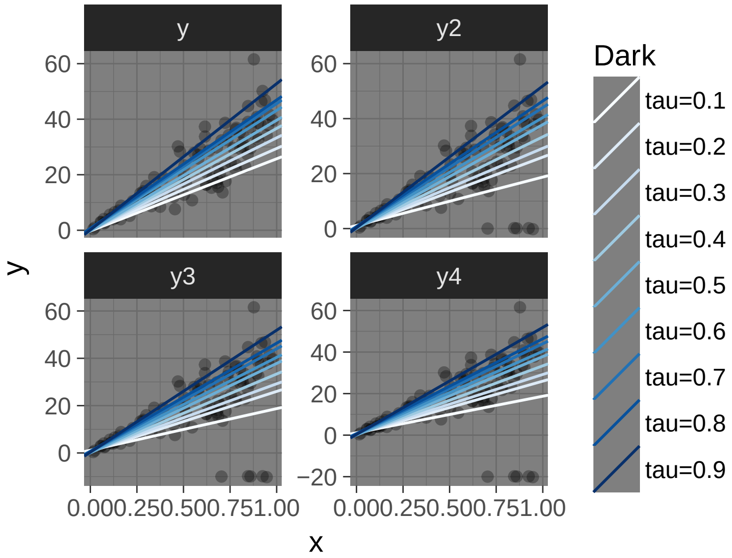

Figure 3: Fitting quantile regression models using simulated data. We keep moving down the outliers in y direction in y2 (y-5), y3 (y-10) and y4 (y-15) to see the change of regression lines.

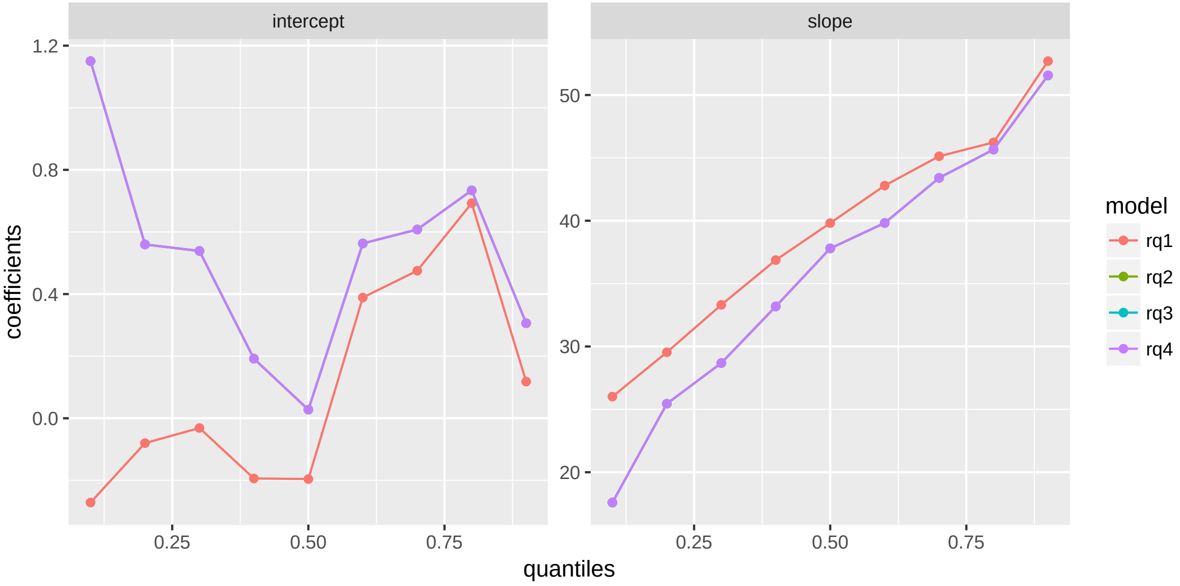

Figure 4: Fitting quantile regression models using simulated data. We keep moving down the outliers in y direction getting datasets with variable y2 (=y-5), y3 (=y-10) and y4 (=y-15). Calculating the estimated coefficients in each experiment and results show that in single predictor case, outliers moving down in y make no difference to the quantile regression coefficients estimations

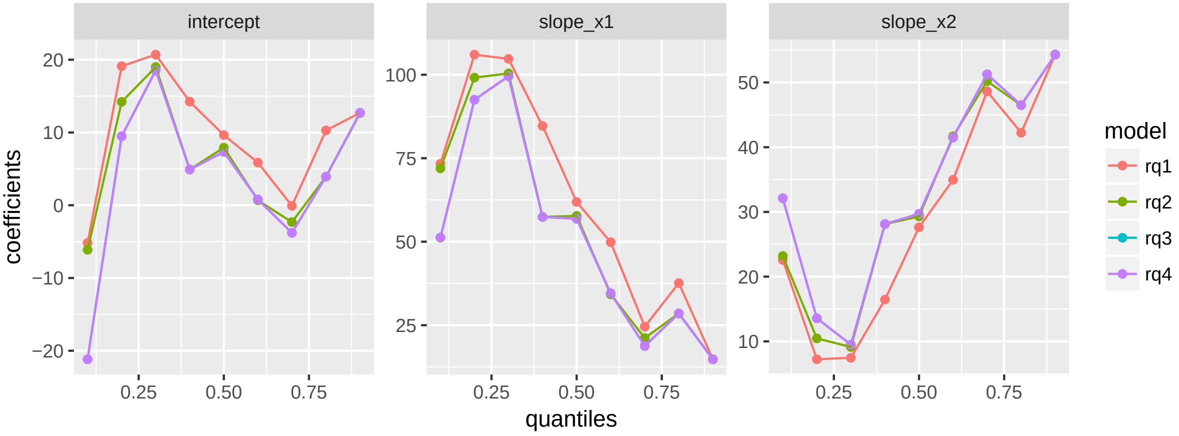

Figure 5: Fitting quantile regression models using simulated data. We keep moving down the outliers in y direction getting three datasets with different locations of outliers (changing in y-aixs, y2 (=y-5), y3 (=y-10) and y4 (=y-15)). Results show that in multi predictors case, outliers moving down in y make small change to the quantile regression coefficients estimations

If moving outliers in same pattern moving on x direction, slopes change greatly every time outliers move. To go further, each move has different effect on different quantiles.

In conclusion, quantile regression response differently to outliers comparing mean regression in two aspects: (a) not all models on each quantile will be affected when outliers exist. If we are interested in model on particular quantile, the effect of outliers should be carefully considered. (b) quantile regression model do not have robustness properties to so called leverage points.

4 Outlier Diagnostic Methods for Quantile Regression

In this section we briefly introduce diagnostic methods used in `quokar`. These methods are well discussed in recent literatures and performed well in our application. We assume a basic knowledge of quantile regerssion model and Baysian methods.

