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

2 Introduction

Quantile regression model has been widely used in many research areas such as economy, finance and social science. (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 $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 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 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

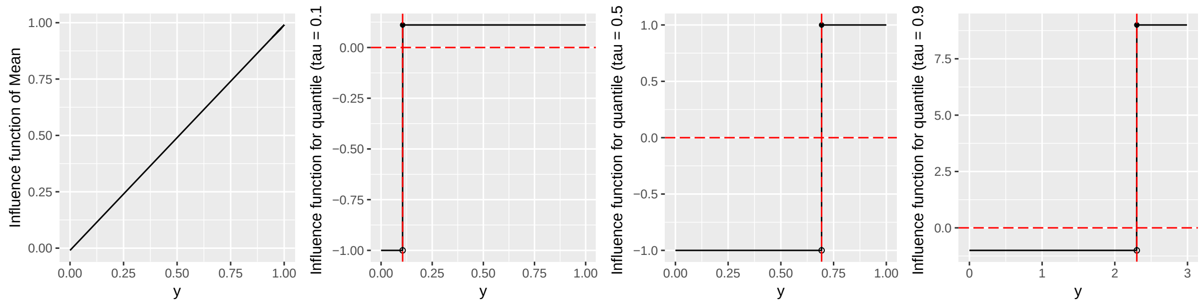

Figure 1: Visualization of influence function for Mean and Quantile. It is obviously that quantile influence functions on quantile 0.1, 0.5 and 0.9 are bounded which indicat that quantile is more robust then Mean. The boundaries of influence function on low and high quantile are asymmetrical.

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

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.

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 also conducted experiments to visualize the robustness of quantile regression. In two simulation studies, we generate 100 sample observation and 3 outliers. The outliers are distributed in two locations in each case. We fitted quantile regression based on these dataset to observe how do outliers affect model coefficients.

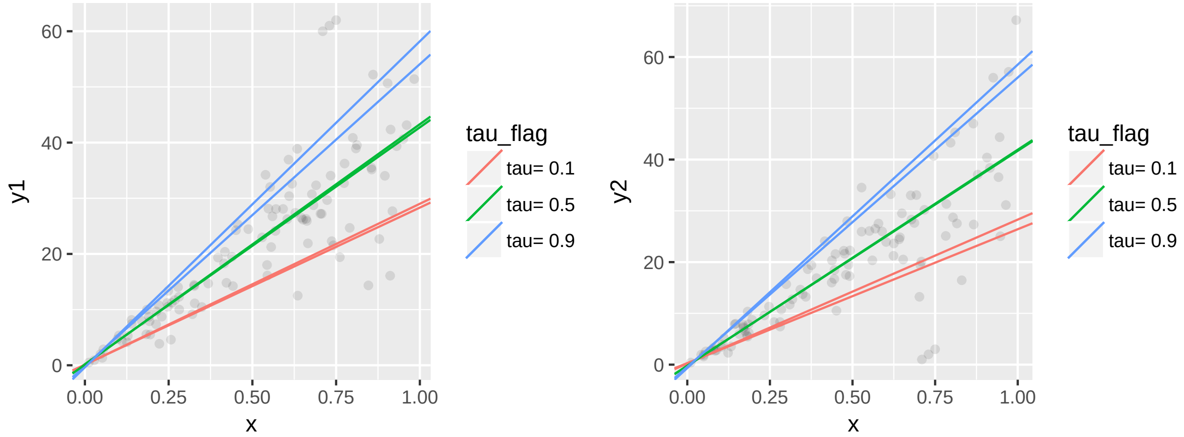

Figure 2: Fitting quantile regression model on quantile 0.1, 0.5 and 0.9 using simulated datasets with and without outliers. The outliers located at the top-left of the original dataset. Results show that outliers pull up the slope of the 0.9 and 0.1 regression line. When outliers located at the bottom-right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

We also conduct simulations to observe the robustness of quantile regression. These simulation studies are extended to multi-variable model. We generate 100 data which contains 5 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.

We also observed the change of coefficients in multi-variable model. The results show that coefficients changes slowly when keep moving down the outliers in y-direction.

