Diagnosing Outliers and Visualization of Quantile Regression Models

Wenjing Wang, Di Cook, Earo Wang

August 2017

Working Paper no/yr

Diagnosing Outliers and Visualization of Quantile Regression Models

Wenjing Wang
xxx,
Renmin University of China, xxx
China.
Email: wenjingwang1990@163.com
Corresponding author

Di Cook
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: dicook@monash.edu

Earo Wang
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: earo.wang@monash.edu

20 August 2017

JEL classification: C10,C14,C22

Diagnosing Outliers and Visualization of Quantile Regression Models

Abstract

Quantile regression models have been widely used in numerous applications, and the subject of substantial research. In a recent review article (Koenker (2017)) that surveys the research in the for the past 40 years, one recommendation is that more diagnostics should be provided for modeling. Estimation has been the priority, but diagnosing the models is lacking. This paper provides diagnostics for assessing the robustness of quantile regression models and implements outlier detection methods in the R (R Core Team (2017)) package `quokar`. The package contains functions and plots for detecting outliers in quantile regression.

Keywords: blah, blah

1 Introduction

Quantile regression model has been widely used in many research areas such as economy, finance and social science (see Autor, Houseman, and Kerr (2017), Mitchell et al. (2017), Gallego-Álvarez and Ortas (2017), Maciejowska, Nowotarski, and Weron (2016)). Quantile regression provides improvements on mean regression because, (a) observed covariates can describe the full distribution of a response variable, and (b) estimators can maintain optimal properties in the presence of heteroscedasticity or heavy tailed distributions.

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 to use l^1 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 (Galvao (2011), Canay (2011)) also construct quantile regression model using panel data. Panel data potentially

integrate 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 possesed, researches also interested in ebbeding it in other models to enhance model features or conduct better results analysis. Geraci and Bottai (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.

Along with the continous progresses in model improvement and application, extensive model inferencing research has been done. Gutenbrunner et al. (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 studied at the same time (Hahn (1995), Buchinsky (1995), Feng, He, and Hu (2011)). Koenker and Machado (1999) used Kolmogorov-Smironov method to measure the goodness of fit of quantile regression. To tackle the "Durbin problem", Koenker and Xiao (2002) developed location shift and location-scale shift test for quantile regression model. The inferencing study also have been extended to Bayesian framework (Yu and Moyeed (2001), Yu and Stander (2007), Kozumi and Kobayashi (2011), Santos and Bolfarine (2016)).

Effective toolboxes conduct fitting and inferencing for quantile regression model has been developed based on the theoretical researches mentioned above. 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 diagnostic methods were proposed for quantile regression and no toolbox for model diagnostic were implemented in R. Gu and Koenker (2017) pointed out more work needs to be done to develop better diagnostic tools.

Outlier detection is one important aspect of model diagnostic. It is unignorable in regression analysis because the results of regression can be sensitive to outliers. Data used for fitting regression model may contain special points located far away from others either in response variable or in the space of the predictors. The latter are also called leverage points. In univariate

model, outliers can be easily observed by scatter plot of predictor and responsor. Difficulty lies in high-dimensional situation, where statistical methods should be used. Various methods for detecting outliers have been explored (Rousseeuw and Van Zomeren (1990), Gather and Becker (1997)). Commonly used statistics include residuals, leverage value, studentized residuals and jacknife residuals.

In regression context, classic least ordinary square estimation of linear regresssion can be expressed as $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$. where, $\boldsymbol{H}$ is called hat matrix. Residuals can be written as $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{1} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{1} - \boldsymbol{H})\boldsymbol{\epsilon}$. Hence, if consider 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 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.

Estimating process of quantile regression differs from mean regression which results no explicit solution form $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. In addition, quantile regressions can be fitted on any quantile interested , which add difficulties in applying diagnosing methods and displaying results simutaneously. Sánchez, Lachos, and Labra (2013) 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. Toolbox for outlier diagnostic can only be found in statistical software SAS in procedure `QUANTREG`. To the authors' knowledge, these methods are still not be implemented in R. In order to fill the gap, related 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.

Visualizations of high-dimensional data are well developed and explored. Model visualizations are recently been studied. Wickham, Cook, and Hofmann (2015) first proposed the idea of plotting model in data space. He pointed out that visual model can anwser questions such as what does the model look like? how well does the model fit the data? how does the shape of the model compare to the shape of the data? is the model fitting uniformly good, or good in some regions but poor in other regions. Integrating model and original data in one plot is a straight-forward way to observe outliers for fitted model. To visualize data and models in high dimensions, we need good tools:

- The grand tour. The grand tour generating a sequence of 2d projections of an p-d object. It randomly chooses new projections and rotates between them. Through this process, we can observe the distribution of data (clusted or sparse) and identify outliers.
- Brushing. We use a brush to colour observations and models in one plot. In our case, regression models on different quantiles are brushed.

In this paper, we propose a general framework of quantile regression visualization with `GGobi`. Linear quantile regression model in 3D and 4D data space and non-linear models are displayed as examples.

This article aims to introduce R package `quokar` and propose general framework of visualizing quantile regression models in high dimensional space. The remainder of this article is organized as follows: In Section 2, we provide a general introduction to quantile regression model and explore its robusness. In Section 3 we give a tour of outlier diagnostic methods used in package `quokar`. In section 4 we illustrate the usefulness of the package through a real data set example. In section 5, we provide the geralize framework of visualizing quantile regression. The summary and future research directions are discussed in Section 6.

2 Robustness of Quantile Regression

Koenker and Bassett Jr (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 minimizing a sum of asymmetrically weighted absolute residuals, the objective function of quantile regression 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 $\hat{\boldsymbol{\beta}}_\tau$ is asymptotically normal, which can be expressed as,

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

where $\boldsymbol{D} = E(f(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}\boldsymbol{X}^{'})$ and $\boldsymbol{\Omega}_x = E(\boldsymbol{X}^{'}\boldsymbol{X})$.

Quantile is more robust than mean when extreme values exist in the dataset interested. This property applies equally in regression context. The robustness of quantile and quantile regression can be expressed by influence function. F is distribution function of interested variable. Set T as a function 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 (4) and (5), the latter is a bounded function with y changing in its domain. Figure 1 is an example plot of influence funtion. Data are generated from distribution function $F(x) = 1 - e^{-\lambda x}, x > 0$, and the corresponding density function and inverse distribution function are $f(x) = e^{-x}$, $Q(\tau) = -ln(1-\tau)$ respectively.

For quantile regression, suppose F represent the joint distribution of the pairs (x, y), the influence function is Equation (6). 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.