4.1 Residual-Robust Distance

In quantile regression, we can not use the famous "Hat Matrix" to detect leverage points since the coefficient estimation of quantile regression do not satisfy $\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$. One way to identify possible leverage points is to calculate a distance from each point to a "center" of the data. Leverage point would then be the one with a distance larger than some predetermined cutoff. A conventional measurement is Mahalanobi distance:

$$MD(x_i) = [(x_i - \bar{x})^{'}\bar{\mathbf{C}}(\mathbf{A})^{-1}(x_i - \bar{x})]^{1/2} \qquad (9)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{\mathbf{C}}(\mathbf{A}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^{'}(x_i - \bar{x})$ are the empirical multivariate location and scale respectively. However, the standard sample location and scale parameters are not robust to outliers. In addition, datasets with multiple outliers or clusters of outliers are subject to problems of masking and swamping (Pearson and Chandra Sekar 1936). Such problems of unrobust, masking and swamping can be resolved by using robust estimates of shape and location, which by definition are less affected by outliers (Rousseeuw and van Zomeren

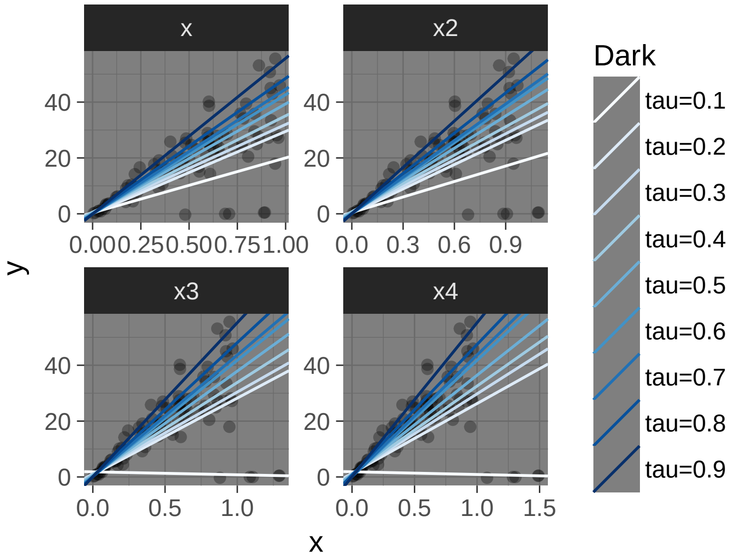

Figure 6: Fitting quantile regression models using simulated data. We keep moving the outliers to the right in x direction getting three datasets with different locations of outliers (changing in x-aixs, x2 (=x+0.2), x3 (=x+0.4) and x4 (=x+0.6)) to see the change of regression lines.

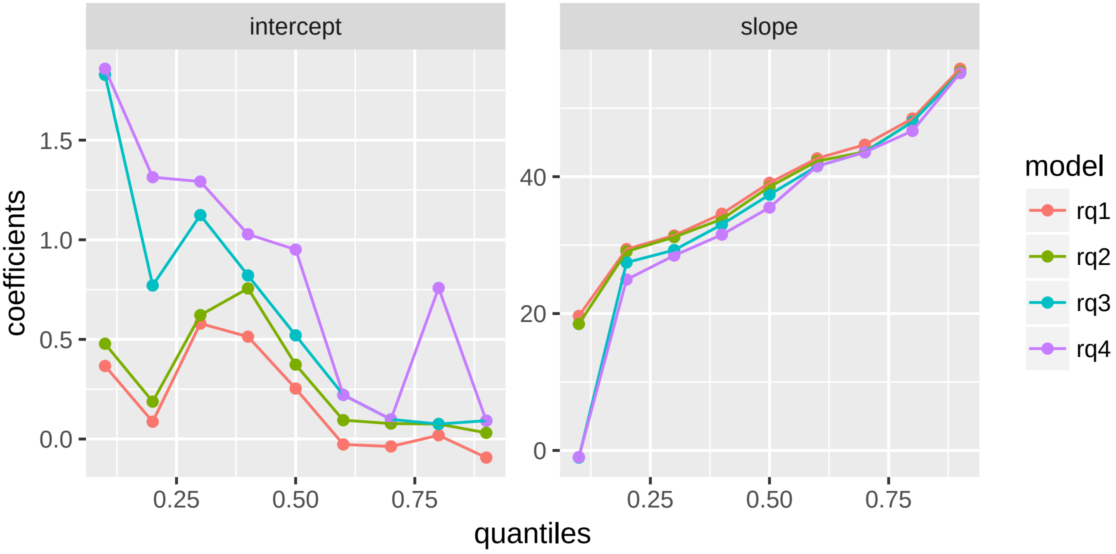

Figure 7: Fitting quantile regression models using simulated data. Keep moving the outliers to the right in x direction getting three datasets with different locations of outliers (changing in x-aixs, x2 (=x+0.2), x3 (=x+0.4) and x4 (=x+0.6)). Calculating the estimated coefficients in each experiment and results show that outliers moving in x aixes make larger difference to the quantile regression coefficients then moving in y aixes.

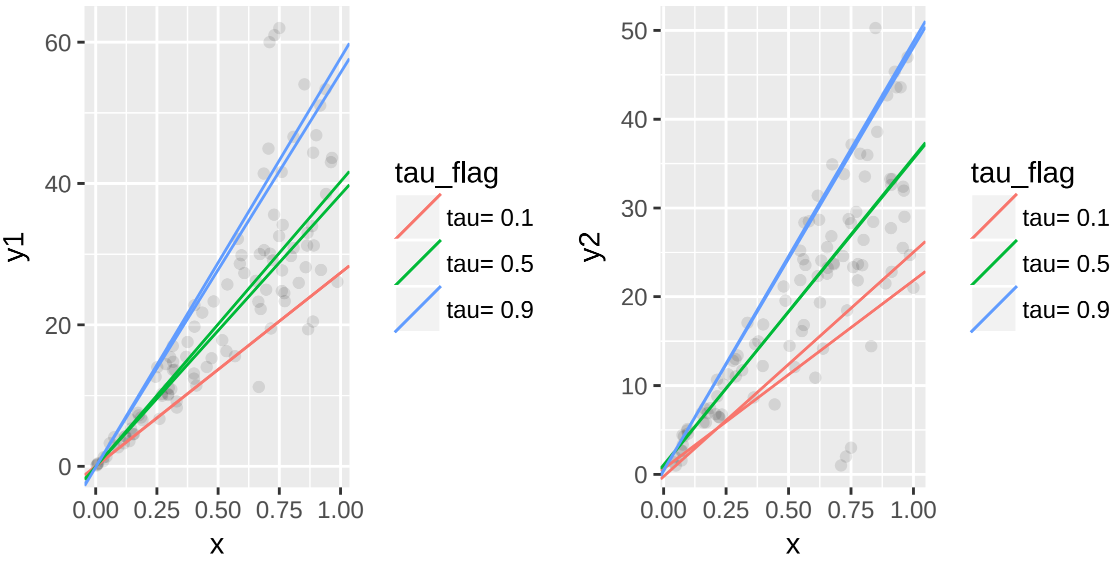

Figure 8: Fitting quantile regression models using simulated data. Keep moving the outliers to the right in x direction getting three datasets with different locations of outliers (changing in x-aixs, x2 (=x+0.2), x3 (=x+0.4) and x4 (=x+0.6)). Calculating the estimated coefficients in each experiment and results show that outliers moving in x aixes make larger difference to the quantile regression coefficients then moving in y aixes.

(1991)). We use Rousseeuw's minimum covariance determinant (MCD) proposed by Rousseeuw and Van Driessen (1999) to estimate the location and scale of the data.

The MCD estimator can be defined as:

$$MCD = (\bar{X}_h^*, S_h^*) \tag{10}$$

where X and S stand for location and scale. $h = p : |S_h^*| < |S_k^*|, |k| = p$, $\bar{X}_h^* = \frac{1}{p}\sum_{i \in p} x_i$, $S_p^* = \frac{1}{p}\sum_{i \in p}(x_i - \bar{X}_p^*)(x_i - \bar{X}_p^*)'$.

