# Package ‘quokar’

November 10, 2017

**Title** Quantile Regression Outlier Diagnostics with K Left Out Analysis

**Version** 0.1.0

**Author**
Wenjing Wang <wenjingwangr@gmail.com>, Di Cook <visnut@gmail.com>, Earo Wang <earo.wang@gmail.com>

**Maintainer** Wenjing Wang <wenjingwangr@gmail.com>

**Description** Diagnostics methods for quantile regression models for detecting influential observations: robust distance methods for general quantile regression models; generalized Cook's distance and Q-function distance method for quantile regression models using aymmetric Laplace distribution. Reference
of this method can be found in Luis E. Benites, Víctor H. Lachos, Filidor E. Vilca (2015) <arXiv:1509.05099v1>;
mean posterior probability and Kullback–Leibler divergence methods for Bayes quantile regression model.
Reference of this method is Bruno Santos, Heleno Bolfarine (2016) <arXiv:1601.07344v1>.

**Depends** R (>= 3.3.0)

**License** GPL (>= 2)

**Encoding** UTF-8

**Imports** stats, quantreg, purrr, magrittr, ALDqr, bayesQR, MCMCpack, ggplot2, knitr, gridExtra, GIGrvg, dplyr, tidyr, robustbase, ald

**Type** Package

**NeedsCompilation** yes

**LazyLoad** false

**VignetteBuilder** knitr

**RoxygenNote** 6.0.1

**URL** https://github.com/wenjingwang/quokar

**BugReports** https://github.com/wenjingwang/quokar/issues

**LazyData** true

**Suggests** testthat, rmarkdown

**Repository** CRAN

**Date/Publication** 2017-11-10 10:21:36 UTC

# R topics documented:



---

`ah` *House Price*

---

### Description

House Price of America

---

`ais` *Australia Institute of Sport data*

---

### Description

Data on 102 male and 100 female athletes collected at the Austrialian Institute of Sports

### Format

A data frame with 202 observations on the following 14 variables

**Sex** 0 = male, 1 = female

**Ht** height(cm)

**Wt** weightkg

**LBM** lean body mass

**RCC** red cell count

**WCC** white cell count

**Hc** Hematocrit

**Hg** Hemoglobin

**Ferr** plasma ferritin concentration

**BMI** body mass index, weight/(height)^2

**SSF** sum of skin folds

**Bfat** Percent body fat

**Label** Case Lables

**Sport** Sport

### References

S.Weisberg(2005). Applied Linear Regression, 3rd edition. New York, Section 6.4.

---

## `ALDqr_case_deletion` *Calculate the case-deletion coefficience of the MLE estimation of quantile regression*

---

### Description

Calculate the case-deletion coefficience of the MLE estimation of quantile regression

### Usage

```
ALDqr_case_deletion(y, x, tau, error, iter)
```

### Arguments

| | |
|---|---|
| `y` | Response variable in quantile regression model |
| `x` | Predictors in quantile regression model. Note that: x is the independent variable matrix which including the intercept. That means, if the dimension of independent variables is p and the sample size is n, x is a n times p+1 matrix with the first column being one. |
| `tau` | Quantile |
| `error` | The EM algorithm accuracy of error used in MLE estimation |
| `iter` | The iteration frequancy for EM algorithm used in MLE estimation |

---

`ALDqr_GCD` *Generalized Cook's distance for each observation in quantile regression model*

---

## Description

Generalized Cook's distance for each observation in quantile regression model

## Usage

```
ALDqr_GCD(y, x, tau, error, iter)
```

## Arguments

| | |
|---|---|
| `y` | Dependent variable in quantile regression. Note that: we suppose y follows asymmetric laplace distribution. |
| `x` | indepdent variables in quantile regression. Note that: x is the independent variable matrix which including the intercept. That means, if the dimension of independent variables is p and the sample size is n, x is a n times p+1 matrix with the first column being one. |
| `tau` | quantile |
| `error` | the EM algorithm accuracy of error used in MLE estimation |
| `iter` | the iteration frequancy for EM algorithm used in MLE estimation |

## Details

Gerneralized Cook's distance is a commonly used estimate of the influence of a data point when performing regression analysis. It involves the log-likelihood function based on the complete data and case-deletion data. To assess the influence of the $i$th case with estimate $\hat{\theta}$, we compare $\theta\hat{(i)}$ and $\hat{\theta}$, and if $\theta\hat{(i)}$ is far from $\theta\hat{(i)}$, then the $i$th case is regarded as influential. We consider here the following generalized Cook's distance:

$$GCD_i = (\theta\hat{_{(i)}} - \hat{\theta}i)' - Q(\hat{\theta}|\hat{\theta})(\theta\hat{_{(i)}} - \hat{\theta}i)$$

$$Q_{(i)}(\theta|\hat{\theta}) = E_{\hat{\theta}}[l_c(\theta|Y_{c(i)})|y]$$

More details please refer to the paper in references

## 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*.

## See Also

`ALDqr_QD`

ALDqr_QD — *Q-function distance for each observation in quantile regression model*