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

where

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

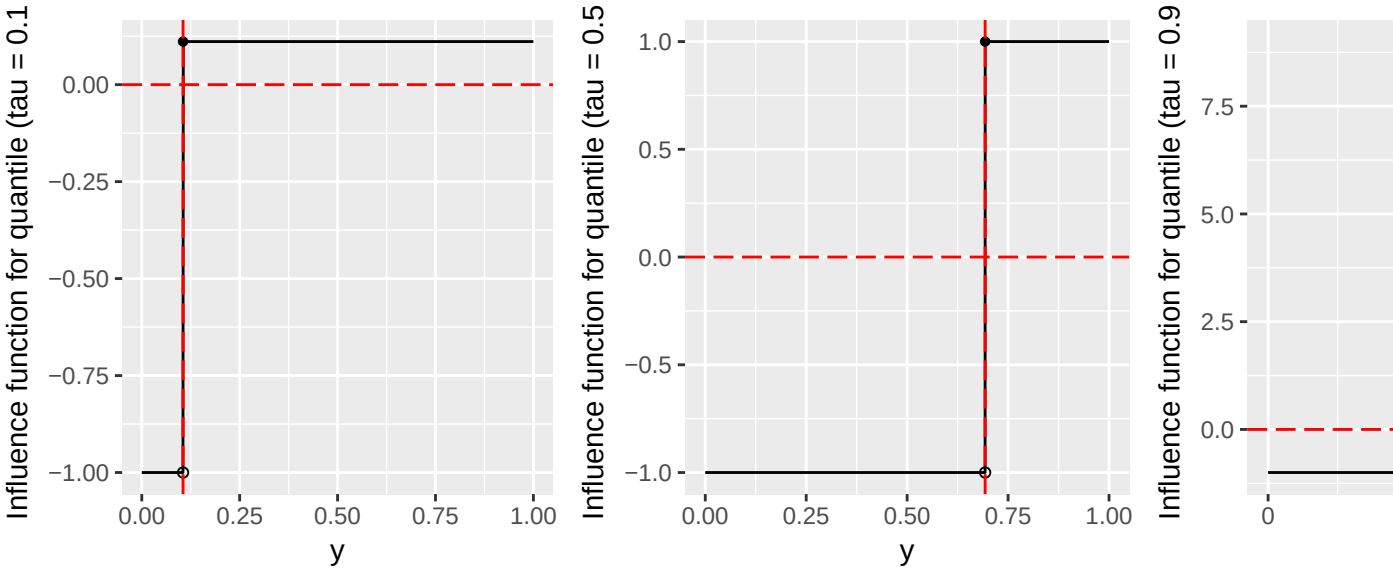

Figure 1: *Visualization of influence function for Mean and Quantile. It is obviously that quantile influence functions on quantile 0.1, 0.5 and 0.9 are bounded which indicat that quantile is more robust then Mean. The boundaries of influence function on low and high quantile are asymmetrical.*

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

To illustrate the model robustness indicated by influence function, we conduct the following simulation study.

- Fitting model with data condaminated by outliers

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.

- Moving outliers location in Y-aixs and X-aixs

To visualize the robustness of quantile regression, we simulate 100 data with 5 condaminated points considered as outliers. We conduct two experiments to test the boundness of quantile regression towards outliers. The first experiment is moving outliers in Y-aixs to observe the change of regression estimated coefficients, and the other is moving them in X-aix. In the former experiment, Figure 4 show that when outliers moving down in Y-aixs for 10 unit, they pull down the slope on every quantile (Comparing the result of $y_1 = x + \epsilon$ and $y_2 = x + \epsilon$). However, figure 4 also show that keeping moving down the outliers does no change to the regression

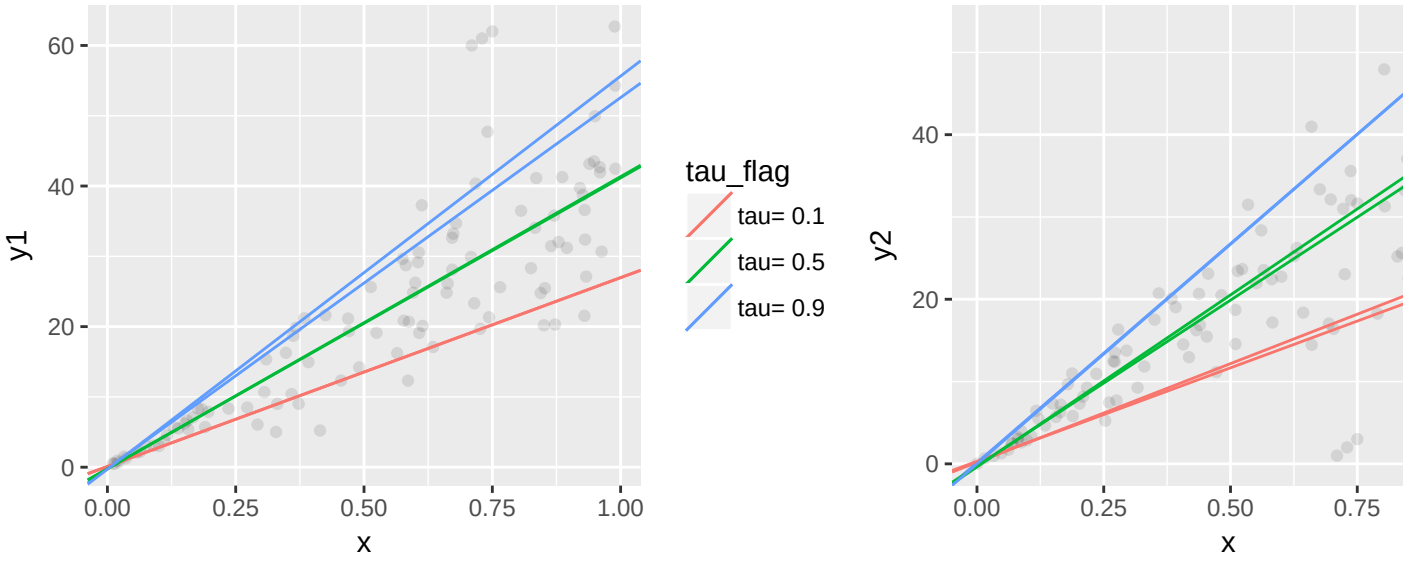

Figure 2: *Fitting quantile regression model on quantile 0.1, 0.5 and 0.9 using simulated datasets with and without outliers. The outliers located at the top-left of the original dataset. Results show that outliers pull up the slope of the 0.9 and 0.1 regression line. When outliers located at the bottom-right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

slopes. This reflect the boundness of influence function and the robustness of quantile regression. To observed the change of coefficients in multi-variable model, we fit quantile regression model $y = x_1 + x_2 + \epsilon$. Figure 5 show that coefficients changes slowly when moving down the outliers in Y-axis. In the other experiment, we follow the similar procedure above while moving outliers in X-axis. Figure 6 and Figure 7 show estimated coefficients change every time outliers move in X-aixs, and the isolate outliers are, the greater of the change. Besides every move has different effect on each quantiles.

In conclusion, quantile regression response differently to outliers comparing mean regression in three 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) the effect of outliers in Y-aixs on quantile regression has boundary; (c) quantile regression have weak robustness to leverage points.

