GSoC 2018 project: Diagnostic statistics and visualization for quantile regression

Wenjing Wang

March 26, 2018

1 Project Info

- Project Title: Diagnostic statistics and visualization for quantile regression
- Project short Title: An extension to R package **quokar**
- URL of project idea page: https://github.com/rstats-gsoc/gsoc2018/wiki/Diagnostic-statistics-and-visualization-for-quantile-regression

2 Bio of Student

I am a third year PhD student at the Department of Finance at Vrije Universiteit Brussel (VUB) and the Department of Statistics at Renmin University of China (RUC). My supervisors are Prof. Kris Boudt at VUB and Prof. Minxue Gao at RUC.

One part of my PhD researches focusses on the outlier diagnostics for quantile regression model. Quantile regression (QR) model was first introduced by Koenker and Bassett Jr (1978). Comparing to traditional regression models, it is capable of modeling the entire conditional distribution of dependent variable rather than only the mean value. For application, QR models are widely used in process control and risk management. I am interested in the diagnostics of influential data of QR model. Outlier diagnostic is one of the crutial parts in regression modeling. Influential observations in regression model may indicate the sample peculiarity or data entry error problems. They will affect the estimation of the coefficients.

In the course of my research, I find that quantile regression has a variety of coefficient estimation methods. Many of these methods have already been implemented in R packages such as **quantreg**, **gbm**, **quantregForest**, **qrnn**, **ALDqr** and **bayesQR**. Literatures in outlier diagnosing for quantile regression model is

relatively new (Chen (2005), Benites et al. (2015), Santos and Bolfarine (2016)), and these methods are implemented in R package **quokar**.

Currently, in this package we have several methods implemented in function **frame_distance**, **frame_mle** and **frame_bayes**. More detail, absolute residual and robust distance method has no distribution assumptions on the error term. Genearal Cook's distance and Q-function distance method are based on asymmetric Laplace distribution error term assumption in QR model. Mean posterior probability and Kullback-Leibler divergence are developed under the Baysian quantile regression framework.

Recently, the research on sensitivity analysis of quantile regression has attracted more and more attention. This project aims to extend diagnostic statistics in **quokar** package. And this project will provide users with more choices for observing the outliers of quantile regression model, thereby reducing the risk outliers brings to model estimation inaccuracy.

To improve the improve the functionality of **frame_distance** in **quokar**, new studentized residual and leverage statistics based on elemental sets (ES) will be used. Elemental sets consists exactly minimum number of observations to fit the regression model parameters. The elemental sets method involves performing many fits to a data set, and each fit made to a subsample of size just large enough to estimate the parameters in the model. Elemental sets method have been proposed as a computational device to approximate estimators in the areas of high breakdown regression and multivariate location/scale estimation. The quantile regression coefficients estimator can be obtained as solutions to an optimization problem which have close relationship with the elemental regression. In QR framework, we will calculate the elemental set based on the LP optimization solution of quantile regression and develop outlier diagnostic statistics.

In this GSoC project I will use elemental set method to improve the functionality of functions **frame_distance** in R package **quokar**.

Further, I will add new functions to calculate depth quantiles estimator. I will use this robust regression as a tool for analyzing outliers in QR. The basic procedures are first fit robust regression that dose justice to the majority of the data and then discovers the outliers by observing the residuals.

As the maintainer of R package **quokar**, I am familiar with the structure and functions in this package. I also use R to develop and test the methodologies developed in my research.

3 Contact Information

- Student name: Wenjing Wang

- Student postal address: A204, Avenue de la Couronne 365, 1050 Brussels
- Telephone: Personal +32(0)486206338
- Email(s): Wenjing.Wang@vub.ac.be, wenjingwang1990@ruc.edu.cn
- Skype: Wenjing Wang

4 Student Affiliation

- Institution: Vrije Universiteit Brussel
- Program: Applied Economics
- Contact to verify: Prof. Kris Boudt (kris.boudt@vub.be)

5 Schedule Conflicts

I do not have any kind of schedule conflicts during the summer for the GSoC 2018 project.