## Description

Q-function distance for each observation in quantile regression model

## Usage

```
ALDqr_QD(y, x, tau, error, iter)
```

## Arguments

| | |
|---|---|
| y | Dependent variable in quantile regression. Note that: we suppose y follows asymmetric laplace distribution. |
| x | Indepdent variables in quantile regression. Note that: x is the independent variable matrix which including the intercept. That means, if the dimension of independent variables is p and the sample size is n, x is a n times p+1 matrix with the first column is one. |
| tau | Quantile |
| error | The EM algorithm accuracy of error used in MLE estimation |
| iter | The iteration frequancy for EM algorithm used in MLE estimation |

## Details

Measure of the influence of the $i$th case is the following Q-distance function, similar to the likelihood distance $LD_i$ (Cook and Weisberg, 1982), defined as

$$QD_i = 2Q(\hat{\theta}|\hat{\theta}) - Q(\hat{\theta_{(i)}})$$

## 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*.

## See Also

ALDqr_GCD

`baseball` *Baseball Hitter Data*

## Description

Major League Baseball Data from the 1986 and 1987 seasons

## Format

Data frame with 322 rows and 22 columns

**AtBat** Number of times at bat in 1986

**Hits** Number of hits in 1986

**HmRun** Number of home runs in 1986

**Runs** Number of runs in 1986

**RBI** Number of runs batted in in 1986

**Walks** Number of walks in 1986

**Years** Number of years in the major leagues

**CAtBat** Number of times at bat during his career

**CHits** Number of hits during his career

**CHmRun** Number of home runs during his career

**CRuns** Number of runs during his career

**CRBI** Number of runs batted in during his career

**CWalks** Number of walks during his career

**League** A factor with levels A and N indicating player's league at the end of 1986

**Division** A factor with levels E and W indicating player's division at the end of 1986

**PutOuts** Number of put outs in 1986

**Assists** Number of assists in 1986

**Errors** Number of errors in 1986

**Salary** 1987 annual salary on opening day in thousands of dollars

**NewLeague** A factor with levels A and N indicating player league at the beginning of 1987 bayesKL *Kullback-Leibler divergence for each observation in Baysian quantile regression model*

### Description

Kullback-Leibler divergence for each observation in Baysian quantile regression model

### Usage

```
bayesKL(y, x, tau, M, burn)
```

### Arguments

| | |
|---|---|
| `y` | vector, dependent variable in quantile regression |
| `x` | matrix, design matrix in quantile regression. |
| `tau` | quantile |
| `M` | the iteration frequancy for MCMC used in Baysian Estimation |
| `burn` | burned MCMC draw |

### Details

Method to address the differences between the posterior distributions from the distinct latent variables in the model, we suggest the use of the Kullback- Leibler divergence as a more precise method of measuring the distance between those latent variables in the Bayesian quantile regression framework. In this posterior information, the divergence is defined as

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

where $f_i$ could be the posterior conditional distribution of $v_i$ and $f_j$ the poserior conditional distribution of $v_j$. We should average this divergence for one observation based on the distance from all others, i.e,

$$KL(f_i) = \frac{1}{n-1}\sum K(f_i, f_j)$$

We expect that when an observation presents a higher value for this divergence, it should also present a high probability value of being an outlier. Based on the MCMC draws from the posterior of each latent vaiable, we estimate the densities using a normal kernel and we compute the integral using the trapezoidal rule.

More details please refer to the paper in references

### References

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

### See Also

`bayesProb`

---

`bayesProb` *Mean posterior probability for each observation in Baysian quantile regression model*

---

### Description

Mean posterior probability for each observation in Baysian quantile regression model

### Usage

```
bayesProb(y, x, tau, M, burn)
```

### Arguments

| | |
|---|---|
| `y` | vector, dependent variable in quantile regression |
| `x` | matrix, design matrix in quantile regression |
| `tau` | quantile |
| `M` | MCMC draws |
| `burn` | burned MCMC draws |

### Details

If we define the variable O_i, which takes value equal to 1 when ith observation is an outlier, and 0 otherwise, then we propose to calculate the probability of an observation being an outlier as:

$$P(O_i = 1) = \frac{1}{n-1} \sum P(v_i > v_j | data) \quad (1)$$

We believe that for points, which are not outliers, this probability should be small, possibly close to zero. Given the natrual ordering of the residuals, it is expected that some observations present greater values for this probability in comparison to others. What we think that should be deemed as an outlier, ought to be those observations with a higher $P(O_i = 1)$, and possibly one that is particularly distant from the others.

The probability in the equation can be approximated given the MCMC draws, as follows

$$P(O_i = 1) = \frac{1}{M} \sum I(v_i^{(l)} > max v_j^k)$$

where $M$ is the size of the chain of $v_i$ after the burn-in period and $v_j^{(l)}$ is the $l$th draw of chain.