The value p can be thought of as the minimum number of points which must not be outliers. The MCD has its highest possible breakdown at $h = [\frac{n+p+1}{2}]$ where [.] is the greatest integer function. Because we are interested in outlier detection, we will use h at its highest possible breakdown. $h = [\frac{n+p+1}{2}]$ in our calculations, and we refer to a sample of size h as a "half sample" The MCD is omputed from the "closet" half sample, and therefore, the outlying points will have little affect on the MCD location or shape estimate. With MCD, we can calculate robust distance which was defined as,

$$RD(x_i) = [(x_i - \mathbf{T}(\mathbf{A}))'\mathbf{C}(\mathbf{A})^{-1}(x_i - \mathbf{T}(\mathbf{A}))]^{1/2} \tag{11}$$

Where $\mathbf{T}(\mathbf{A})$ and $\mathbf{C}(\mathbf{A})$ are robust multivariate location and scale estimates that are computed according to the MCD.

Package `quokar` provide Mahalanobi distance and robust distance to detect leverage points in quantile regression. Residuals that are based on quantile regression estimates are used to detect vertical outliers.

4.2 Cook's Distance and Likelihood Distance

Case-deletion diagnostics such as Cook's distance or Likelihood distance have been successfully applied to various statistical models. Based on the research of Sánchez, Lachos and Labra (2013), we calculate Cook's distance and Likelihood distance for quantile regression in package `quokar`. More specify process will be discussed as follows.

Yu and Moyeed (2001) proposed random variable Y distributed as asymmetric Laplace distribution with location parameter μ, scale parameter $\sigma > 0$ and skewness parameter $\tau \in (0, 1)$ has density function:

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} exp{-}\rho_p(\frac{(y-\mu)}{\sigma}) \tag{12}$$

where $\rho_\tau(.)$ is the loss function mentioned above.

Suppose that $y_i \sim ALD(\mathbf{x}_i^{'}\beta_p, \sigma, \tau)$, $i = 1, ..., n$ are independent. The likelihood function for n observations is

$$L(\beta, \sigma|y) = \frac{\tau^n(1-\tau)^n}{\sigma^n} exp{-}\sum_{i=1}^{n} \rho_\tau(\frac{y_i - \mathbf{x}_i^{'}}{\sigma}) \tag{13}$$

For note, a quantity with a subscript '[i]' means the relevant quantity with the ith observation deleted. Let $\hat{\theta}$ and $\hat{\theta}^*_{[i]}$ be the maximum likelihood estimator of $theta$ based on $L(\theta|Y)$ and $L(\theta|Y_{[i]})$ respectively. Cook's distance CD_i is given by **??**. For external norms, M is usually chosen to be $-L(\ddot{Y}|\theta)$.

$$CD_i = (\hat{\theta}^1_{[i]} - \hat{\theta})^{'} M(\hat{\theta}^1_{[i]} - \hat{\theta}) \tag{14}$$

Alternatively, another measure of difference between θ and $\theta^*_{[i]}$ is the observed data likelihood function which is defined as Likelihood distance.

$$LD_i = L(\hat{\theta}|Y) - L(\hat{\theta}^1_{[i]}|Y) \tag{15}$$

The ith observation is regarded as influential if the value of Cook's distance or Likelihood distance is relatively large. Sánchez and Lachos (2015) proposed a EM algorithm to calculate the above Cook's distance and Likelihood distance which reduced the calculation burden. They used the expectation of likelihood function.

$$Q(\theta|\hat{\theta}) = E\{L(\theta|Y)|\hat{\theta}\} \tag{16}$$

To assess the influence of the ith case, we will consider the function

$$Q_{[i]}(\theta|\hat{\theta}) = E\{L(\theta|Y_{[i]})|\hat{\theta}\} \tag{17}$$

Let $\hat{\theta}_{[i]}$ be the maximiser of $Q_{[i]}(\theta|\hat{\theta})$. The one-step approximation $\hat{\theta}_{[i]}$ is

$$\hat{\theta}_{[i]} = \hat{\theta} + \{-\ddot{Q}(\hat{\theta}|\hat{\theta})\}^{-1} \dot{Q}_{[i]}(\hat{\theta}|\hat{\theta}) \tag{18}$$

where

$$\ddot{Q}(\hat{\theta}|\hat{\theta}) = \frac{\partial^2 Q(\theta|\hat{\theta})}{\partial\theta\partial\theta^T}|_{\theta=\hat{\theta}}$$

$$\dot{Q}_{[i]}(\hat{\theta}|\hat{\theta}) = \frac{\partial Q_{[i]}(\theta|\hat{\theta})}{\partial\theta}|_{\theta=\hat{\theta}}$$

are the Hessian matrix and the gradient vector evaluated at $\hat{\theta}$, respectively.

The Cook's distance is

$$GD_i = (\hat{\theta}_{[i]} - \hat{\theta})^T\{-Q(\hat{\theta}|\hat{\theta})\}(\hat{\theta}_{[i]} - \hat{\theta}), i = 1, ..., n \tag{19}$$

The measurement of the influence of the ith case is based on the Q function, similar to the likelihood distance LD_i which was defined as

$$QD_i = 2\{Q(\hat{\theta}|\hat{\theta}) - Q(\hat{\theta}_{[i]}|\hat{\theta})\} \tag{20}$$

4.3 Mean Posterior Probability and Kullback-Leibler Divergence

In Bayesian quantile regression framework, Kozumi and Kobayashi (2011) proposed a location-scale mixture representation of the asymmetric Laplace distrbution, as follows

$$Y|v \sim N(\mu + \theta v, \phi^2\sigma v) \tag{21}$$

where $\theta = (1-2\tau)/(\tau(1-\tau))$, $\phi^2 = 2/(\tau(1-\tau))$. v is a latent variable which prior distribution is exponential and the full conditional posterior distribution for each vi follows generalized inverse Gaussian distribution with parameters

$$v = \frac{1}{2}, \quad \delta_i^2 = \frac{(y_i - x_i^{'}\beta(\tau))^2}{\phi^2\sigma}, \quad \gamma^2 = \frac{2}{\sigma} + \frac{\theta^2}{\phi^2\sigma} \tag{22}$$

Parameters of v_i in @ref{eq:parameters} show two characters of latent variable v: (a) each random variable v_i has different distributions due to parameter δ^2 changes among obvervations. (b) distribution of v_i depended on weighted squared residual of the quantile fit. Based on the above two characters, we propose to compare the posterior distribution of its latent variable to detect outliers. We implete two methods in `quokar`, one is mean posterior prability and the other is Kullback-Leibler divergence.

We define variable O_i indicating whether observation i is an outlier.

$$O_i = \begin{cases} 1, & i \quad is \quad outlier \\ 0, & i \quad is \quad normal \end{cases}$$

The mean posterior probability appoximatlly calculated by MCMC draw is

$$P(O_i = 1) = \frac{1}{n-1}\sum_{j \neq i}\frac{1}{M}I(v_i^{(l)} > \max_{k\in 1:M} v_j^{(k)})$$

where M is the size of the chain of v_i after the burn-in perior and $v_i^{(l)}$ is the lth draw of this chain.

Kullback and Leibler (1951) proposed a more precise method of measuring the distance between variables. Suppose f_i is the posterior conditional distribution of v_i and correspondingly f_j is the posterior conditional distribution of v_j. The Kullback-Leibler divergence of f_i and f_j is defined as

$$K(f_i, f_j) = \int log(\frac{f_i(x)}{f_j(x)} f_i(x))dx$$

Similar with calculating mean posterior probability, we average this divergence for one observation based on the distance from all others,

$$KL(f_i) = \frac{1}{n-1} \sum_{j \neq i} K(f_i, f_j)$$

The outliers should show a high probability value for this divergence. We compute the integral using the trapezoidal rule, and the density function are estimated using kernel estimation with Gaussian kernel function.