6 Mentors

- Mentor names: Prof. Dianne Cook and Prof. Kris Boudt
- Mentor emails: dicook@monash.edu; Kris.Boudt@vub.be

7 Coding plan and methods

7.1 Review methods

The diagnotic statistics for quantile regression were first developed by Koenker and Machado (1999). They introduced goodness-of-fit process for quantile regression analogous to the conventional R^2 statistics of least squares regression. On the contrary, outlier diagnostic methods for quantile regression are relatively new in literature. John (2015) first investigated the influence function of quantile regression estimator. Quantile regression inherits robustness property since bounded influence function of quantiles. Santos and Elian (2015) use the asymmetric Laplace distribution to define likelihood displacement method to observe outliers in quantile regression model. Benites et al. (2015) introduced generalized Cook's distance method to detect influence points in QR models. Santos and Bolfarine (2016) proposed methods based on the posterior distribution of the latent variable in Bayesian quantile regression model to diagnose influencial data.

Noh et al. (2013), Ranganai and Nadarajah (2017) introduced elemental sets method to analyze the quality fit in the framework of quantile regression and provide a predictive leverage statistic. Ranganai (2016) proposed studentized residual for quantile regression models based on elemental sets.

Other literatures focus on developing robust estimators for quantile regression which also can be used for influential diagnostic (see Rousseeuw and Hubert (1999), Struyf and Rousseeuw (1999), Van Aelst et al. (2002)).

7.2 Coding plan

The goal of this project is to extend outlier diagnostic methods for quantile regression in R package **quokar**. There are three parts in this project: (1) Improve functionality of function **frame_distance** by using studentized residual and new leverage statistics based on elemental sets; (2) Implementation of the depth quantiles estimators in R function and provide visualizations to easily spot outliers; (3) Add documentation in vignette.

7.2.1 Improve functionality of function

Currently in function **frame_distance**, we use absolute residuals to diagnose outliers in y-space for the regression model on each quantile. And we use robsut distance to detect the leverages in x-space. Robust distance can give us information on the influential observations in covariate matrix by calculating MCD for each observation. However, for quantile regression, we use the same covariate matrix but putting different weights on observations when modeling on different quantiles. Thus, robust distance is not accurate enough to be the leverage diagnotic statistic for QR model. Recently Ranganai and Nadarajah (2017) introduced a more proper leverage diagnostic statistic based on elemental division of the covariate matrix. I will implement this method into R code to improve the functionality of exsiting function **frame_distance**.

Ranganai (2016) developed the studentized residual statistics to detect the outliers in quantile regression model. This methods is more accurate than absolute residuals used in function **frame_distance** in quokar. However, Ranganai (2016) only considered the Normally distributed measurement errors. I will generate this method for multivariate skew Laplace distribution and implement it in R function.

7.2.2 Add new function to diagnose outliers based on depth regression

I will implement depth quantile estimator in R function **depth_qr**. Based on this estimator, we calculate the residuals of all data. Observations with unusual large residuals considered to be outlier for QR model.

Regression depth of a fit θ from regression relative to a data set Z_n is the smallest number of observations that need to be removed to make θ a nonfit. It gives an indication of how well the data surrounds the hyperplane. Van Aelst et al. (2002) prove that deepest regression is a consistent estimator of the median regression quantile. Rousseeuw and Hubert (1999) show that the breakdown value of the deepest regression is around 33% which is more robust comparing to the L1-estimator which has breakdown value 0%.