More details please refer to the paper in references

### References

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

### See Also

`bayesKL`

### Examples

```
## Not run:
ais_female <- subset(ais, Sex == 1)
y <- ais_female$BMI
x <- cbind(1, ais_female$LBM)
tau <- 0.5
M <- 5000
burn <- 1000
prob <- bayesProb(y, x, tau, M, burn)
case <-  1:100
dat <- data.frame(case, prob)
ggplot(dat, aes(case, prob))+
 geom_point() +
 geom_text(data = subset(dat, prob > mean(prob) + 2*sd(prob)),

## End(Not run)
```

---

`frame_ald` *Density function plot of the error term for quantile regression model using asymmetric Laplace distribution*

---

### Description

density function plot of the error term on each quantile

### Usage

```
frame_ald(y, x, tau, smooth, error, iter)
```

### Arguments

| | |
|---|---|
| `y` | vector, dependent variable of quantile regression |
| `x` | matrix, matrix consisted independent variables of quantie regression |
| `tau` | sigle number or vector, quantiles |
| `smooth` | sigular, default is 100, the larger the smoother of density function |
| `error` | the convergence maximum error |
| `iter` | maximum iterations of the EM algorithm |

### Value

dataframe to plot the density function of the error term

### Examples

```
library(ggplot2)
data(ais)
x <- matrix(ais$LBM, ncol = 1)
y <- ais$BMI
tau = c(0.1, 0.5, 0.9)
ald_data <- frame_ald(y, x, tau, smooth = 10, error = 1e-6,
                      iter = 2000)
ggplot(ald_data) +
   geom_line(aes(x = r, y = d, group = obs, colour = tau_flag)) +
   facet_wrap(~tau_flag, ncol = 1, scale = "free") +
   xlab('') +
   ylab('Asymmetric Laplace Distribution Density Function')
```

---

## frame_ald_weight *Weighting Matrix of Quantile regression using Asymmetric Laplace Distrubtion*

---

### Description

This function calulate the weighting matrix

### Usage

```
frame_ald_weight(y, x, tau, error, iter)
```

### Arguments

| | |
|---|---|
| y | dependent variable of quantile regression |
| x | design matrix of quantile regression |
| tau | quantile must be a scaler |
| error | The EM algorithm accuracy of error used in MLE estimation |
| iter | The iteration frequancy for EM algorithm used in MLE estimation |

### Details

In the estimation procedure in EM algorithm, we can see that $\varepsilon$ is inversely proportional to $d_i = |y_i - x_i^{'}\beta_p^{(k)}|/\sigma$. Hence, $u_i(\theta^k) = \varepsilon_{-1i}(\theta^{(k)})$ can be interpreted as a type of weight for $i$th case in the estimates of $\beta_{(k)^p}$, which tends to be small for outlying observations.

### Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

### Examples

```
library(ggplot2)
library(dplyr)
library(ALDqr)
data(ais)
y <- ais$BMI
x <- cbind(1, ais$LBM)
tau <-  c(0.1, 0.5, 0.9)
error <- 1e-06
iter <- 100
weights <- frame_ald_weight(y, x, tau, error, iter)
weights
```

---

## frame_bayes — *Mean probability of posterior distribution and Kullback-Leibler divergence for observations in Bayesian quantile regression model*

---

### Description

This function give the dataframe to plot the mean probability of posterior and Kullback-leibler divergence of quantile regression model with asymmetric laplace distribution based on bayes estimation procedure.

### Usage

```
frame_bayes(y, x, tau, M, burn, method = c("bayes.prob", "bayes.kl"))
```

### Arguments

| | |
|---|---|
| y | vector, dependent variable in quantile regression |
| x | matrix, design matrix for quantile regression. For quantile regression model with intercept, the firt column of x is 1. |
| tau | sigular or vector, quantiles |
| M | the iteration frequancy for MCMC used in Baysian estimation |
| burn | burned MCMC draw |
| method | the diagnostic method for outlier detection |

### Value

Mean probability or Kullback-Leibler divergence for observations in Bayesian quantile regression model

### Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

### Examples

```
## Not run:
library(ggplot2)
ais_female <- subset(ais, Sex == 1)
y <- ais_female$BMI
x <- matrix(ais_female$LBM, 1)
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')
prob_m <- cbind(case, prob)
ggplot(prob_m, aes(x = case, y = value )) +
  geom_point() +
  geom_text(aes(label = case)) +
  facet_wrap(~variable, scale = 'free') +
  xlab("case number") +
  ylab("Mean probability of posterior distribution")
It takes time to run the following code.
kl <- frame_bayes(y, x, tau, M = 50, burn = 10,
                method = 'bayes.kl')
kl_m <- cbind(case, kl)
ggplot(kl_m, aes(x = case, y = value)) +
  geom_point() +
  geom_text(aes(label = case)) +
  facet_wrap(~variable, scale = 'free')+
  xlab('case number') +
  ylab('Kullback-Leibler')

## End(Not run)
```

---

## frame_br *Visualization of quantile regression model fitting: br algorithem*

---

### Description

get the observation used in br algorithem

### Usage

```
frame_br(object, tau)
```

### Arguments

| | |
|---|---|
| `object` | quantile regression model using br method |
| `tau` | quantiles can be a single quantile or a vector of quantiles |

## Details

This is a function that can be used to create point plot for the observations used in quantile regression fitting based on 'br'method.