5 Examining Outlier Detection

We developed R package `quokar` to implete quantile regression outlier diagnostic methods. This package mainly realized two basic features: (a) plot the outlier state; (b) plot data with outliers marked. `quokar` is available from Github at https://github.com/wenjingwang/quokar, so to install and load withn R use:

```r
devtools::install_github("wenjingwang/quokar")
library(quokar)
```

We implete AIS data as an example to introduce this package. AIS data include 14 variables for 100 female atheletes.

5.1 Plot the outlier state

In single variable case, we can use scatter plot to represent the outlier state. The following code showed how to display suspicious outliers based on quantile regression models. Figure 6 showed the potential outlier is case 1 and 75. When comes to multi-variable case, one way to display the outlier state in data by the scatter plot on separate covariants. Figure 7 showed case 56 and 75 are suspicious outliers in the data set.

```r
data(ais)
ais_female <- filter(ais, Sex == 1)
case <- 1 : nrow(ais_female)
ais_female <- cbind(case, ais_female)
coef_rq <- coef(rq(BMI ~ LBM, tau = c(0.1, 0.5, 0.9),
                   data = ais_female, method = "br"))

br_coef <- data.frame(intercept = coef_rq[1, ],
                      coef = coef_rq[2, ],
```

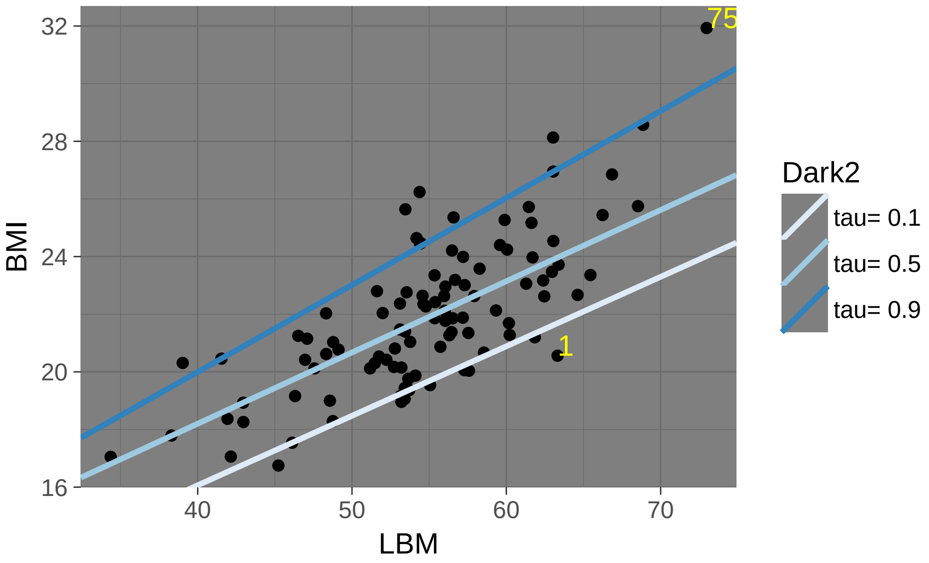

Figure 9: Plot the outlier state for single variable case.

```
                             tau_flag = colnames(coef_rq))
ggplot(ais_female)+
  geom_point(aes(x = LBM, y = BMI)) +
  geom_abline(data = br_coef, aes(intercept = intercept,
                                  slope = coef,
                                  colour = tau_flag), size = 1) +
  geom_text(data = subset(ais_female, case %in% c(1, 75)),
                          aes(x = LBM, y = BMI, label = case),
            colour = "yellow",hjust = 0, vjust = 0) +
  scale_colour_brewer("Dark2") +
  theme_dark()
```

```
ais_female_f <- dplyr::select(ais_female, c(case, BMI, LBM, Bfat))
ais_female_f_long <- tidyr::gather(ais_female_f, variable, value, -case, -BMI)
ggplot(ais_female_f_long, aes(x = value, y = BMI))+
  geom_point(alpha = 0.5) +
  geom_text(data = subset(ais_female_f_long, case %in% c(56, 75)),
                          aes(x = value, y = BMI, label = case),
            colour = "yellow", vjust = 0, hjust = 0) +
  facet_wrap(~ variable, scales = "free_x") +
  scale_colour_brewer("Dark2") +
  theme_dark()
```

5.2 Plot data with outliers marked

Scatter plot has limitations when tackling multi-variable regression. In quokar, we provide functions to do outlier diagnostic which return the dataframe easily to plot data with outliers marked.

- residual-robust distance method

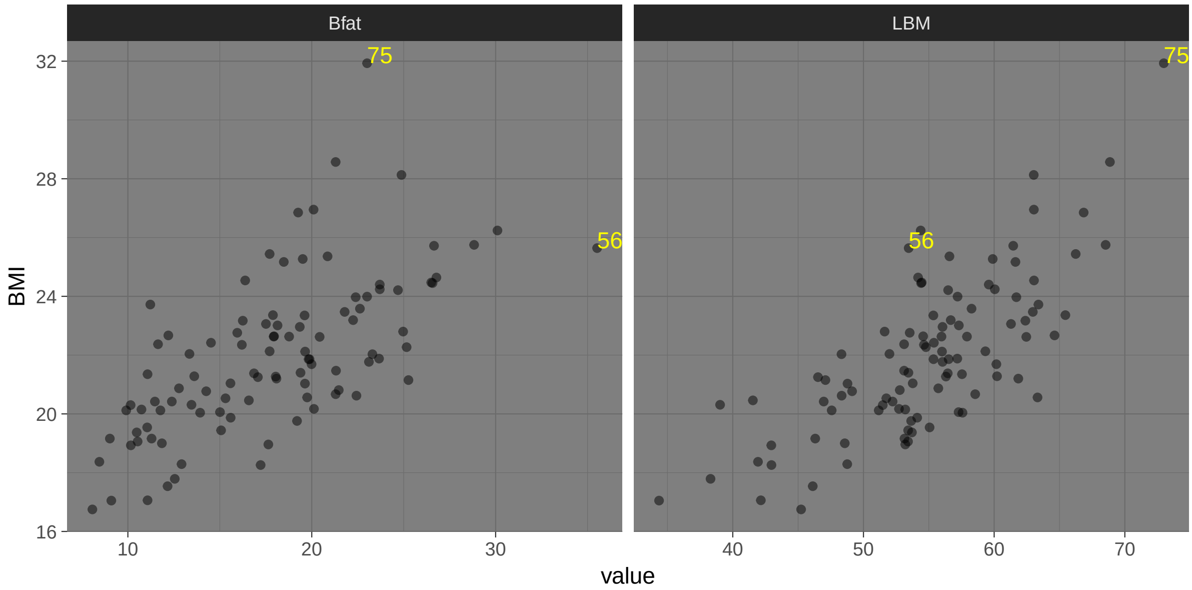

Figure 10: Plot the outlier state for multi-variable regression.