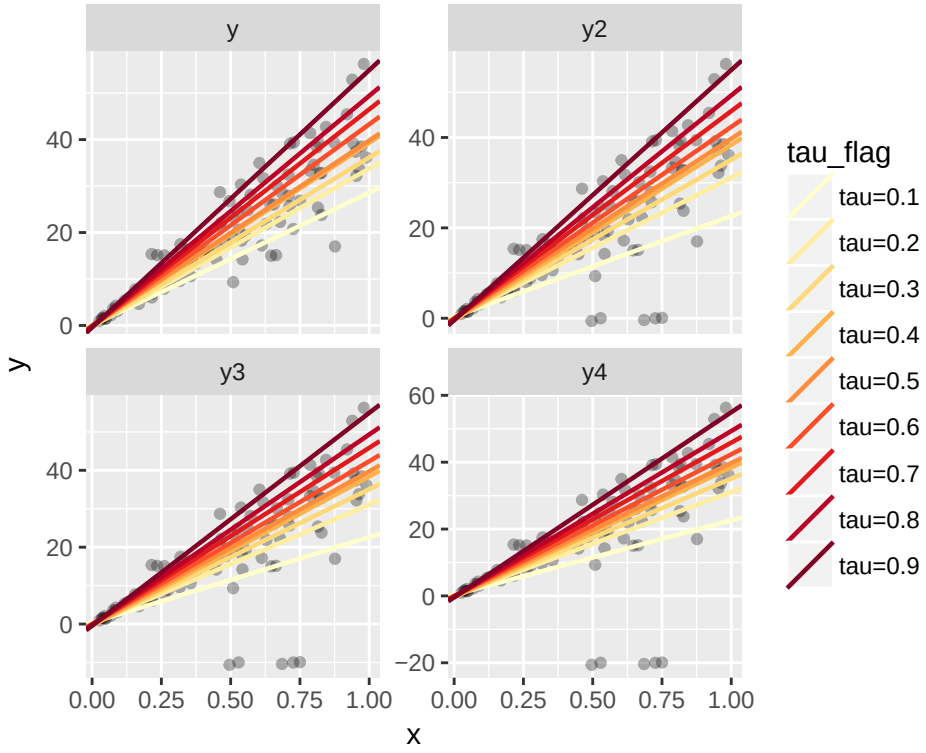

Figure 3: *Fit quantile regression models using simulated data. Keep moving down the outliers in Y-axis to get different longitudinal ordinates values:* $y_2 = y - 5, y_3 = y - 10$ *and* $y_4 = y - 15$*. We are interested in the change of regression lines.*

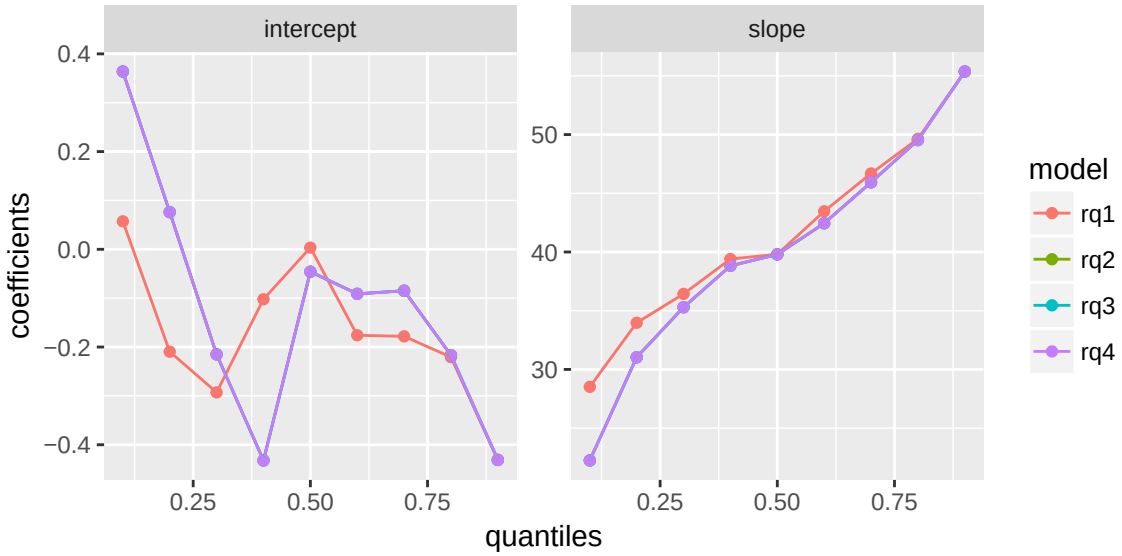

Figure 4: *Fit quantile regression models using simulated data. Keep moving down the outliers in Y-axis to get different longitudinal coordinates values:* $y_2 = y - 5, y_3 = y - 10$ *and* $y_4 = y - 15$*. Calculating the estimated coefficients in each experiment and results show that in single predictor case, outliers moving down in y make no difference to the quantile regression coefficients estimations. This visualization show the bound property of influence function for quantile regression.*

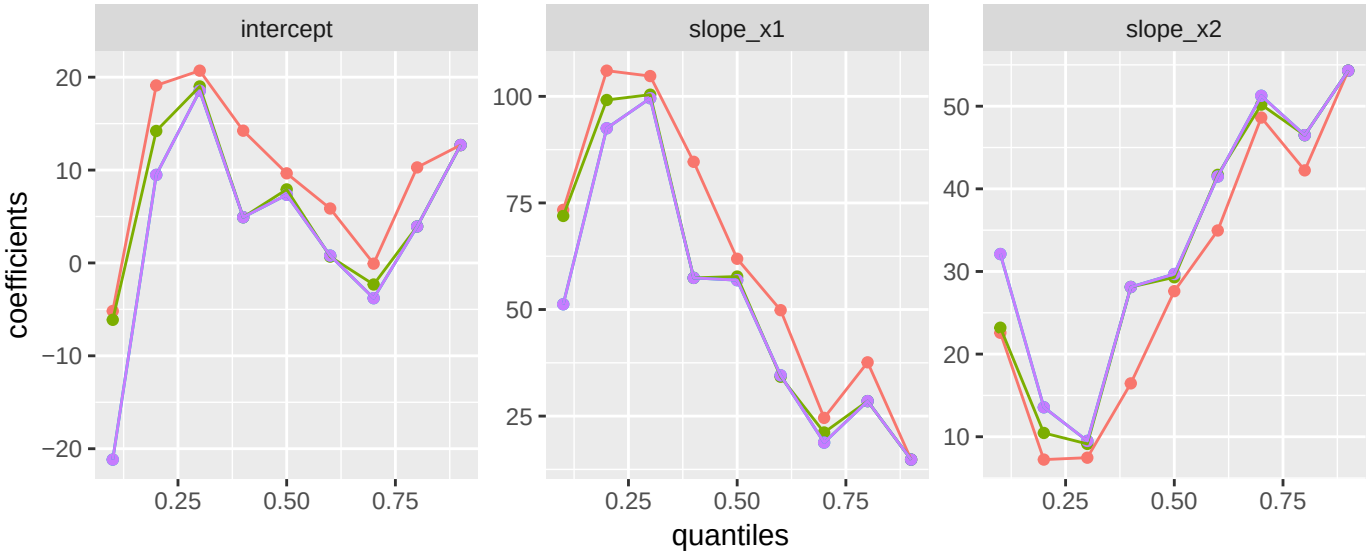

Figure 5: *Fit quantile regression models using simulated data. Keep moving down the outliers in Y-axis to get different longitudinal coordinates values:* $y_2 = y - 5, y_3 = y - 10$ *and* $y_4 = y - 15$. *Results show that in multi predictors case, outliers moving down in Y-axis still makes little change to the quantile regression coefficients estimations.*