If moving outliers in same pattern moving on x direction, slopes change every time outlier moves. To go further, each move does 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

Based on related researches, we implete several methods for outlier detecting in `quokar`.

4.0.1 1. Standard residual-Robust Distance

We can not use the famous “Hat Matrix” to detect leverage points in quantile regression 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:

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

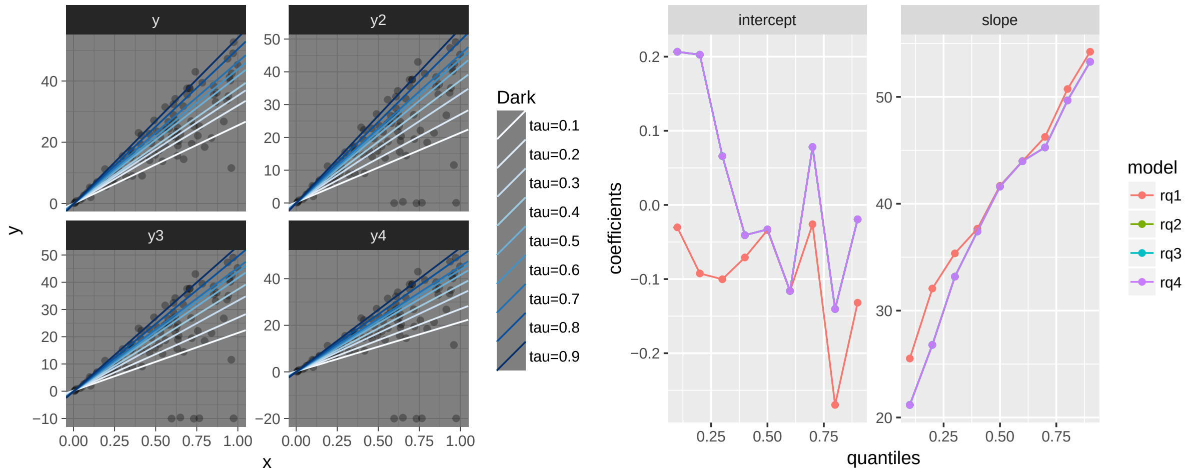

Figure 3: Left fig: Fitting quantile regression models using simulated data. We keep moving down the outliers in y direction in y2 (y-5), y3 (y-10) and y4 (y-15). Right fig: Fitting quantile regression models using simulated data. We keep moving down the outliers in y direction getting datasets with variable y2 (=y-5), y3 (=y-10) and y4 (=y-15). Results show that in single predictor case, outliers moving down in y make no difference to the quantile regression coefficients estimations

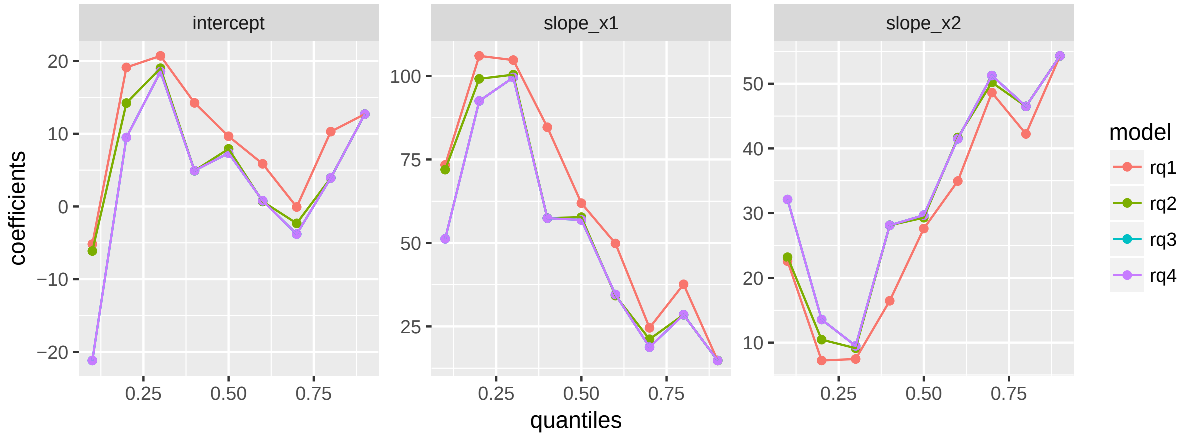

Figure 4: Fitting quantile regression models using simulated data. We keep moving down the outliers in y direction getting three datasets with different locations of outliers (changing in y-aixs, y2 (=y-5), y3 (=y-10) and y4 (=y-15)). Results show that in multi predictors case, outliers moving down in y make small change to the quantile regression coefficients estimations