## Value

All observations and the observations used in quantile regression fitting using br algorithem

## Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

## Examples

```
library(ggplot2)
library(quantreg)
data(ais)
tau <- c(0.1, 0.5, 0.9)
object1 <- rq(BMI ~ LBM, tau, method = 'br', data = ais)
data_plot <- frame_br(object1, tau)$all_observation
choose <- frame_br(object1, tau)$fitting_point
ggplot(data_plot,
 aes(x=value, y=data_plot[,2])) +
 geom_point(alpha = 0.1) +
 ylab('y') +
 xlab('x') +
 facet_wrap(~variable, scales = "free_x", ncol = 2) +
 geom_point(data = choose, aes(x = x, y = y,
                                     group = tau_flag,
                                     colour = tau_flag,
                                     shape = obs))

object2 <- rq(BMI ~ Ht + LBM + Wt, tau, method = 'br',
            data = ais)
data_plot <- frame_br(object2, tau)$all_observation
choose <- frame_br(object2, tau)$fitting_point
ggplot(data_plot,
 aes(x=value, y=data_plot[,2])) +
 geom_point(alpha = 0.1) +
 ylab('y') +
 xlab('x') +
 facet_wrap(~variable, scales = "free_x", ncol = 2) +
 geom_point(data = choose, aes(x = x, y = y,
                                     group = tau_flag,
                                     colour = tau_flag,
                                     shape = obs))
```

## frame_distance — *Residual-robust distance plot of quantile regression model*

### Description

the standardized residuals from quantile regression against the robust MCD distance. This display is used to diagnose both vertical outlier and horizontal leverage points. Function `frame_distance` only work for linear quantile regression model. With non-linear model, use `frame_distance_implement`

### Usage

```
frame_distance(object, tau)
```

### Arguments

| | |
|---|---|
| `object` | model, quantile regression model |
| `tau` | singular or vectors, quantile |

### Details

The generalized MCD algorithm based on the fast-MCD algorithm formulated by Rousseeuw and Van Driessen(1999), which is similar to the algorithm for least trimmed squares(LTS). The canonical Mahalanobis distance is defined as

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

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{C}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x})$ are the empirical multivariate location and scatter, respectively. Here $x_i = (x_{i1}, ..., x_{ip})^T$ exclueds the intercept. The relation between the Mahalanobis distance $MD(x_i)$ and the hat matrix $H = (h_{ij}) = X(X^TX)^{-1}X^T$ is

$$h_{ii} = \frac{1}{n-1}MD_i^2 + \frac{1}{n}$$

The canonical robust distance is defined as

$$RD(x_i) = [(x_i - T(X))^T C(X)^{-1}(x_i - T(X))]^{1/2}$$

where $T(X)$ and $C(X)$ are the robust multivariate location and scatter, respectively, obtained by MCD. To achieve robustness, the MCD algorithm estimates the covariance of a multivariate data set mainly through as MCD $h$-point subset of data set. This subset has the smallest sample-covariance determinant among all the possible $h$-subsets. Accordingly, the breakdown value for the MCD algorithm equals $\frac{(n-h)}{n}$. This means the MCD estimates is reliable, even if up to $\frac{100(n-h)}{n}$ set are contaminated.