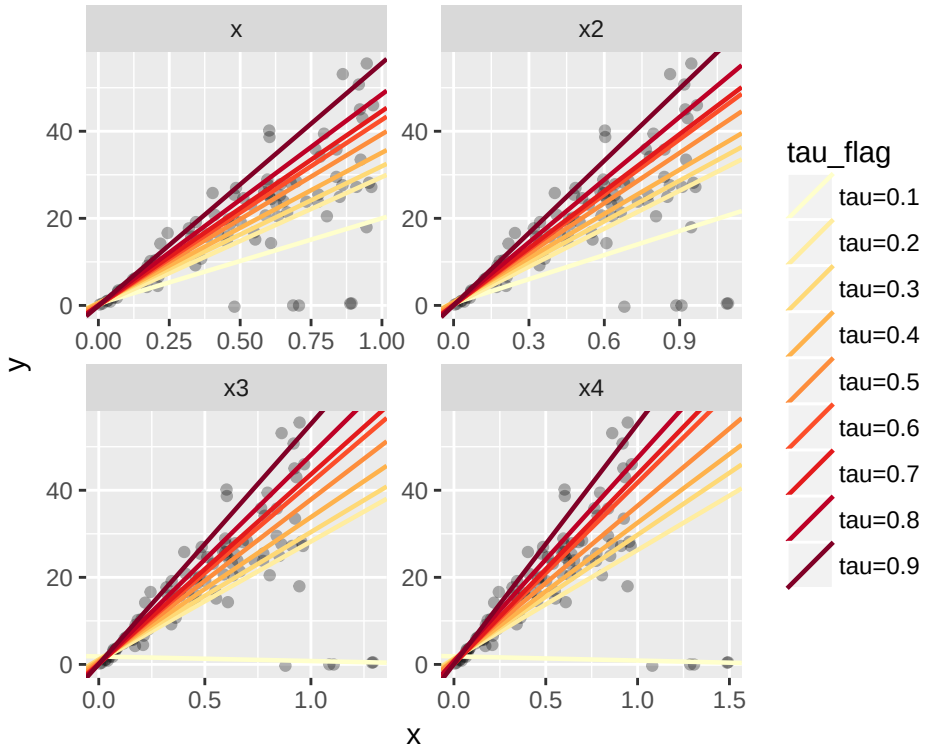

Figure 6: *Fit quantile regression models using simulated data. Keep moving the outliers to the right in X-aixs to get different horizontal ordinate values:* $x_2 = x + 0.2, x_3 = x + 0.4$ *and* $x_4 = x + 0.6$. *We are interested in the change of regression lines.*

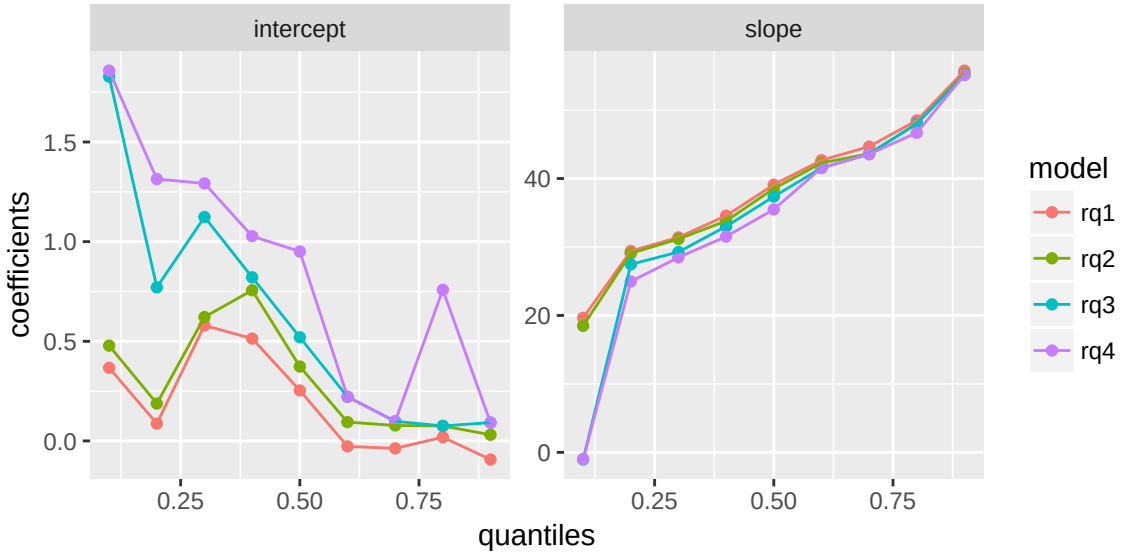

Figure 7: *Fit quantile regression models using simulated data. Keep moving the outliers to the right in X-aixs to get different horizontal ordinate values:* $x_2 = x + 0.2$, $x_3 = x + 0.4$ *and* $x_4 = x + 0.6$. *Calculating the estimated coefficients in each experiment and results show that outliers moving in X-aixs make larger difference to the quantile regression coefficients then moving in Y-aixs.*

3 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 (Sánchez, Lachos, and Labra (2013), Santos and Bolfarine (2016)) and performed well in our application. We assume a basic knowledge of quantile regerssion and Baysian methods.

3.1 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:

$$MD(x_i) = [(x_i - \bar{x})' \bar{\boldsymbol{C}}(\boldsymbol{X})^{-1}(x_i - \bar{x})]^{1/2} \tag{9}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{\boldsymbol{C}}(\boldsymbol{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})'(x_i - \bar{x})$ 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 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 Zomeren (1991)). In quokar, we use minimum covariance determinant (MCD) proposed by Rousseeuw and Driessen (1999) to estimate the above two parameters. The MCD estimator can be defined as

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

where $\bar{X}$ and S represent location and scale. $h = p : |S_h^*| < |S_k^*|, |k| = p$, $\bar{X}_h^* = \frac{1}{p}\sum_{i\in p} x_i$, $S_p^* = \frac{1}{p}\sum_{i\in p}(x_i - \bar{X}_p^*)(x_i - \bar{X}_p^*)'$. p can be thought 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. With MCD, we can calculate robust distance which was defined as

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

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

Package `quokar` implement 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.

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

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

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 θ based on $L(\theta|Y)$ and $L(\theta|Y_{[i]})$ respectively. Cook's distance CD_i is given by (14). For external norms, M is usually chosen to be $-\ddot{L}(Y|\theta)$.

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

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

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

The ith observation is regarded as influential if the value of Cook's distance or likelihood distance is relatively large. Benites, Lachos, and Vilca (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 (Equation q_function) for estimation.

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

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

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

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

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

Hence, the Cook's distance is defined as

$$CD_i = (\hat{\theta}_{[i]} - \hat{\theta})' \{-Q(\hat{\theta}|\hat{\theta})\}(\hat{\theta}_{[i]} - \hat{\theta}), \quad i = 1, ..., n \tag{19}$$

The measurement of the influence of the ith case which based directly on the Q function is similar to the likelihood distance LD_i. It can be defined as,

$$QD_i = 2\{Q(\hat{\theta}|\hat{\theta}) - Q(\hat{\theta}_{[i]}|\hat{\theta})\} \quad i = 1, ..., n \tag{20}$$

3.3 Mean Posterior Probability and Kullback-Leibler Divergence

In Bayesian quantile regression framework, Kullback and Leibler (1951) 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 (22) 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.

$$O_i = \begin{cases} 1, & i \quad is \quad outlier \\ 0, & i \quad is \quad normal \end{cases} \tag{23}$$

The mean posterior probability appoximatlly calculated by MCMC draw is

$$P(O_i = 1) = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{M} I(v_i^{(l)} > \max_{k \in 1:M} v_j^{(k)}) \tag{24}$$

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 in Equation @ref(eq:kl_divergence) and @ref(eq:mean_posterior_kl). 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.

$$K(f_i, f_j) = \int log(\frac{f_i(x)}{f_j(x)}) f_i(x) dx \tag{25}$$

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

$$KL(f_i) = \frac{1}{n-1} \sum_{j \neq i} K(f_i, f_j) \tag{26}$$

4 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:

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

We implete Australia Institution of Sports (AIS) data as an example to introduce this package. AIS is a data frame with 202 observations (102 male and 100 female) on 13 variables. The female data contain outlier which suit for our research.