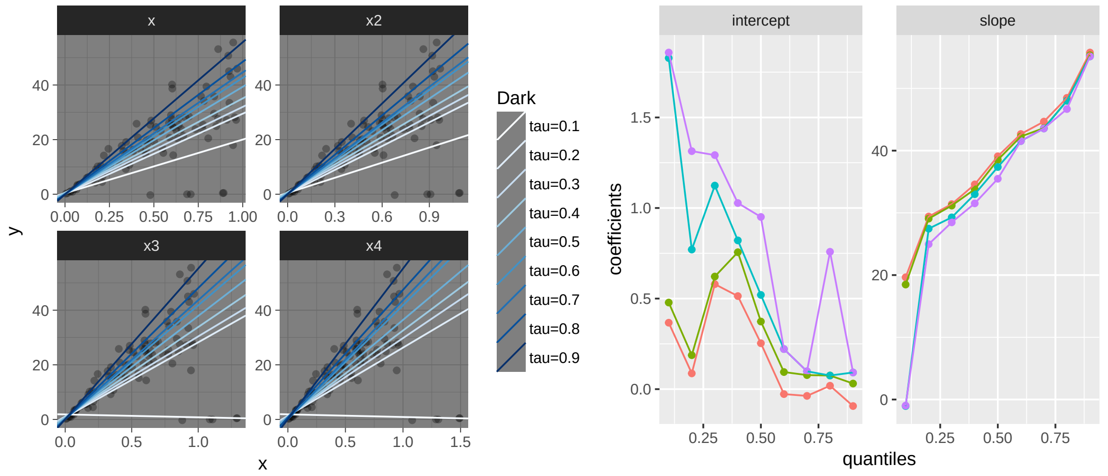

Figure 5: Left fig: Fitting quantile regression models using simulated data. We keep moving the outliers to the right in x direction getting three datasets with different locations of outliers (changing in x-aixs, x2 (=x+0.2), x3 (=x+0.4) and x4 (=x+0.6)). Right fig: Fitting quantile regression models 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)).Results show that in single predictors case, outliers moving right in x make significant change to the quantile regression coefficients estimations.

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 (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.

4.0.2 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

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

4.0.3 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 states; (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:

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 stats

In single variable case, we can use scatter plot to represent the outlier stats. The following code showed how to display suspicious outliers based on quantile regression models.

In single variable case, we can use scatter plot to represent the outlier stats. The following code showed how to display suspicious outliers based on quantile regression models.

```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"))
```

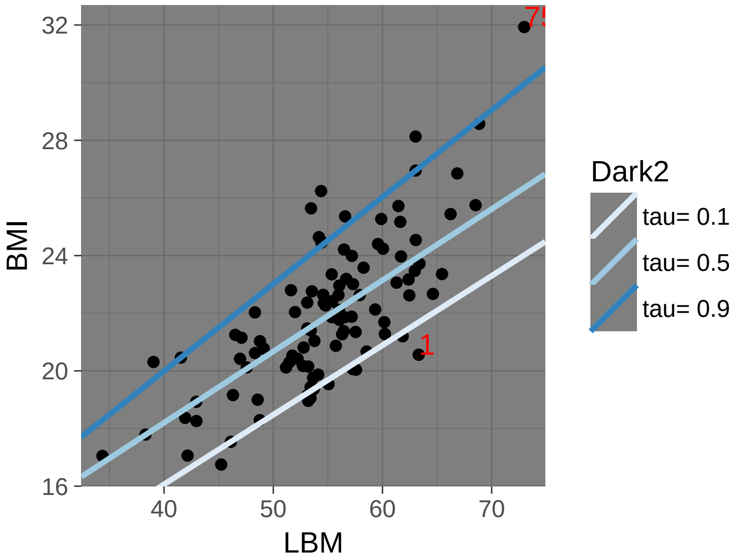

Figure 6: Plot the outlier stats.

```
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("Dark2") +
  theme_dark()
```

5.2 Plot data with outliers marked

Scatter plot has limitations when tackling multi-variable regression cases. 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 in regression models.

```
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 and overlaying the cutoff lines.