First, we calculate residuals, mahananobi distance and robust distance for quantile regression using function `plot_distance`. Simutaneously, it provides the cutoff value for identifying the outliers.

```
tau <- c(0.1, 0.5, 0.9)
object <- rq(BMI ~ LBM + Bfat, data = ais_female, tau = tau)
plot_distance <- frame_distance(object, tau = c(0.1, 0.5, 0.9))
distance <- plot_distance[[1]]
head(distance, 3)
```

```
##          md        rd tau_flag  residuals
## 1 1.2275233 1.3912428   tau0.1 -1.4630550
## 2 0.6988854 0.6486756   tau0.1 -0.9262022
## 3 0.3836449 0.3315911   tau0.1  1.0706377
```

```
cutoff_v <- plot_distance[[2]]; cutoff_v
```

```
## [1] 2.716203
```

```
cutoff_h <- plot_distance[[3]]; cutoff_h
```

```
## [1] 12.450378  6.917875 14.073312
```

Function `plot_distance` returns the tidy data form for plotting data with outliers marked together overlaying the cutoff lines. We use the following code for visualizing the diagnose result. Figure 8 showed, on quantile 0.1, 0.5 and 0.9, case 56, 75, 98 and 100 are detected as leverage points and no outliers in y-direction exsited.

```
n <- nrow(object$model)
case <- rep(1:n, length(tau))
distance <- cbind(case, distance)
distance$residuals <- abs(distance$residuals)
tau_f <- paste("tau", tau, sep="")
text_flag <- 1:length(cutoff_h) %>%
```

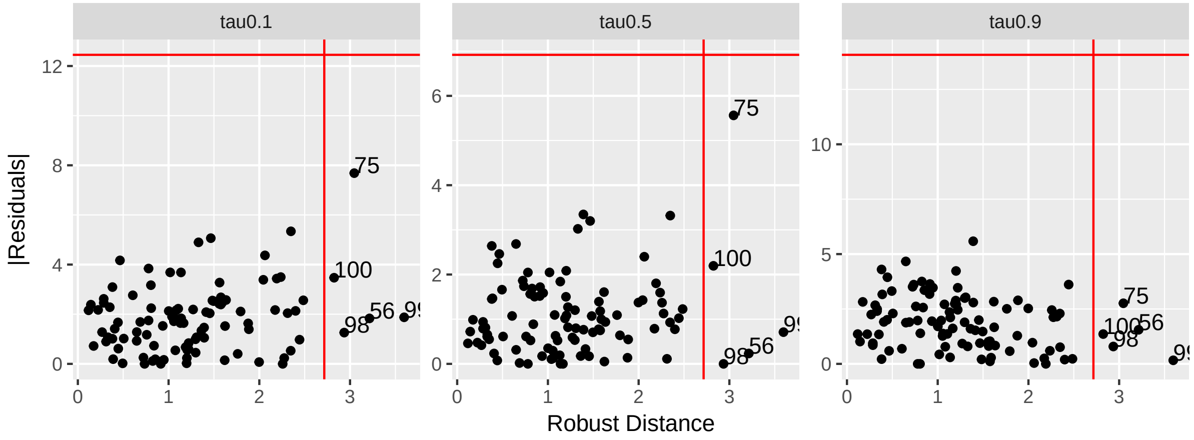

Figure 11: Robust Distance-Residual Plot. Points on the right of vertical cutoff line are considered leverage points and points above the horizental cutoff line are outliers in y-direction.

```
                    map(function(i){
                        distance %>%
                           filter((residuals > cutoff_h[i] |rd > cutoff_v)
                                & tau_flag == tau_f[i])})

text_flag_d <- rbind(text_flag[[1]], text_flag[[2]], text_flag[[3]])
ggplot(distance, aes(x = rd, y = residuals)) +
      geom_point() +
      geom_hline(data = data.frame(tau_flag = paste("tau", tau, sep=""),
                                   cutoff_h = cutoff_h),
                 aes(yintercept = cutoff_h), colour = "red") +
      geom_vline(xintercept = cutoff_v, colour = "red") +
      geom_text(data = text_flag_d, aes(label = case), hjust = 0, vjust = 0) +
      facet_wrap(~ tau_flag, scales = 'free_y') +
      xlab("Robust Distance") +
      ylab("|Residuals|")
```

- Generalized Cook distance and Q function distance

We apply generalized Cook distance and Q function distance methods in function `frame_mle` using AIS data. Methods `bayes.prob` and `bayes.kl` in function `frame_bayes` return the mean probability and Kullback-Leibler divergence of each observation on each given quantile. The results are also in tidy data structure which can be easily used for plotting the two distances with outliers marked. Figure 9 and 10 show regression model on 0.1 quantile has outlier case 1, and case 75 is the potential outlier of regression models on quantile 0.5 and 0.9.

```
y <- ais_female$BMI
x <- cbind(1, ais_female$LBM, ais_female$Bfat)
case <- rep(1:length(y), length(tau))
GCD <- frame_mle(y, x, tau, error = 1e-06, iter = 10000,
                  method = 'cook.distance')
GCD_m <- cbind(case, GCD)
ggplot(GCD_m, aes(x = case, y = value )) +
    geom_point() +
    facet_wrap(~variable, scale = 'free_y') +
```

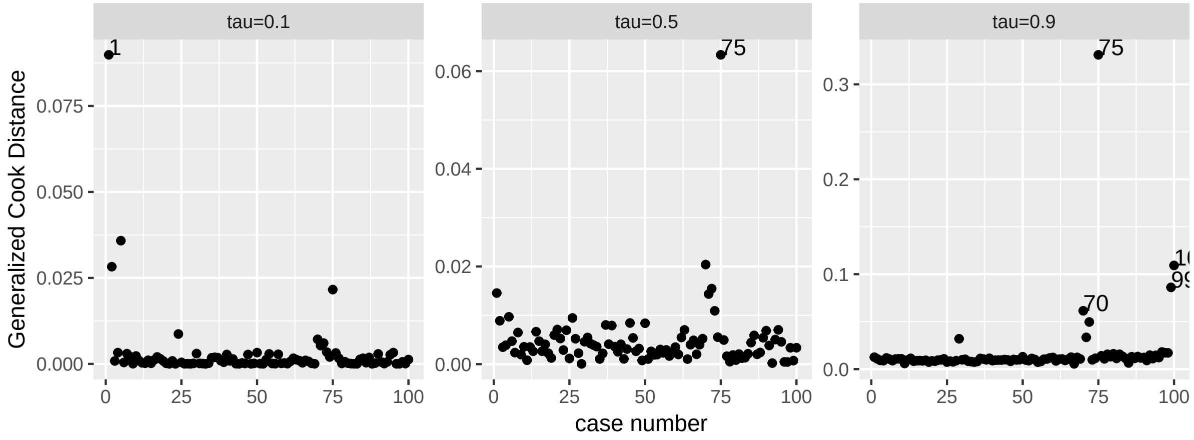

Figure 12: Generalized Cook distance of each observation on quantile 0.1, 0.5 and 0.9. Case 75 has relative large Cook distance-funtion distance to other points