The τth depth quantile θ_τ can be estimated by maximizing function

$$\inf_{\gamma \in S^p} \sum_{i=1}^{n} \Psi_\tau(y_i - x^{'}\theta)\text{sign}(x^{'}\gamma) \tag{1}$$

where $\Psi_\tau(u) = \tau - I(u < 0)$, $S^p = \{r \in R^p : ||r|| = 1\}$,
and

$$\begin{aligned} L^+(v) &= \#\{j; x_j \leq v \text{ and } r_j \geq 0\} \\ R^-(v) &= \#\{j; x_j > v \text{ and } r_j < 0\} \end{aligned} \tag{2}$$

where $r_i(\theta)$ denotes residuals of regression.

Struyf and Rousseeuw (1999) construct an algorithm to compute the regression depth of a plane relative to a three-dimensional data set in $O(n^2 \log n)$ time. And for data sets with large n or p, they propose an approximate algorithm which computes the depth of a regression fit in $O(mp^3 + mpn + mn \log n)$ time. The maximal depth (or deepest regression) estimator has been studied in Van Aelst et al. (2002). Fortran code examples to calculate the deepest regression estimator can be found in https://wis.kuleuven.be/stat/robust/Programs. Based on the previous work, I will extended this algorithm to estimate general depth quantiles, and implement it in R function. Then we can calculate residuals for models on each quantile and easily detect outliers.

The geometrical explaination of regression depth is the smallest number of observations the coefficients hyperplane has to pass in order to turn the hyperplane into vertical position. Hyperplanes with high regression depth fit the data better than hyperplanes with low depth. Based on this clear geometrical explaination, visualization can be a good way to illustrate the model fit together with spotting the outliers. I will also provide some visualization methods to spot the outlier in the framework of depth quantiles. I will assign the diagnose results and plot related data to a depth quantile method. In the summary method, the previously calculated result will be displayed. Similarly, in the plot.method implementation, users can directly use this method to draw.

7.3 Documentationn and vignette

I will write the vignette to explain the new methods and perform some real data examples to show the usage of each function.

8 Time table

The workload is organized into three main time periods:

- From 14/05 to 15/06: during this period I will improve the functionality of **frame_distance**. I will implement elemental set method for studentize residual and leverage statistics into function and compare their performance with the former function. Then, I will extend the normal distribution assumption in the leverage statistics to asymmetric Laplace distribution and implement the improvment.

- From 16/06 to 22/07: during this period I will write function to implement depth quantile estimator for estimating QR model. Provide plot methods to spot the outliers. Testing the performance of the robust estimator.

- From 23/07 to 22/08: during this time I will do bug corrections and extensive testing of all the methods. I will document all the new methods into the vignette, and document clearly the applicable conditions and assumptions of each method. Then, I will give real data examples to analyze the performance of various methods.

9 Management of Coding Project

I plan to start coding from now. A shared folder with mentors on Github page will be made to store the code. The mentors will trace all the development of the project at any time. Additionally, I will provide a progress report at the end of each month which includes sample code. From June onwards I will have weekly meetings with Kris Boudt in Brussels and online with Dianne Cook.

10 Test

10.1 Test 1

I implement ER method in Ranganai (2016) to diagnose outliers for QR model. To observe the function performance, I use ais data set in package **quokar** to test the performance of this method. Figure (1) and (2) show the scatter plot and regression lines.

Figure (3) shows that the elemental set method introduced in Ranganai (2016) detect observation 75 in low and middle quantiles together with other potential ob-