```
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`. This function returns generalized cook or q function distance for regression model 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.

```
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') +
```

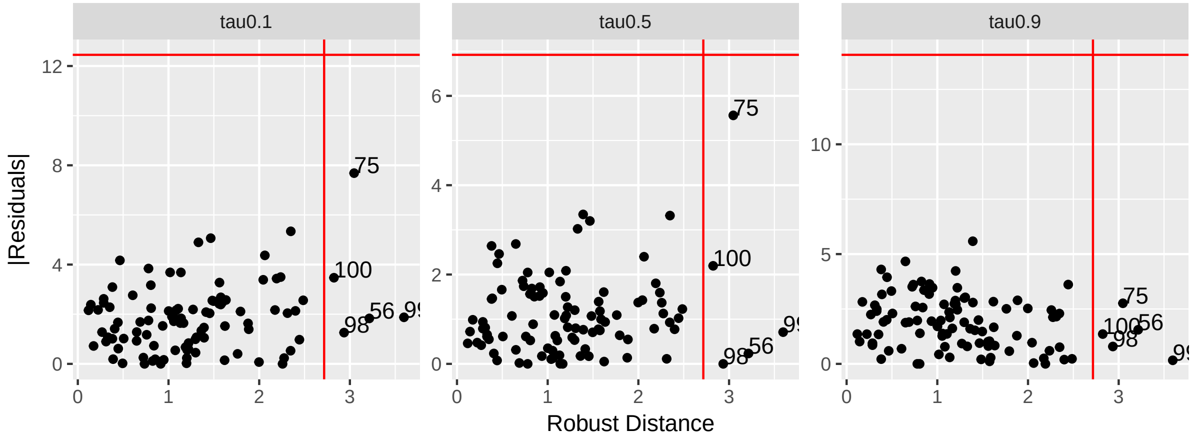

Figure 7: Robust Distance-Residual Plot. Points on the right of vertical cutoff line are considered leverage points and points above the horizental cutoff line are outliers in y-direction.

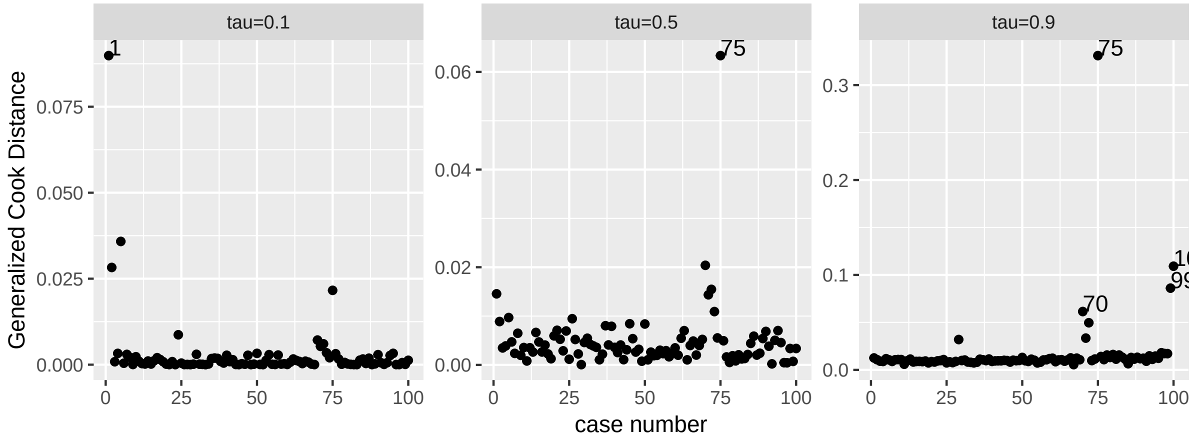

Figure 8: Generalized cook distance of each observation on quantile 0.1, 0.5 and 0.9. Case 75 has relative large cook distance-funtion distance to other points

```
    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")
```