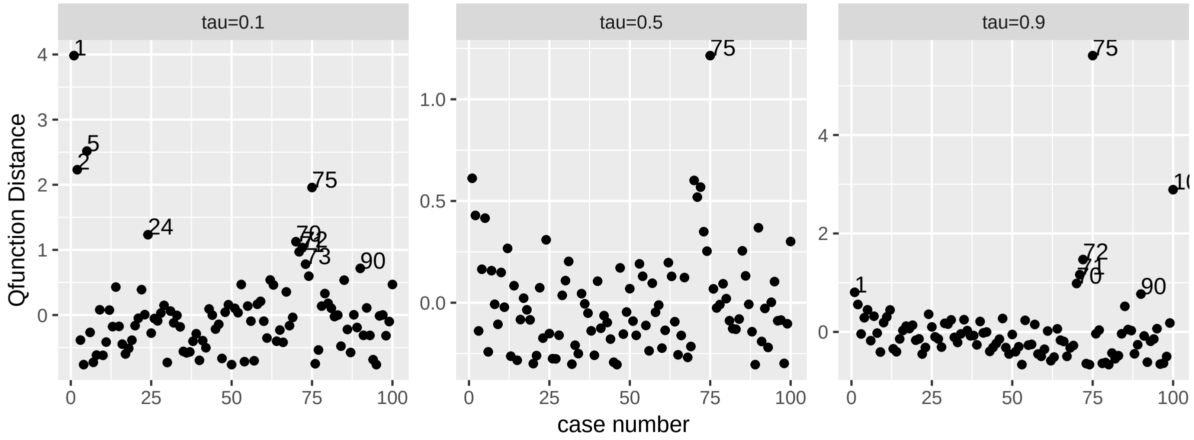

Figure 13: Q function distance of each observation on quantile 0.1, 0.5 and 0.9. Case 75 has relative large Q function distance to other points

```
    geom_text(data = subset(GCD_m, value > mean(value) + 2*sd(value)),
              aes(label = case), hjust = 0, vjust = 0) +
    xlab("case number") +
    ylab("Generalized Cook Distance")
```

```
QD <- frame_mle(y, x, tau, error = 1e-06, iter = 10000,
                method = 'qfunction')
QD_m <- cbind(case, QD)
ggplot(QD_m, aes(x = case, y = value)) +
 geom_point() +
 facet_wrap(~variable, scale = 'free_y')+
 geom_text(data = subset(QD_m, value > mean(value) + sd(value)),
           aes(label = case), hjust = 0, vjust = 0) +
 xlab('case number') +
 ylab('Qfunction Distance')
```

```
y <- ais_female$BMI
x <- matrix(c(ais_female$LBM, ais_female$Bfat), ncol = 2, byrow = FALSE)
```

```
tau <- c(0.1, 0.5, 0.9)
case <- rep(1:length(y), length(tau))
prob <- frame_bayes(y, x, tau, M =  10, burn = 1,
                 method = 'bayes.prob')
head(prob)
```

```
##   variable       value
## 1  tau=0.1 0.087542088
## 2  tau=0.1 0.108866442
## 3  tau=0.1 0.098765432
## 4  tau=0.1 0.067340067
## 5  tau=0.1 0.005611672
## 6  tau=0.1 0.185185185
```

```
kl <- frame_bayes(y, x, tau, M = 10, burn = 1,
                  method = 'bayes.kl')
head(kl)
```

```
##   variable     value
## 1  tau=0.1 0.2992471
## 2  tau=0.1 0.3253945
## 3  tau=0.1 0.4297409
## 4  tau=0.1 0.1994685
## 5  tau=0.1 0.4520945
## 6  tau=0.1 0.7489053
```

With the result which is long data form returned by function `frame_bayes`, we provide visualization of the mean posterior probability and Kullback-Leibler divergence of each observation with outlier marked. Figure 11 and 12 show that the potential outlier is case 75.

```
prob_m <- cbind(case, prob)
ggplot(prob_m, aes(x = case, y = value )) +
   geom_point() +
   facet_wrap(~variable, scale = 'free') +
  geom_text(data = subset(prob_m, value > mean(value) + 2*sd(value)),
            aes(label = case), hjust = 0, vjust = 0) +
   xlab("case number") +
   ylab("Mean probability of posterior distribution")
```

```
kl_m <- cbind(case, kl)
ggplot(kl_m, aes(x = case, y = value)) +
  geom_point() +
  facet_wrap(~variable, scale = 'free')+
  geom_text(data = subset(kl_m, value > mean(value) + sd(value)),
            aes(label = case), hjust = 0, vjust = 0) +
  xlab('case number') +
  ylab('Kullback-Leibler')
```

6 Generalized Framework for Visualizing Regrssion Model

Visualization is particularly useful and comprehensive way to explore data and model. It is also a extremely straight-forward way to detect outlier by observing the location of data and model. In regression context, a fitted regression model is not only judged by its prediction error, rather other questions worth to consider, such as do the data space is too inseparablely or too sparesly to be represented by the data; are there some regions that are difficultly for model to fit. For quantile regression, we are also curious to know what is the relative location of models on different quantiles in data space.

Use model visualization to discover useful information in fitted quantile regression and high dimensional data set is a challenging problem. There exists no work which aims to visualize the quantile regression itself. In this section, we propose approach aimed to visualize the whole data set together with the quantile regression models fitted on different quantile in one plot. In this way, we can analyze model fitting, model performance and model comparison simultaneously.

Given our visualization results, the exploring and observing aspects are organized into the following steps.

- Is data clustered or sparsely distributed? How do quantile regression models deal with that?
- Do model overfit/underfit exist?
- Are there potential outlier exist in the data and how methods treat these?
- How do models fitted on different quantiles located in different regions of data space?
- What about the relative locations of quantile regression models?

Our visualizations are realized by software `GGobi` (Swayne, Temple Lang, Buja and Cook 2003). `GGobi` is a free software for interactive and dynamic graphics which can be used with R via package `rggobi`. With `GGobi`, we can extend the limit of 2D visualization of quantile regression model and break the visualization barrier in 3D or much higher dimension.

We proposed a feasible framework to visualize quantile regression model in `GGobi`. Assumming given data set including points $\mathbf{x}_i \in \mathbf{X}$. In high dimension case, $\mathbf{x}_i$ is a vector and $\mathbf{X}$ is matrix. The ith value of responsor $\mathbf{Y}$ is y_i. The quantile regression model $f_\tau : \mathbf{X} \to \mathbf{Y}$ will be fitted on the given data set.

- Use grid method to generate data in data space bounded by $\mathbf{X}$. The generated data form data set $\mathbf{Z}$.
- Fit quantile regression models f_τ on every interested quantile and get the estimated parameters β_τ.
- Calculate quantile regression model using $f_\tau = \mathbf{Z}\beta_\tau$. In non-linear case, calculate model based on the non-linear curve form.
- Tidy data set $(\mathbf{Z}, f_\tau)$ for each quantile into long data form and add tag representing quantile.

6.1 Linear Case Result

In two predictors case (3D data space), quantile regression models are planes in space. We use ais data to fit models and visualize them with `GGobi`.

Figure 8 show that one point is isolated in data space from other points and three fitted quantile regression models which indicating this point is potential outliers for the models fitted. The three quantile regression models are not paralleled in the data space and they respectively fitted the data set on quantile.