4.1 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 **??** 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 9 showed case 56 and 75 are suspicious outliers in the data set.

```
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, ],
                      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 = "red",hjust = 0, vjust = 0) +
  scale_colour_brewer(palette="YlOrRd")+
  theme_grey()
```

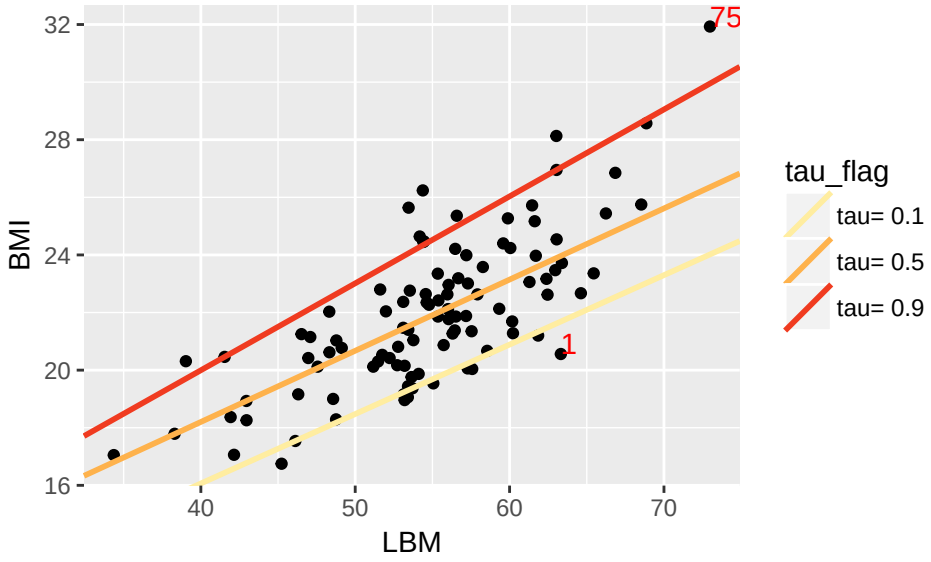

Figure 8: *Plot the outlier state for single variable case.*

```
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 = "red", vjust = 0, hjust = 0) +
  facet_wrap(~variable, scales = "free_x") +
  scale_colour_brewer(palette="YlOrRd")+
  theme_grey()
```

4.2 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

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.

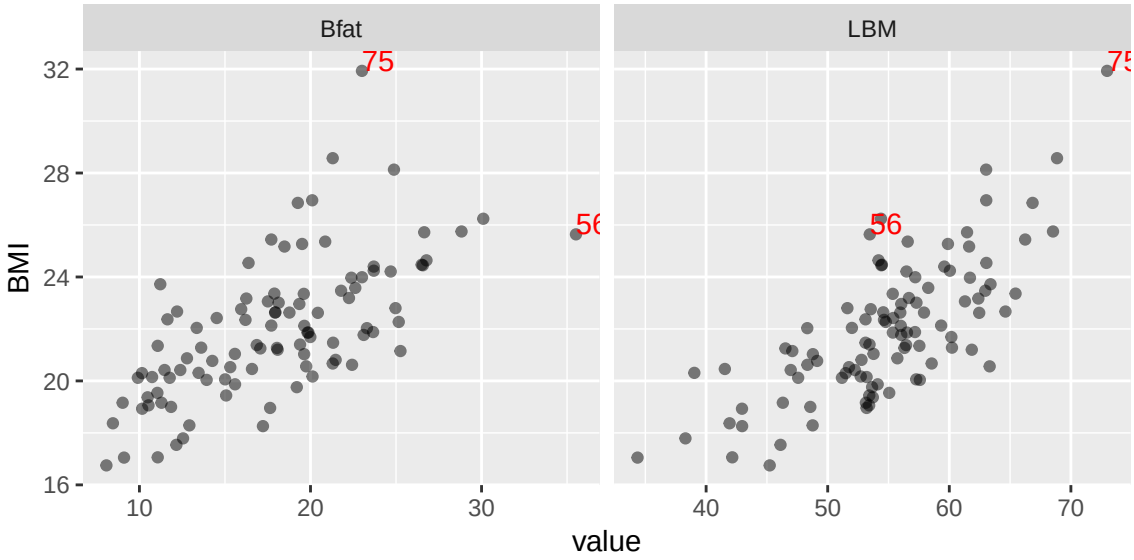

Figure 9: *Plot the outlier state for multi-variable regression.*

```
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) %>%
                    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))
```

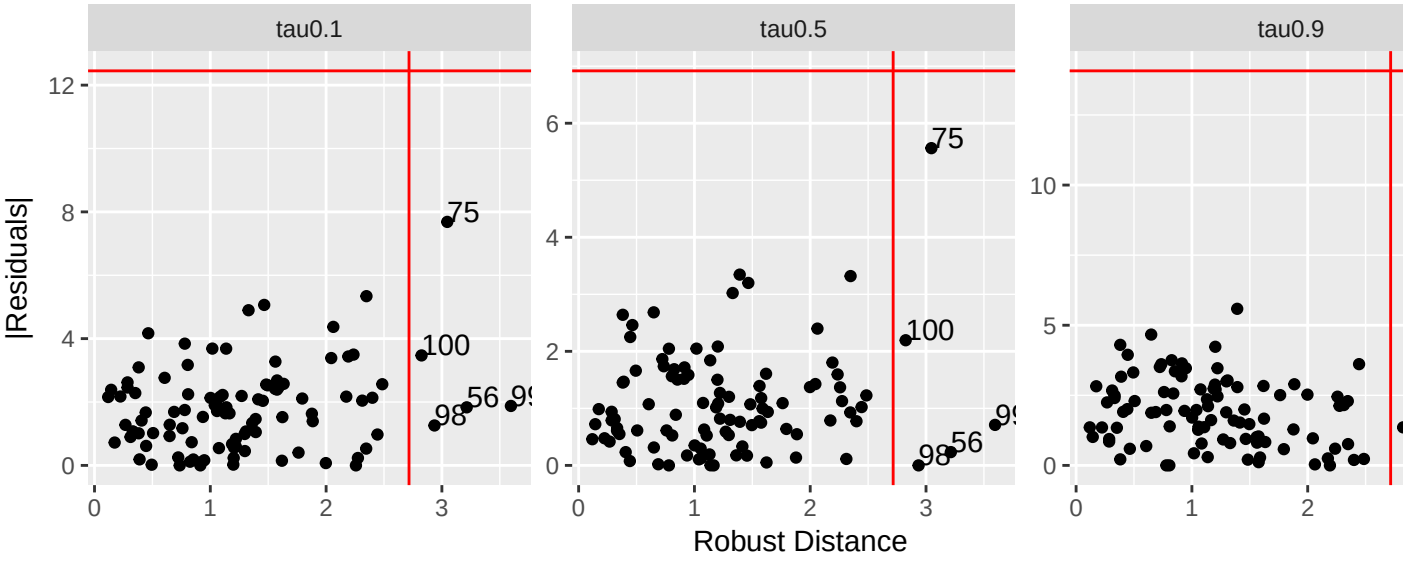

Figure 10: *Robust Distance-Residual Plot. Points on the right of vertical cutoff line are considered leverage points and points above the horizontal cutoff line are outliers in y-direction.*

```
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') +
    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) +
```