The same, visualization of Q function diagnostic results are shown in fig,

```
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')
```

Same as above, we also applied mean post probability, KL divergence to diagnose,

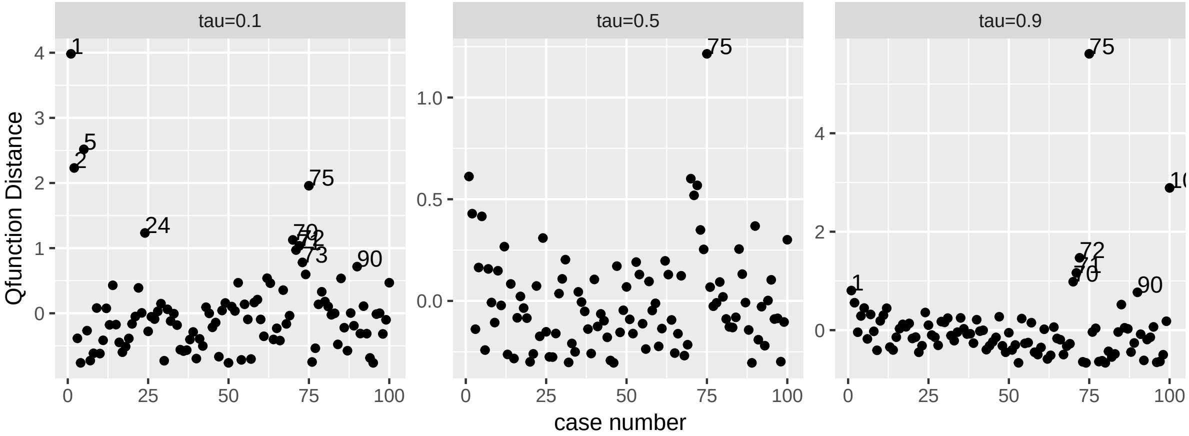

Figure 9: Q function distance of each observation on quantile 0.1, 0.5 and 0.9. Case 75 has relative large Q function distance to other points

6 Visualizing quantile regression

Visualization of quantile regression will help us understand the questions 'How does the shape of the model compare to the shape of the data?'. In addition, we can have a good impression of the location of models on each quantile. We use GGobi to visualize quantile regression model.

6.1 Linear quantile regression model

In two predictors case, quantile regression models are lines in space. We use ais data fitting models and visualize them with GGobi.

In three predictor case, quantile regression models are cuboids in space which were displayed as follows,

6.2 Non-linear quantile regression model

In non-linear case, we use elliptic hyperboloid and hyperbolic paraboloid as examples.

7 Future work

high-dimensional and extreme quantile work.

8 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.