### Value

dataframe for residual-robust distance plot

### Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

### See Also

function frame_distance_complex

### Examples

```
library(quantreg)
library(ggplot2)
library(ALDqr)
library(purrr)
library(robustbase)
library(tidyr)
library(gridExtra)
tau = c(0.1, 0.5, 0.9)
ais_female <- subset(ais, Sex == 1)
object <- rq(BMI ~ LBM + Ht, data = ais_female, tau = tau)
plot_distance <- frame_distance(object, tau = c(0.1, 0.5, 0.9))
distance <- plot_distance[[1]]
cutoff_v <- plot_distance[[2]]
cutoff_h <- plot_distance[[3]]
n <- nrow(object$model)
case <- rep(1:n, length(tau))
distance <- cbind(case, distance)
distance$residuals <- abs(distance$residuals)
distance1 <- subset(distance, tau_flag == "tau0.1")
p1 <- ggplot(distance1, aes(x = rd, y = residuals)) +
 geom_point() +
 geom_hline(yintercept = cutoff_h[1], colour = "red") +
 geom_vline(xintercept = cutoff_v, colour = "red") +
 geom_text(data = subset(distance1, residuals > cutoff_h[1]|rd > cutoff_v),
           aes(label = case), hjust = 0, vjust = 0) +
 xlab("Robust Distance") +
 ylab("|Residuals|")

distance2 <- subset(distance, tau_flag == "tau0.5")

p2 <- ggplot(distance1, aes(x = rd, y = residuals)) +
 geom_point() +
 geom_hline(yintercept = cutoff_h[2], colour = "red") +
 geom_vline(xintercept = cutoff_v, colour = "red") +
 geom_text(data = subset(distance1, residuals > cutoff_h[2]|rd > cutoff_v),
          aes(label = case), hjust = 0, vjust = 0) +
 xlab("Robust Distance") +
 ylab("|Residuals|")
distance3 <- subset(distance, tau_flag == "tau0.9")

p3 <- ggplot(distance1, aes(x = rd, y = residuals)) +
 geom_point() +
 geom_hline(yintercept = cutoff_h[3], colour = "red") +
```

```
 geom_vline(xintercept = cutoff_v, colour = "red") +
 geom_text(data = subset(distance1, residuals > cutoff_h[3]|rd > cutoff_v),
         aes(label = case), hjust = 0, vjust = 0) +
xlab("Robust Distance") +
 ylab("|Residuals|")
grid.arrange(p1, p2, p3, ncol = 3)
```

---

`frame_distance_complex` *Residual-robust distance plot of quantile regression model*

---

### Description

the standardized residuals from quantile regression against the robust MCD distance. This display is used to diagnose both vertical outlier and horizontal leverage points. Function `frame_distance` only work for linear quantile regression model. With non-linear model, use `frame_distance_complex`

### Usage

```
frame_distance_complex(x, resid, tau)
```

### Arguments

| | |
|---|---|
| `x` | matrix, covariate of quantile regression model |
| `resid` | matrix, residuals of quantile regression models |
| `tau` | singular or vectors, quantile |

### Details

The generalized MCD algorithm based on the fast-MCD algorithm formulated by Rousseeuw and Van Driessen(1999), which is similar to the algorithm for least trimmed squares(LTS). The canonical Mahalanobis distance is defined as

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

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{C}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x})$ are the empirical multivariate location and scatter, respectively. Here $x_i = (x_{i1}, ..., x_{ip})^T$ excluеds the intercept. The relation between the Mahalanobis distance $MD(x_i)$ and the hat matrix $H = (h_{ij}) = X(X^TX)^{-1}X^T$ is

$$h_{ii} = \frac{1}{n-1}MD_i^2 + \frac{1}{n}$$

The canonical robust distance is defined as

$$RD(x_i) = [(x_i - T(X))^T C(X)^{-1}(x_i - T(X))]^{1/2}$$

where $T(X)$ and $C(X)$ are the robust multivariate location and scatter, respectively, obtained by MCD. To achieve robustness, the MCD algorithm estimates the covariance of a multivariate data set mainly through as MCD $h$-point subset of data set. This subset has the smallest sample-covariance determinant among all the possible $h$-subsets. Accordingly, the breakdown value for the MCD algorithm equals $\frac{(n-h)}{n}$. This means the MCD estimates is reliable, even if up to $\frac{100(n-h)}{n}$ set are contaminated.

### Value

dataframe for residual-robust distance plot

### Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

---

`frame_fn_obs` *Visualization of quantile regression model fitting: interior point algorithm*

---

### Description

observations used in quantile regression fitting

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

where $\rho_\tau(r) = r[\tau - I(r < 0)]$ for $\tau \in (0, 1)$. This yields the modified linear program

$$min(\tau e^{'} u + (1-\tau) e^{'} v | y = Xb + u - v, (u, v) \in R_{+}^{2n})$$

Adding slack variables, $s$, satisfying the constrains $a + s = e$, we obtain the barrier function

$$B(a, s, u) = y^{'} a + \mu \sum_{i=1}^{n} (loga_i + logs_i)$$

which should be maximized subject to the constrains $X^{'} a = (1 - \tau) X^{'} e$ and $a + s = e$. The Newton step $\delta_a$ solving

$$maxy^{'} \delta_a + \mu \delta_a^{'} (A^{-1} - S^{-1}) e - \frac{1}{2} \mu \delta_a^{'} (A^{-2} + S^{-2}) \delta_a$$

subject to $X' \delta_a = 0$, satisfies

$$y + \mu (A^{-1} - S^{-1}) e - \mu (A^{-2} + S^{-2}) \delta_a = Xb$$

for some $b \in R^p$, and $\delta_a$ such that $X^{'} \delta_a = 0$. Using the constraint, we can solve explicitly for the vector $b$,

$$b = (X^{'}WX)^{-1}X^{'}W[y + \mu(A^{-1} - S^{-1})e]$$

where $W = (A^{-2} + S^{-2})^{-1}$. This is a form of the primal log barrier algorithm described above. Setting $\mu = 0$ in each step yields an affine scaling variant of the algorithm. The basic linear algebra of each iteration is essentially unchanged, only the form of the diagonal weighting matrix $W$ has chagned.