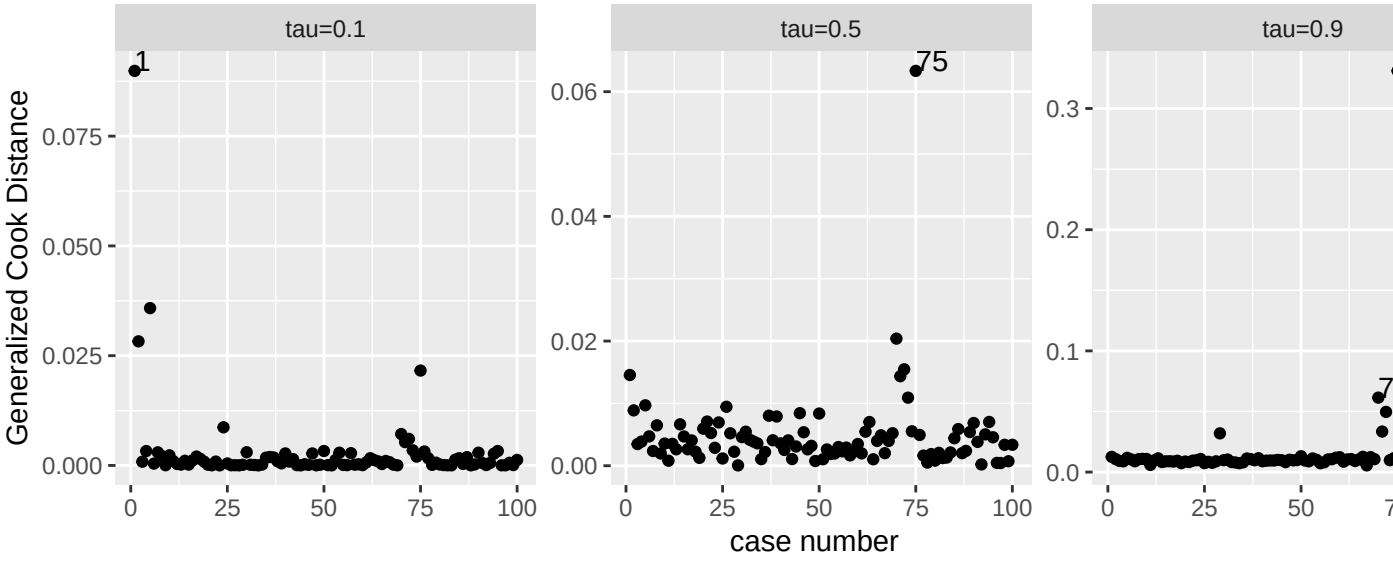

Figure 11: *Generalized Cook distance of each observation on quantile 0.1, 0.5 and 0.9. Case 75 has relative large Cook distance-funtion distance to other points*

```
xlab('case number') +
ylab('Qfunction Distance')
```

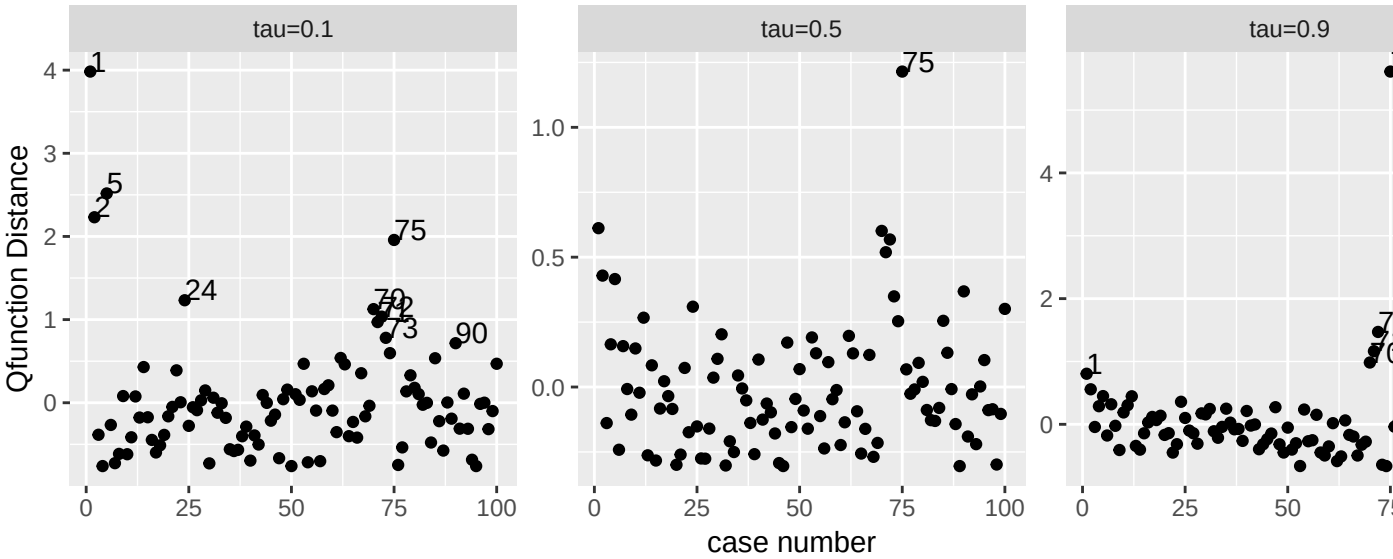

Figure 12: *Q function distance of each observation on quantile 0.1, 0.5 and 0.9. Case 75 has relative large Q function distance to other points*

```
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 =  5000, burn = 1000,
```

```
                  method = 'bayes.prob')


kl <- frame_bayes(y, x, tau, M = 5000, burn = 1000,
                   method = 'bayes.kl')
```

```
head(prob)
head(kl)
```