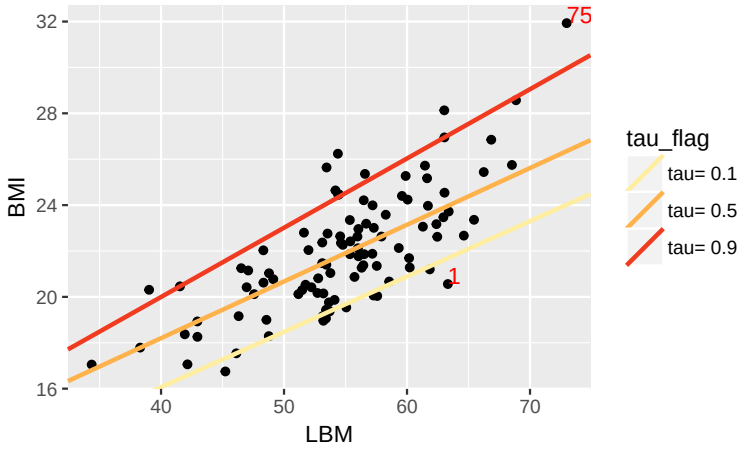

Figure 1: Observation 1, 75 are possible outliers in regression BMI $\sim$ LBM

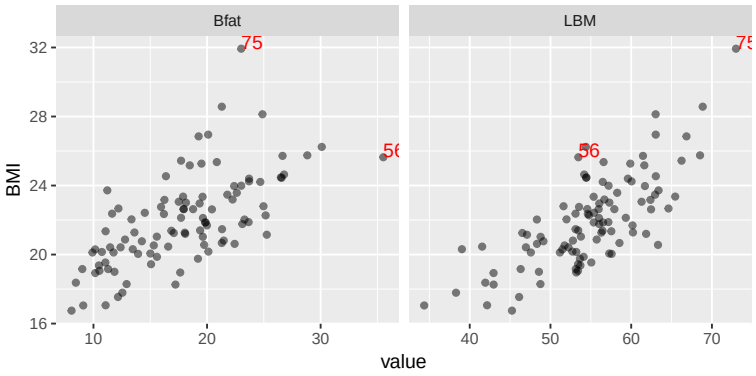

Figure 2: Observation 75, 56 are possible outlier in multivariate regression BMI $\sim$ LBM + Bfat

servations. Figure (4) shows the result from function **frame_distance**. Comparing this two methods, we can conclude that the elemental set methods is more sensitive to influential values on lower quantile while absolute residual and robust distance method is more sensitive to influential values on higher quantile. My next step is to improve elemental set methods to suit for asymmetric Laplace distribution on measure error assumption which is more widely used in QR model frame.

10.2 Test 2

To visualize hyperplane in space, I will use projection method. I explored to plot quantile regression model surface in hyperspace using R package **rggobi**

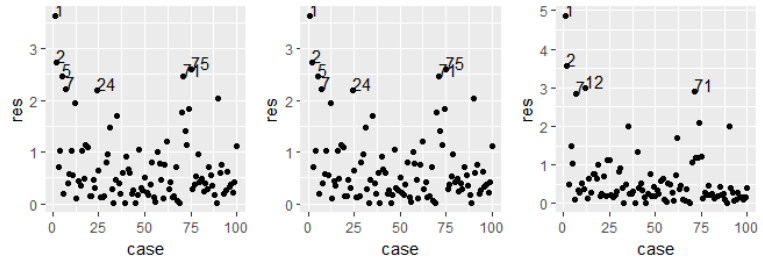

Figure 3: Elemental set method in Ranganai (2016)

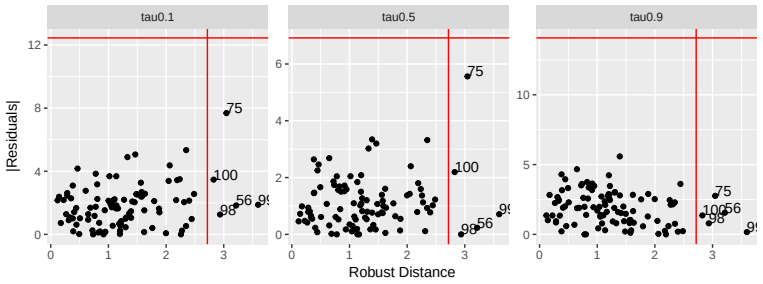

Figure 4: Absolute residual and robust distance method to detect outliers in low, middle and high quantiles.

(Figure 5). This method also will be used in visualizing depth quantiles hyperplane.