## Usage

```
frame_fn_obs(object, tau)
```

## Arguments

| | |
|---|---|
| `object` | quantile regression model using interior point method for estimating |
| `tau` | quantile |

## Details

This function used to illustrate data used in fitting process of quantile regression based on interior point method. Koenker and Bassett(1978) introduced asymmetric weight on positive and negative residuals, and solves the slightly modified l1-problem.

## Value

Weighted observations in quantile regression fitting using interior point algorithm

## Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

## References

Portnoy S, Koenker R. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 1997, 12(4): 279-300.

## Examples

```
library(ggplot2)
library(quantreg)
library(tidyr)
library(dplyr)
library(gridExtra)
data(ais)
tau <- c(0.1, 0.5, 0.9)
object <- rq(BMI ~ LBM + Ht, data = ais, tau = tau, method = 'fn')
fn <- frame_fn_obs(object, tau)
##For tau = 0.1, plot the observations used in quantile regression
##fitting based on interior point method
fn1 <- fn[ ,1]
```

```
case <- 1:length(fn1)
fn1 <- cbind(case, fn1)
m <- data.frame(y = ais$BMI, x1 = ais$LBM, x2 = ais$Ht, fn1)
p <- length(attr(object$coefficients, "dimnames")[[1]])
m_f <- m %>% gather(variable, value, -case, -fn1, -y)
mf_a <- m_f %>%
 group_by(variable) %>%
 arrange(variable, desc(fn1)) %>%
 filter(row_number() %in% 1:p )
p1 <- ggplot(m_f, aes(x = value, y = y)) +
 geom_point(alpha = 0.1) +
 geom_point(data = mf_a, size = 3) +
 facet_wrap(~variable, scale = "free_x")
## For tau = 0.5, plot the observations used in quantile regression
##fitting based on interior point method
fn2 <- fn[,2]
case <- 1: length(fn2)
fn2 <- cbind(case, fn2)
m <- data.frame(y = ais$BMI, x1 = ais$LBM, x2 = ais$Ht, fn2)
p <- length(attr(object$coefficients, "dimnames")[[1]])
m_f <- m %>% gather(variable, value, -case, -fn2, -y)
mf_a <- m_f %>%
   group_by(variable) %>%
   arrange(variable, desc(fn2)) %>%
   filter(row_number() %in% 1:p )
p2 <- ggplot(m_f, aes(x = value, y = y)) +
   geom_point(alpha = 0.1) +
   geom_point(data = mf_a, size = 3) +
   facet_wrap(~variable, scale = "free_x")
## For tau = 0.9
fn3 <- fn[,3]
case <- 1: length(fn3)
fn3 <- cbind(case, fn3)
m <- data.frame(y = ais$BMI, x1 = ais$LBM, x2 = ais$Ht, fn3)
p <- length(attr(object$coefficients, "dimnames")[[1]])
m_f <- m %>% gather(variable, value, -case, -fn3, -y)
mf_a <- m_f %>%
  group_by(variable) %>%
  arrange(variable, desc(fn3)) %>%
  filter(row_number() %in% 1:p )
p3 <- ggplot(m_f, aes(x = value, y = y)) +
  geom_point(alpha = 0.1) +
  geom_point(data = mf_a, size = 3) +
  facet_wrap(~variable, scale = "free_x")
grid.arrange(p1, p2, p3, ncol = 1)
```

## frame_fn_path *Visualization of the fitting path of quantile regression: interior point method*

## Description

observations used in quantile regression fitting

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

where $\rho_\tau(r) = r[\tau - I(r < 0)]$ for $\tau \in (0, 1)$. This yields the modified linear program

$$min(\tau e^{'} u + (1-\tau)e^{'} v | y = Xb + u - v, (u, v) \in R_{+}^{2n})$$

Adding slack variables, $s$, satisfying the constrains $a + s = e$, we obtain the barrier function

$$B(a, s, u) = y^{'} a + \mu \sum_{i=1}^{n} (log a_i + log s_i)$$

which should be maximized subject to the constrains $X^{'} a = (1 - \tau)X^{'} e$ and $a + s = e$. The Newton step $\delta_a$ solving

$$max y^{'} \delta_a + \mu \delta_a^{'} (A^{-1} - S^{-1})e - \frac{1}{2} \mu \delta_a^{'} (A^{-2} + S^{-2})\delta_a$$

subject to $X' \delta_a = 0$, satisfies

$$y + \mu(A^{-1} - S^{-1})e - \mu(A^{-2} + S^{-2})\delta_a = Xb$$

for some $b \in R^p$, and $\delta_a$ such that $X^{'} \delta_a = 0$. Using the constraint, we can solve explicitly for the vector $b$,

$$b = (X^{'} W X)^{-1} X^{'} W [y + \mu(A^{-1} - S^{-1})e]$$

where $W = (A^{-2} + S^{-2})^{-1}$. This is a form of the primal log barrier algorithm described above. Setting $\mu = 0$ in each step yields an affine scaling variant of the algorithm. The basic linear algebra of each iteration is essentially unchanged, only the form of the diagonal weighting matrix $W$ has chagned.