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')
```

5 Generalized Framework for Visualizing Quantile 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 et al. (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 $\boldsymbol{x}_i \in \boldsymbol{X}$. In high dimension case, $\boldsymbol{x}_i$ is a vector and $\boldsymbol{X}$ is matrix. The ith value of responsor $\boldsymbol{Y}$ is y_i. The quantile regression model $f_\tau : \boldsymbol{X} \to \boldsymbol{Y}$ will be fitted on the given data set.

- Use grid method to generate data in data space bounded by $\boldsymbol{X}$. The generated data form data set $\boldsymbol{Z}$.
- Fit quantile regression models f_τ on every interested quantile and get the estimated parameters $\mathbf{fi}_\tau$.
- Calculate quantile regression model using $f_\tau = \mathbf{Zfi}_\tau$. In non-linear case, calculate model based on the non-linear curve form.

- Tidy data set $(\boldsymbol{Z}, f_\tau)$ for each quantile into long data form and add tag representing quantile.

5.1 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`.

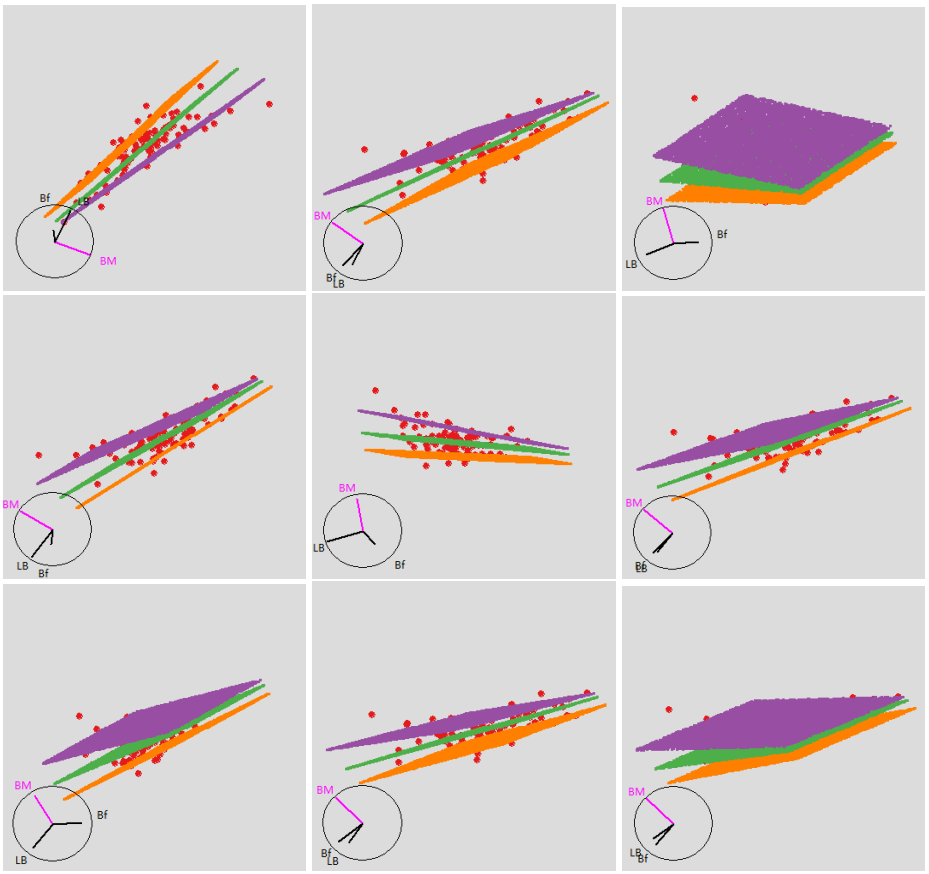

Figure 13: *Linear quantile regression model with 2 response variables. Models on quantile 0.1, 0.5 and 0.9 corresponds to color orange, green and purple.*

Figure 13 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.

In three predictor case (4D data space), quantile regression models are cuboids 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.

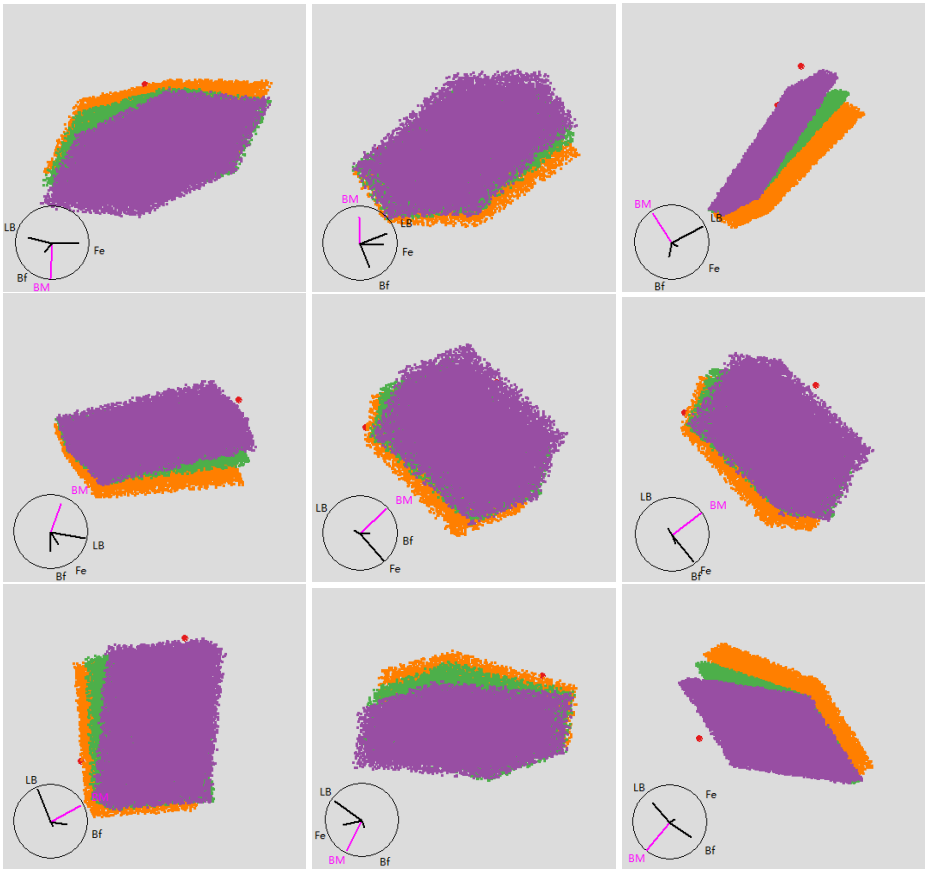

Figure 14: *Linear quantile regression model with 3 response variables. Models on quantile 0.1, 0.5 and 0.9 corresponds to color orange, green and purple.*

5.2 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.

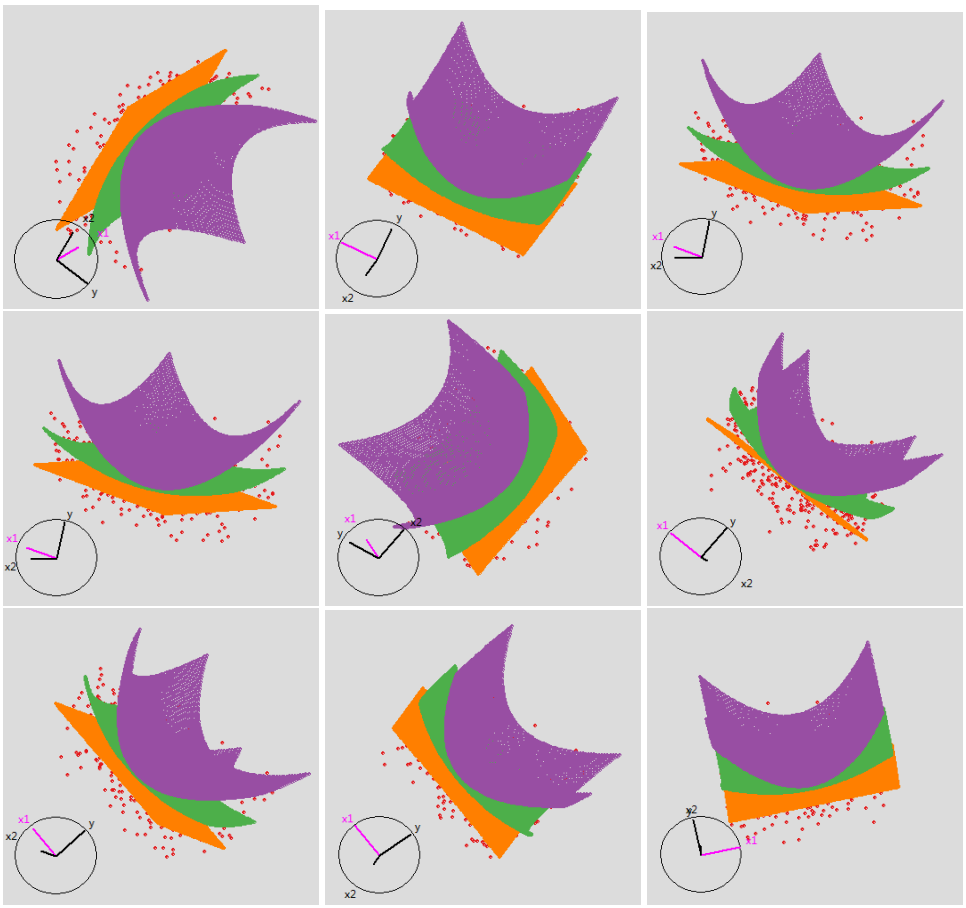

Figure 15: *Non-linear quantile regression model on elliptic hyperboloid. Models on quantile 0.1, 0.5 and 0.9 corresponds to color orange, green and purple.*

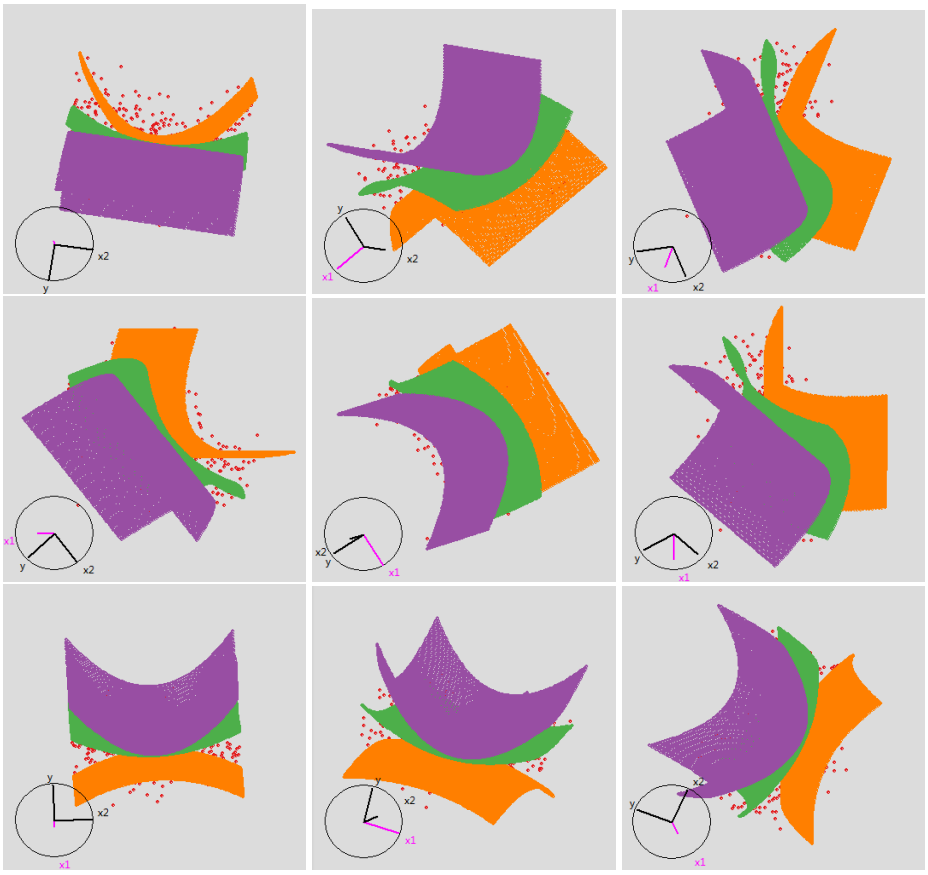

Figure 16: *Non-linear quantile regression model on elliptic hyperboloid. Models on quantile 0.1, 0.5 and 0.9 corresponds to color orange, green and purple.*

6 Summary and Future Work

The first part of this paper presents 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 data form which can be directly used for plot. In addition, we provide some visualization methods for the diagnosed results.

In data example, it was shown that `quokar` provides convenient tools 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 provide integrated plot of quantile regression models and original data. Our future work will continue to explore visual methods for the outlier diagnostic models in data space. We are trying to do model performance comparison with `GGobi`.

References

Autor, DH, SN Houseman, and SP Kerr (2017). The Effect of Work First Job Placements on the Distribution of Earnings: An Instrumental Variable Quantile Regression Approach. *Journal of Labor Economics* **35**(1), 149–190.

Benites, LE, VH Lachos, and FE Vilca (2015). Case-Deletion Diagnostics for Quantile Regression Using the Asymmetric Laplace Distribution. *arXiv preprint arXiv:1509.05099*.

Buchinsky, M (1995). Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study. *Journal of Econometrics* **68**(2), 303–338.

Canay, IA (2011). A simple approach to quantile regression for panel data. *The Econometrics Journal* **14**(3), 368–386.

Chernozhukov, V and C Hansen (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics* **132**(2), 491–525.

Feng, X, X He, and J Hu (2011). Wild bootstrap for quantile regression. *Biometrika* **98**(4), 995.