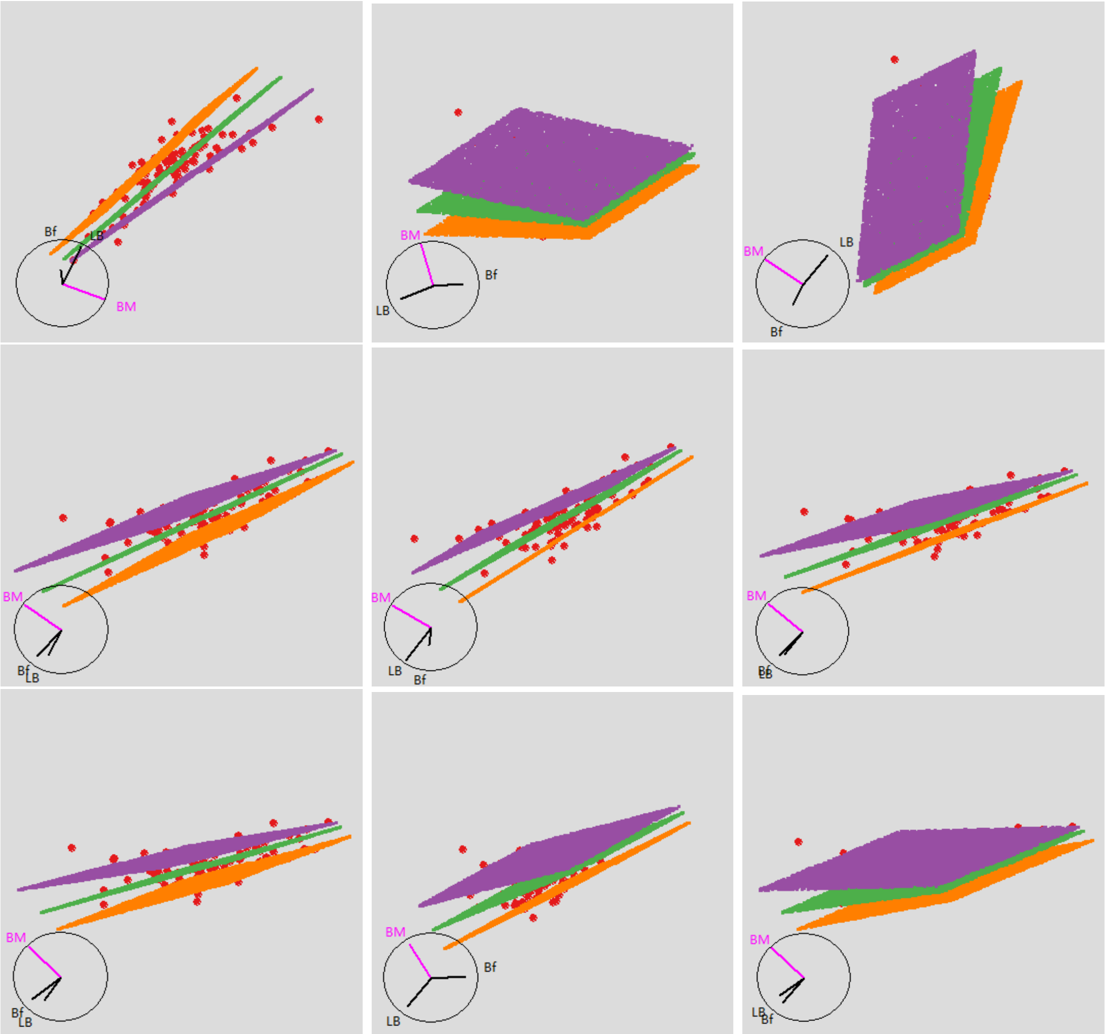

Figure 10: Markdown supported string as caption 1

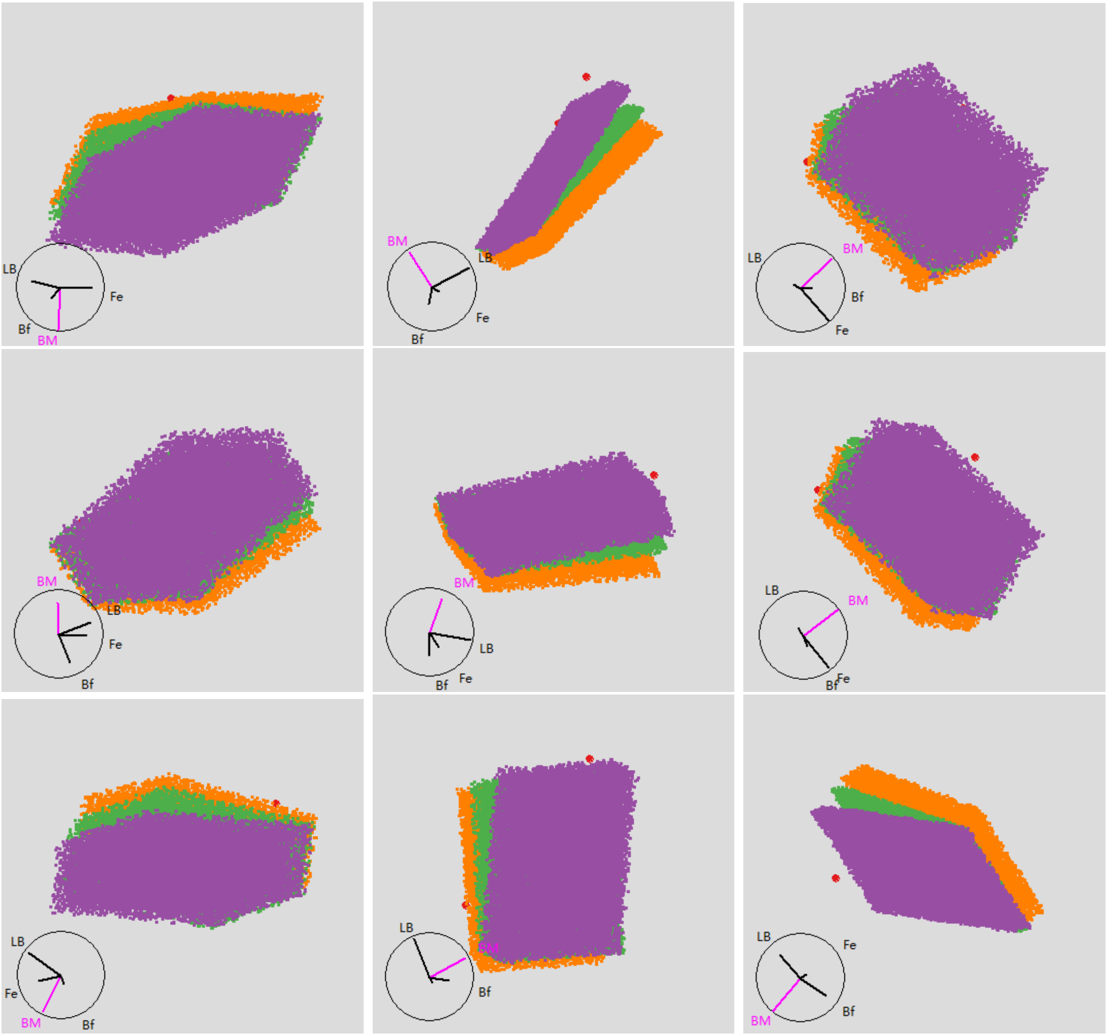