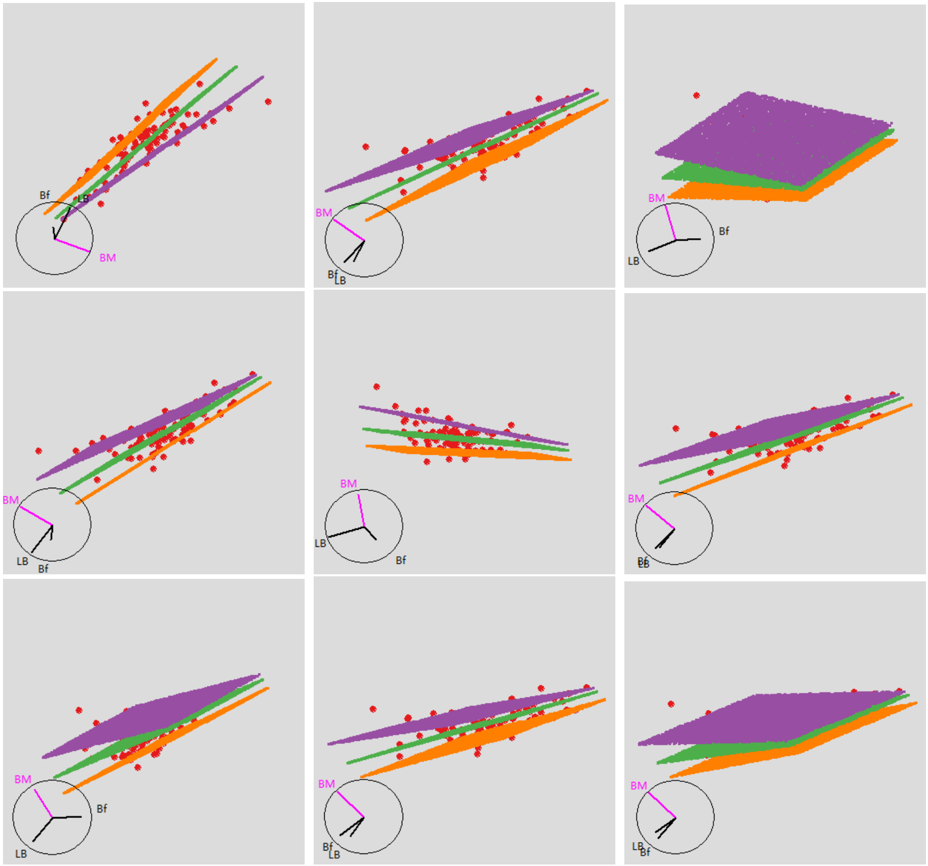

Figure 14: Linear quantile regression model with 2 response variables. Models on quantile 0.1, 0.5 and 0.9 corresponds to color orange, green and purple.

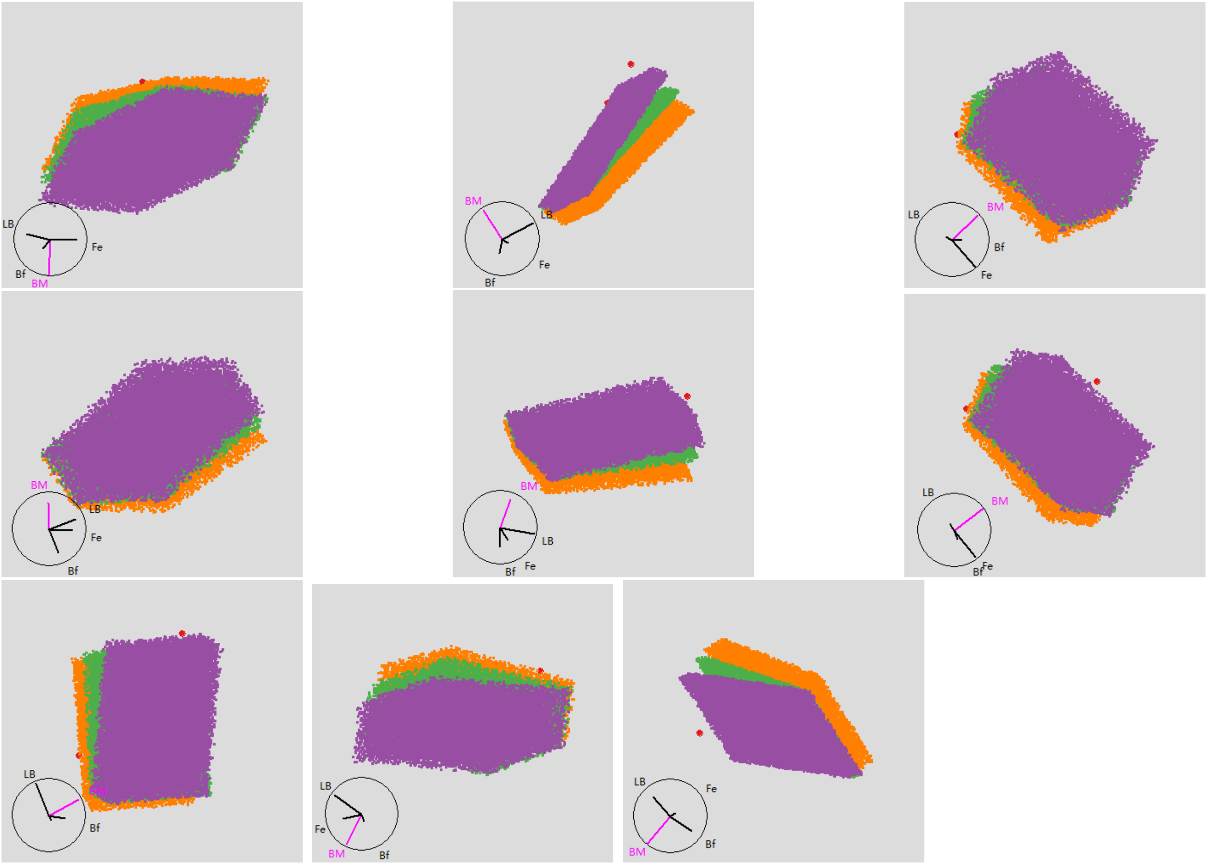

Figure 15: Linear quantile regression model with 3 response variables. Models on quantile 0.1, 0.5 and 0.9 corresponds to color orange, green and purple.

In three predictor case (4D data space), quantile regression models are cuboids in space which were displayed in Figure 14. We identified one point being the potential outliers and the relative location of the three regression models maintained in the data space.

6.2 Non-linear Case Result

In non-linear case, we use elliptic hyperboloid and hyperbolic paraboloid as examples. Figure 15 and 16 display interesting information: (a) non-linear models show different shape on different quantiles. For the elliptic hyperboloid, on high quantile, the non-linear model have largest curvature comparing to models on quantile 0.5 and 0.1, while model on quantile 0.1 has the smallest curvature. For the hyperbolic paraboloid, the curvature of models various among quantiles much larger. (b) The relative locations of models are maintained based on the quantile of data. (c) no clustered or sparsed region exists in data and no suspicious outlier exists.

7 Summary and Future Work

This paper presents the R package `quokar` for outlier diagnostic of quantile regression. The package contains methods for outlier detecting. We considered diagnostic methods corresponding to estimation with none error term distribution assumption, error term with asymmetric Laplace distribution assumption and Bayesian estimating framework. The results are provided in tidy