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

Galvao, AF (2011). Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics* **164**(1), 142–157.

Gather, U and C Becker (1997). *Convergence rates in multivariate robust outlier identification*. Tech. rep. Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Geraci, M and M Bottai (2006). Use of auxiliary data in semi-parametric spatial regression with nonignorable missing responses. *Statistical Modelling* **6**(4), 321–336.

Geraci, M and M Bottai (2014). Linear quantile mixed models. *Statistics and computing* **24**(3), 461–479.

Gu, J and R Koenker (2017). Unobserved heterogeneity in income dynamics: an empirical Bayes perspective. *Journal of Business & Economic Statistics* **35**(1), 1–16.

Gutenbrunner, C, J Jurečková, R Koenker, and S Portnoy (1993). Tests of linear hypotheses based on regression rank scores. *Journaltitle of Nonparametric Statistics* **2**(4), 307–331.

Hahn, J (1995). Bootstrapping quantile regression estimators. *Econometric Theory* **11**(01), 105–121.

Koenker, R (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91**(1), 74–89.

Koenker, R (2017). Quantile Regression: 40 Years On. *Annual Review of Economics* **9**(1), null.

Koenker, R and G Bassett Jr (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R and JA Machado (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association* **94**(448), 1296–1310.

Koenker, R and Z Xiao (2002). Inference on the quantile regression process. *Econometrica* **70**(4), 1583–1612.

Kozumi, H and G Kobayashi (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of statistical computation and simulation* **81**(11), 1565–1578.

Kullback, S and RA Leibler (1951). On information and sufficiency. *The annals of mathematical statistics* **22**(1), 79–86.

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

Mitchell, JA, M Dowda, RR Pate, K Kordas, K Froberg, LB Sardinha, E Kolle, and A Page (2017). Physical Activity and Pediatric Obesity: A Quantile Regression Analysis. *Medicine and science in sports and exercise* **49**(3), 466.

Parente, PM and J Santos Silva (2016). Quantile regression with clustered data. *Journal of Econometric Methods* **5**(1), 1–15.

PEARSON, ES and CC Sekar (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika* **28**(3-4), 308–320.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Rousseeuw, PJ and KV Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**(3), 212–223.

Rousseeuw, PJ and BC Van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association* **85**(411), 633–639.

Rousseeuw, PJ and BC van Zomeren (1991). "Robust distances: simulations and cutoff values". In: *Directions in Robust Statistics and Diagnostics*. Springer, pp.195–203.

Sánchez, B, H Lachos, and V Labra (2013). Likelihood based inference for quantile regression using the asymmetric Laplace distribution. *Journal of Statistical Computation and Simulation* **81**, 1565–1578.

Santos, B and H Bolfarine (2016). On Bayesian quantile regression and outliers. *arXiv preprint arXiv:1601.07344*.

Swayne, DF, DT Lang, A Buja, and D Cook (2003). GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis* **43**(4), 423–444.

Wickham, H, D Cook, and H Hofmann (2015). Visualizing statistical models: removing the blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **8**(4), 203–225.

Yu, K and RA Moyeed (2001). Bayesian quantile regression. *Statistics & Probability Letters* **54**(4), 437–447.

Yu, K and J Stander (2007). Bayesian analysis of a Tobit quantile regression model. *Journal of Econometrics* **137**(1), 260–276.