Figure 11: Markdown supported string as caption 2

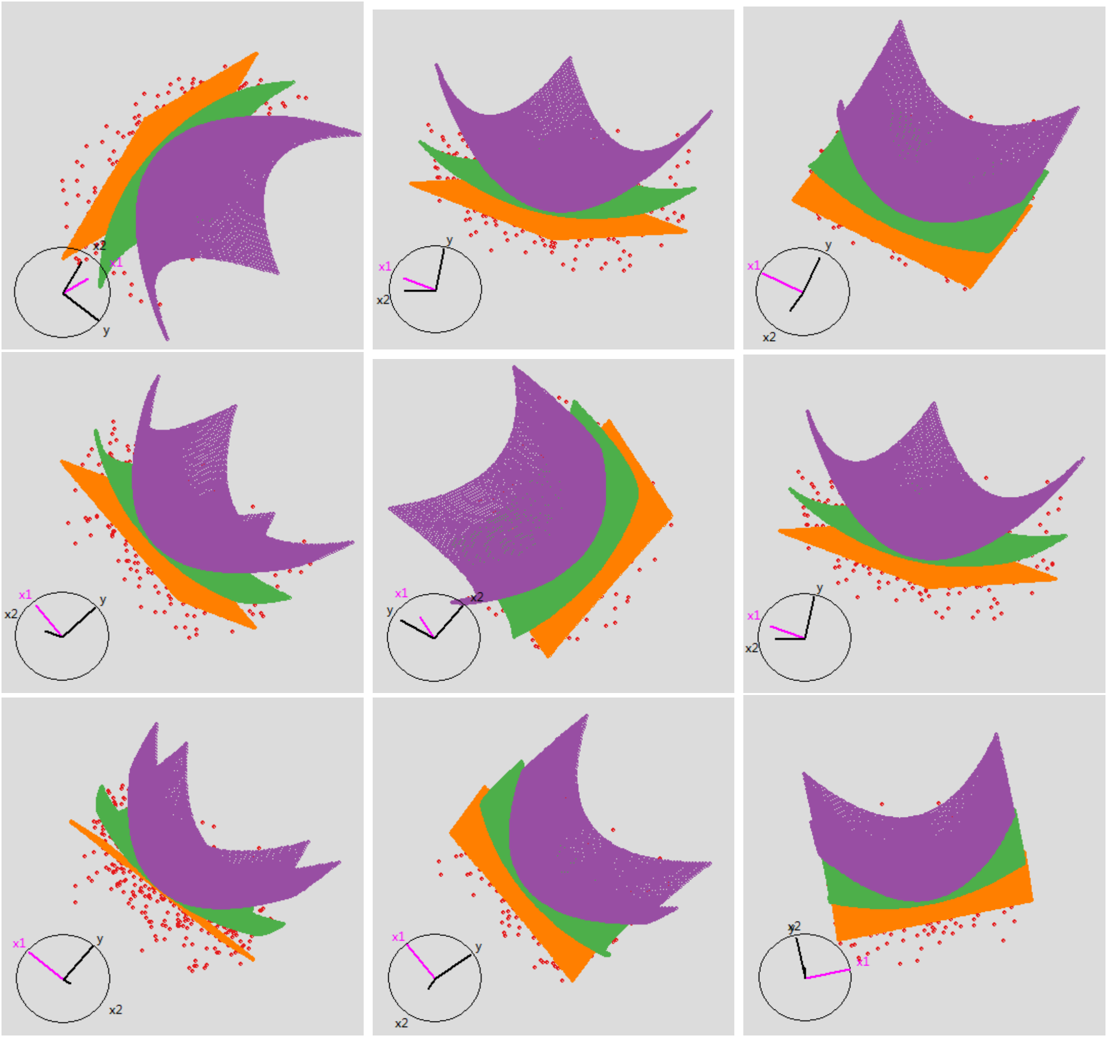

Figure 12: Markdown supported string as caption 3