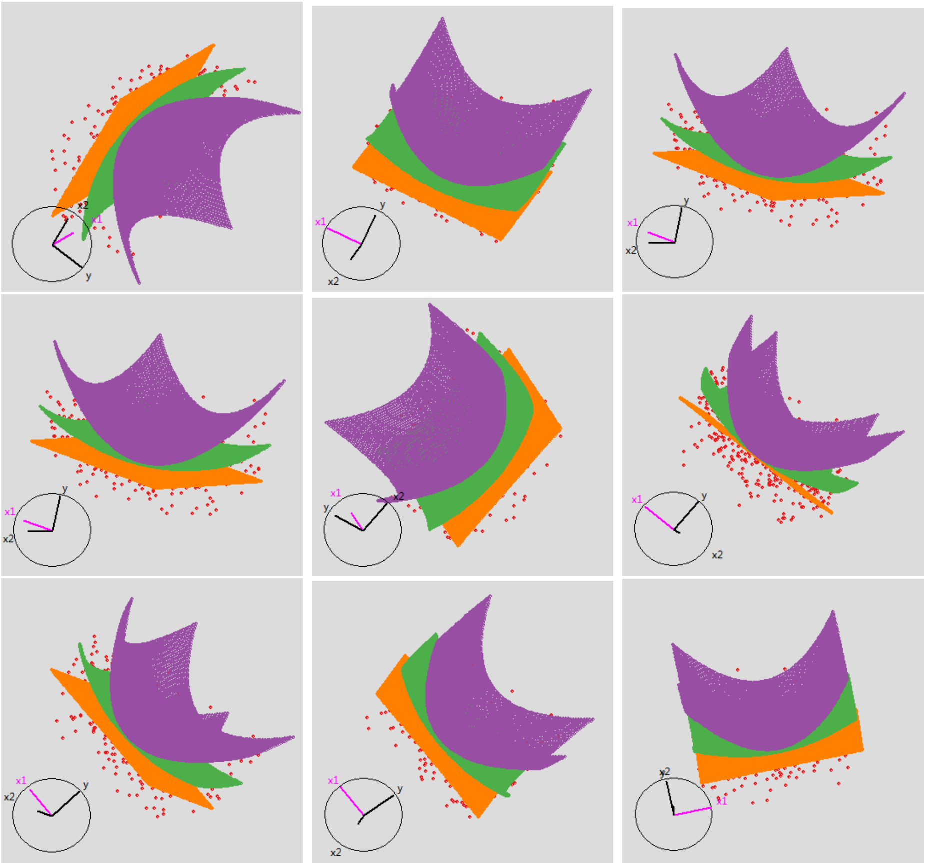

Figure 16: Non-linear quantile regression model on elliptic hyperboloid. Models on quantile 0.1, 0.5 and 0.9 corresponds to color orange, green and purple.

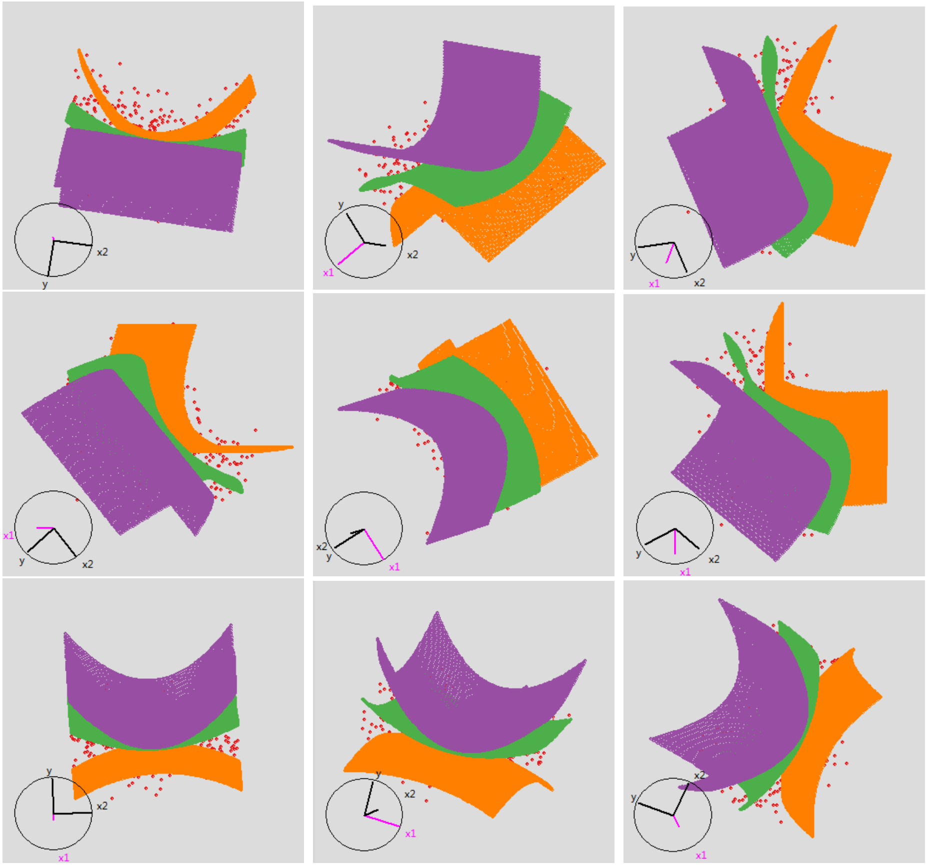

Figure 17: Non-linear quantile regression model on elliptic hyperboloid. Models on quantile 0.1, 0.5 and 0.9 corresponds to color orange, green and purple.

data form which can be directly used for plot. In addition, we provide visualization of the outlier state in original data and diagnostic results with outlier marked.

In data example, it was shown that the `quokar` package is a convenient way to detect suspicious outliers in quantile regression. Future versions of the package will focus on supporting other diagnotic methods such as methods for high dimensional data or extreme quantiles and improving computational efficiency.

Another contribution of this paper is proposed a general framework to visualize quantile regression in high dimensional data space. Our visualization tool is `GGobi`. We organized the problems can be answered by visualizing the integrated plot of quantile regression models and original data. Our future work will continue to explore visualization methods for the outlier diagnostic models. We are trying to do model performance comparison by visualizing using `GGobi`.

Reference

Koenker R, Machado J A F. Goodness of fit and related inference processes for quantile regression[J]. Journal of the american statistical association, 1999, 94(448): 1296-1310.

Fitzenberger B. The moving blocks bootstrap and robust inference for linear least squares and quantile regressions[J]. Journal of Econometrics, 1998, 82(2): 235-287.

Chernozhukov V, Hansen C. Instrumental variable quantile regression: A robust inference approach[J]. Journal of Econometrics, 2008, 142(1): 379-398.

Geraci M, Bottai M. Quantile regression for longitudinal data using the asymmetric Laplace distribution[J]. Biostatistics, 2007, 8(1): 140-154.

Koenker R. Quantile regression for longitudinal data[J]. Journal of Multivariate Analysis, 2004, 91(1): 74-89.

Korobilis D. Quantile regression forecasts of inflation under model uncertainty[J]. International Journal of Forecasting, 2017, 33(1): 11-20.

Autor D H, Houseman S N, Kerr S P. The Effect of Work First Job Placements on the Distribution of Earnings: An Instrumental Variable Quantile Regression Approach[J]. Journal of Labor Economics, 2017, 35(1): 149-190.

Mitchell J A, Dowda M, Pate R R, et al. Physical Activity and Pediatric Obesity: A Quantile Regression Analysis[J]. Medicine and science in sports and exercise, 2017, 49(3): 466.

Gallego-Álvarez I, Ortas E. Corporate environmental sustainability reporting in the context of national cultures: A quantile regression approach[J]. International Business Review, 2017, 26(2): 337-353.

Maciejowska K, Nowotarski J, Weron R. Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging[J]. International Journal of Forecasting, 2016, 32(3): 957-965.

Parente P M D C, Santos Silva J. Quantile regression with clustered data[J]. Journal of Econometric Methods, 2016, 5(1): 1-15.

Galvao A F, Kato K. Smoothed quantile regression for panel data[J]. Journal of Econometrics, 2016, 193(1): 92-112.

Arellano M, Bonhomme S. Nonlinear panel data estimation via quantile regressions[J]. The Econometrics Journal, 2016, 19(3).

Canay I A. A simple approach to quantile regression for panel data[J]. The Econometrics Journal, 2011, 14(3): 368-386.

Geraci M. Linear quantile mixed models: the lqmm package for Laplace quantile regression[J]. Journal of Statistical Software, 2014, 57(13): 1-29.

Chernozhukov V, Hansen C. Instrumental quantile regression inference for structural and treatment effect models[J]. Journal of Econometrics, 2006, 132(2): 491-525.