## Usage

```
frame_fn_path(object, tau)
```

## Arguments

| | |
|---|---|
| `object` | quantile regression model using interior point method |
| `tau` | quantile |

## Details

This function used to illustrate the fitting process of quantile regression using interior point method. Koenker and Bassett(1978) introduced asymmetric weight on positive and negative residuals, and solves the slightly modified l1-problem.

## Value

The fitting path of quantile regression model using interior point method

## Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

## References

Portnoy S, Koenker R. The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 1997, 12(4): 279-300.

## Examples

```
## Not run:
library(ggplot2)
library(quantreg)
data(ais)
tau <- c(0.1, 0.5, 0.9)
object <-rq(BMI ~ LBM + Ht, tau = tau, data = ais, method = 'fn')
frame_fn <- frame_fn_path(object, tau)
#plot the path
frame_fn1 <- frame_fn[[1]]
ggplot(frame_fn1, aes(x = value, y = objective)) +
   geom_point() +
   geom_path() +
   facet_wrap(~ variable, scale = 'free')

## End(Not run)
```

---

`frame_mle` *General Cook's distance or Q-function distance of quantile regression*

---

## Description

dataframe used to plot generalized Cook's distance or Q-function distance for observations.

## Usage

```
frame_mle(y, x, tau, error = 1e-06, iter = 100,
  method = c("cook.distance", "qfunction"))
```

## Arguments

| | |
|---|---|
| y | vector, dependent variable for quantile regression |
| x | matrix, design matrix for quantile regression. For quantile regression model with intercept, the firt column of x is 1. |
| tau | sigular or vector, quantiles |
| error | the EM algorithm accuracy of error used in MLE estimation |
| iter | the iteration frequancy for EM algorithm used in MLE estimation |
| method | use method 'cook.distance' or 'qfunction' |

## Details

Gerneralized Cook's distance and Q-function distance are commonly used in detecting the influence data point when performing regression analysis. They involve the log-likelihood function and estimations of based on the complete and case-deletion data. We used EM algorithm to estimate the coefficiences of quantile regression with asymmetric Laplace distribution.

## Value

generalized Cook's distance or Q-function distance for multiple quantiles

## Author(s)

Wenjing Wangwenjingwangr@gmail.com

## Examples

```
library(ggplot2)
data(ais)
ais_female <- subset(ais, Sex == 1)
y <- ais_female$BMI
x <- cbind(1, ais_female$LBM, ais_female$Bfat)
tau <- c(0.1, 0.5, 0.9)
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') +
  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')+
geom_text(data = subset(QD_m, value > mean(value) + sd(value)),
          aes(label = case), hjust = 0, vjust = 0) +
xlab('case number') +
ylab('Qfunction Distance')
```

---

`frame_nlrq` *Visualization of fitting process of non-linear quantile regression: interior point algorithm*

---

## Description

This function explore the fitting process of nonlinear quantile regression

## Usage

```
frame_nlrq(formula, data, tau, start)
```

## Arguments

| | |
|---|---|
| `formula` | non-linear quantile regression model |
| `data` | data frame |
| `tau` | quantiles |
| `start` | the initial value of all parameters to estimate, must be a list |

## Details

To extentd the linear programming method to the case of non-linear response functions, Koenker & Park(1996) considered the nonlinear $l_1$ problem

$$min_{t \in R^p} \sum |f_i(t)|$$

where, for example,

$$f_i(t) = y_i - f_0(x_i, t)$$

As noted by El Attar et al(1979) a necessary condition for $t*$ to solve $min_{t \in R^p} \sum |f_i(t)|$ is that there exists a vector $d \in [-1, 1]^n$ such that

$$J(t*)^{'} d = 0$$

$$f(t*)^{'} d = \sum |f_i(t*)|$$

where $f(t) = (f_i(t))$ and $J(t) = (\partial f_i(t) / \partial t_j)$. Thus, as proposed by Osborne and Watson(1971), one approach to solving $min_{t \in R^p} \sum |f_i(t)|$ is to solve a succession of linearized $l_1$ problems minimizing

$$\sum |f_i(t) - J_i(t)^{'} \delta|$$

### Value

Weighted observations in non-linear quantile regression model fitting using interior algorithm

### Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

### Examples

```
library(tidyr)
library(ggplot2)
library(purrr)
x <- rep(1:25, 20)
y <- SSlogis(x, 10, 12, 2) * rnorm(500, 1, 0.1)
Dat <- data.frame(x = x, y = y)
formula <- y ~ SSlogis(x, Aysm, mid, scal)
nlrq_m <- frame_nlrq(formula, data = Dat, tau = c(0.1, 0.5, 0.9))
weights <- nlrq_m$weights
m <- data.frame(Dat, weights)
m_f <- m %>% gather(tau_flag, value, -x, -y)
ggplot(m_f, aes(x = x, y = y)) +
  geom_point(aes(size = value, colour = tau_flag)) +
  facet_wrap(~tau_flag)
```

---

## qrod_bayes — *Outlier Dignostic for Quantile Regression Based on Bayesian Estimation*

---

### Description

This function cacluate the mean probability of posterior of Baysian quantile regression model with asymmetric laplace distribution

### Usage

```
qrod_bayes(y, x, tau, M, burn, method = c("bayes.prob", "bayes.kl"))
```

### Arguments

| | |
|---|---|
| y | dependent variable in quantile regression |
| x | matrix, design matrix for quantile regression. For quantile regression model with intercept, the firt column of x is 1. |
| tau | quantile |
| M | the iteration frequancy for MCMC used in Baysian Estimation |
| burn | burned MCMC draw |
| method | the diagnostic method for outlier detection |

## Details

If we define the variable Oi, which takes value equal to 1 when ith observation is an outlier, and 0 otherwise, then we propose to calculate the probability of an observation being an outlier as:

$$P(O_i = 1) = \frac{1}{n-1} \sum P(v_i > v_j | data) \quad (1)$$

We believe that for points, which are not outliers, this probability should be small, possibly close to zero. Given the natrual ordering of the residuals, it is expected that some observations present greater values for this probability in comparison to others. What we think that should be deemed as an outlier, ought to be those observations with a higher $P(O_i = 1)$, and possibly one that is particularly distant from the others.

The probability in the equation can be approximated given the MCMC draws, as follows

$$P(O_i = 1) = \frac{1}{M} \sum I(v_i^{(l)} > maxv_j^k)$$

where $M$ is the size of the chain of $v_i$ after the burn-in period and $v_j^{(l)}$ is the $l$th draw of chain.

Another proposal to address these differences between the posterior distributions from the distinct latent variables in the model, we suggest the use of the Kullback- Leibler divergence proposed by Kullback and Leibler(1951), as a more precise method of measuring the distance between those latent variables in the Bayesian quantile regression framework. In this posterior information, the divergence is defined as

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

where $f_i$ could be the posterior conditional distribution of $v_i$ and $f_j$ the poserior conditional distribution of $v_j$. Similar to the probability proposal in the previous subsection, we should average this divergence for one observation based on the distance from all others, i.e,

$$KL(f_i) = \frac{1}{n-1} \sum K(f_i, f_j)$$

We expect that when an observation presents a higher value for this divergence, it should also present a high probability value of being an outlier. Based on the MCMC draws from the posterior of each latent vaiable, we estimate the densities using a normal kernel and we compute the integral using the trapezoidal rule.

## Value

Mean probability or Kullback-Leibler divergence for observations in Bayesian quantile regression model

## Author(s)

Wenjing Wang <wenjingwangr@gmail.com>

### 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*.

Hawkins D M, Bradu D, Kass G V.(1984)“Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26(3), 197-208.

Koenker R, Bassett Jr G.(1978)“ Regression quantiles, *Econometrica*, 1, 33-50.

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

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

### See Also

`qrod_mle`

---

## qrod_mle — *Outlier Dignostic for Quantile Regression with Asymmetric Laplace Distribution*

---

### Description

This function cacluate the generalized cook distance and q function distance of quantile regression model with asymmetric laplace distribution.

### Usage

```
qrod_mle(y, x, tau, error, iter, method = c("cook.distance", "qfunction"))
```

### Arguments

| | |
|---|---|
| `y` | Dependent variable in quantile regression |
| `x` | Indepdent variables in quantile regression. Note that: x is the independent variable matrix which including the intercept. That means, if the dimension of independent variables is p and the sample size is n, x is a n times p+1 matrix with the first column being one. |
| `tau` | quantile |
| `error` | The EM algorithm accuracy of error used in MLE estimation |
| `iter` | the iteration frequancy for EM algorithm used in MLE estimation |
| `method` | the diagnostic method for outlier detection |

### Details

please refer to the reference paper

### Value

Generalized Cook’s distance or Q-function distance for multiple quantiles `trout` *Fish habbit of trout*

### Description

The data set trout, which follows, includes the average numbers of Lahontan cutthroat trout per meter of stream, and the width-to-depth ratios for 71 samples

### Format

A data frame with with 71 rows and 2 columns

**wdratio** Width-to-depth ratio of trout

**density** Numbers of Lahontan cutthroat trout per meter of stream

# Index