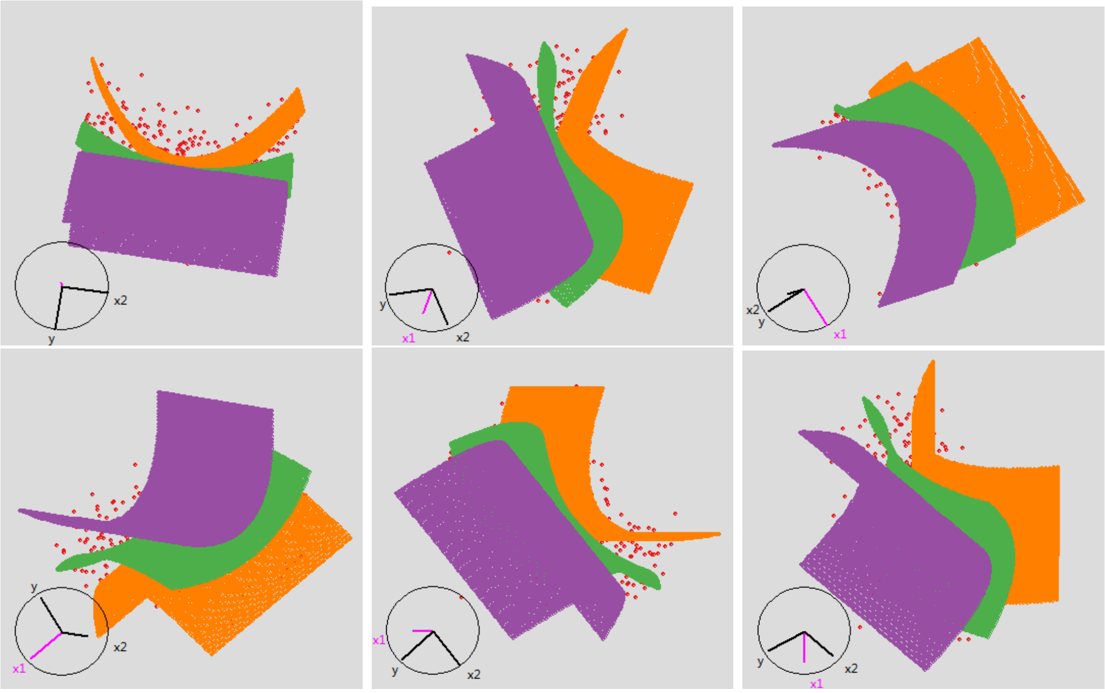

Figure 13: Markdown supported string as caption 4

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.