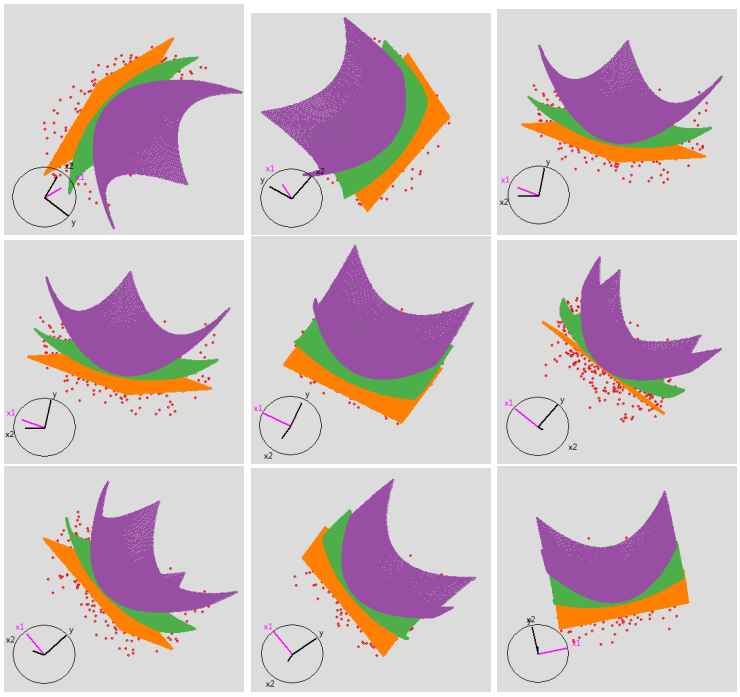

Figure 5: Quantile regression model surface in space. Purple, green and orange representing low quantile, median and high quantile.

References

Benites, L.E., Lachos, V.H., Vilca, F.E., 2015. Case-deletion diagnostics for quantile regression using the asymmetric laplace distribution. arXiv preprint arXiv:1509.05099 .

Chen, C., 2005. An introduction to quantile regression and the quantreg procedure, in: Proceedings of the Thirtieth Annual SAS Users Group International Conference, SAS Institute Inc. Cary, NC.

John, O., 2015. Robustness of quantile regression to outliers. American Journal of Applied Mathematics and Statistics 3, 86–88.

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

Koenker, R., Machado, J.A., 1999. Goodness of fit and related inference processes for quantile regression. Journal of the american statistical association 94, 1296–1310.

Noh, H., Ghouch, A.E., KEILEGOM, I.V., 2013. Quality of fit measures in the framework of quantile regression. Scandinavian Journal of Statistics 40, 105–118.

Ranganai, E., 2016. Quality of fit measurement in regression quantiles: an elemental set method approach. Statistics & Probability Letters 111, 18–25.

Ranganai, E., Nadarajah, S., 2017. A predictive leverage statistic for quantile regression with measurement errors. Communications in Statistics-Simulation and Computation 46, 6385–6398.

Rousseeuw, P.J., Hubert, M., 1999. Regression depth. Journal of the American Statistical Association 94, 388–402.

Santos, B., Bolfarine, H., 2016. On bayesian quantile regression and outliers. arXiv preprint arXiv:1601.07344 .

Santos, B.R., Elian, S.N., 2015. Influence measures in quantile regression models. Communications in Statistics-Theory and Methods 44, 1842–1853.

Struyf, A.J., Rousseeuw, P.J., 1999. Halfspace depth and regression depth characterize the empirical distribution. Journal of Multivariate Analysis 69, 135–153.

Van Aelst, S., Rousseeuw, P.J., Hubert, M., Struyf, A., 2002. The deepest regression method. Journal of Multivariate Analysis 81, 138–166.

Wang, W., Cook, D., Wang, E., . quokar: Quantile Regression Outlier Diagnostics with K Left Out Analysis. URL: `https://github.com/wenjingwang/quokar`. r package version 0.1.0.9000.