DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1

Article 33

5-1-2017

JMASM43: TEEReg: Trimmed Elemental Estimation (R)

Wei Jiang
University of Kansas Medical Center, willjiang29@gmail.com

Matthew S. Mayo
University of Kansas Medical Center

Recommended Citation

Jiang, W. & Mayo, M. S. (2017). JMASM43: TEEReg: Trimmed Elemental Estimation (R). Journal of Modern Applied Statistical Methods, 16(1), 613-629. doi: 10.22237/jmasm/1493598780

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 613-629.
doi: 10.22237/jmasm/1493598780

ISSN 1538 − 9472

JMASM43: TEEReg: Trimmed Elemental Estimation (R)

Wei Jiang
University of Kansas Medical Center
Kansas City, KS

Matthew S. Mayo
University of Kansas Medical Center
Kansas City, KS

Trimmed elemental regression is robust to outliers and violations of model assumptions. Its properties and statistical inference were evaluated using bias-corrected and accelerated bootstrap confidence intervals. An R package named TEEReg is developed to compute the trimmed elemental estimates and the corresponding bootstrap confidence intervals. Two examples are provided to demonstrate its usage.

Keywords: Trimmed elemental estimator, robust linear regression, R, bias-corrected and accelerated bootstrap confidence interval

Introduction

Linear regression is useful in discovering relationships between observations and covariates. Assume that $\mathbf{Y}$ is an n-dimensional vector of dependent variables, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters, $\boldsymbol{\epsilon}$ is an n-dimensional vector of random errors with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, and $\mathbf{X}$ is a design matrix with n rows and p columns, the multiple linear regression model can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

For the ordinary least square (OLS) approach, the estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

minimizes the sum of squares of the residuals

Dr. Jiang is a Ph.D candidate. Email them at: willjiang29@gmail.com. Dr. Mayo is a professor in the Department of Biostatistics.

$$\hat{\boldsymbol{\epsilon}}^t\hat{\boldsymbol{\epsilon}} = \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^t \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

Although the OLS approach has advantages of easy calculation and well-developed statistical inference, it is sensitive to outliers and violations of model assumptions.

The weighted least square (WLS) and iterative reweighted least square (IRLS) are commonly employed alternatives to the OLS approach to deal with unequal variances of the error terms and influential outlying observations; see Kutner, Nachtsheim, Neter, and Li (2005) for a complete review. Other examples of IRLS can be found in Schlossmacher (1973), Sposito, Kennedy, and Gentle (1977), Krasker and Welsch (1983), Carroll and Ruppert (1988), and Street, Carroll, and Ruppert (1988). There are some other available alternatives to OLS. In 1760, Boscovich first introduced the absolute values estimator that was put into a more structured form later by Laplace (Dielman, 2005). The concept of regression quantiles was generalized by Koenker and Bassett (1978); see also Koenker and D'Orey (1987), Gutenbrunner and Jureckova (1992), Koenker (1994), and Koenker (2005). The least median of squares regression was developed by Rousseeuw (1984), and Hawkins (1993) introduced the globally best estimator and the best elemental estimator. Most of these alternatives were developed based on modifying fitting criteria.

The trimmed elemental (TE) estimator that is robust to outliers and violations of model assumptions was developed by Mayo and Gray (1997). It belongs to a class of regression estimators called leverage-residual weighted elemental (LRWE) estimators (Mayo & Gray, 1997). Hall and Mayo (2008) explored the inference properties of TE approach by investigating the coverage probability of the associated bias-corrected and accelerated (BCa) bootstrap confidence interval (CI). Compared with the traditional bootstrap methods, the BCa approach proposed by Efron (1987) corrected the bias and skewness of the sampling distribution through adjusting the selected percentiles used for constructing CIs.

The purpose of this article is to provide an R-package called TEEReg to compute the TE estimates and the corresponding BCa bootstrap CIs. This package contains two functions, TEE() and TEE.BCa(), and can be obtained at CRAN at http://cran.r-project.org/web/packages/TEEReg/.

TE Estimator and BCa Bootstrap CI

The TE estimator developed by Mayo and Gray (1997) is robust to outlying cases and violations of model assumptions. It is a solution based on the elemental subset and the elemental regressions.

Elemental Subsets and Elemental Regressions

In most situations, the sample size n is much larger than the number of unknown parameters p. Instead of using all n observations, only p are required to obtain estimates of the p-dimensional vector of unknown parameters defined in model (1). In this case, there are $\binom{n}{p}$ distinct subvectors of the data and thus $\binom{n}{p}$ possible solutions for the vector $\boldsymbol{\beta}$ in which each solution provides an exact fit to the corresponding p observations. Let $h = \{i_1, i_2, \ldots, i_p\}$ be a subset containing p distinct values from the n-dimensional set of indices $\{1, 2, \ldots, n\}$, $\mathbf{X}_h$ denote a p-dimensional square matrix constructed by the rows of $\mathbf{X}$ with corresponding indices, and $\mathbf{Y}_h$ denote a $p \times 1$ subvector of $\mathbf{Y}$ of which elements are those in $\mathbf{Y}$ indexed by the subset h. Then, the subset h is an elemental subset of the data and the solution to $\mathbf{X}_h \hat{\boldsymbol{\beta}}_h = \mathbf{Y}_h$, a system of p equations with p unknowns, is called an elemental regression and is given by

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \frac{\sum_h \left|\mathbf{X}_h^t \mathbf{X}_h\right| \hat{\boldsymbol{\beta}}_h}{\sum_h \left|\mathbf{X}_h^t \mathbf{X}_h\right|} = \sum_h \frac{\left|\mathbf{X}_h^t \mathbf{X}_h\right|}{\left|\mathbf{X}^t \mathbf{X}\right|} \hat{\boldsymbol{\beta}}_h = \sum_h w_h \hat{\boldsymbol{\beta}}_h \tag{2}$$

where $|\mathbf{A}|$ denotes the determinant of matrix $\mathbf{A}$. This indicates that the least squares estimate is a weighted average over all possible elemental estimates $\hat{\boldsymbol{\beta}}_h$ with weights

$$w_h = \frac{\left|\mathbf{X}_h^t \mathbf{X}_h\right|}{\left|\mathbf{X}^t \mathbf{X}\right|}$$

Moreover, Mayo and Gray (1997) demonstrated that the WLS estimator can be formed as a function of elemental regressions. Let v_i denote the weight for observation i, $\mathbf{V}$ be a diagonal matrix containing the weights v_i, and $\mathbf{V}_h$ be a $p \times p$

submatrix of **V** corresponding to the elemental subset *h*. After some calculations, the WLS estimator can be equivalently written as

$$\hat{\boldsymbol{\beta}}_{\text{WLS}} = \frac{\sum_h \left|\mathbf{X}_h^t \mathbf{V}_h \mathbf{X}_h\right| \hat{\boldsymbol{\beta}}_h}{\sum_h \left|\mathbf{X}_h^t \mathbf{V}_h \mathbf{X}_h\right|} = \sum_h \frac{\left|\mathbf{X}_h^t \mathbf{V}_h \mathbf{X}_h\right|}{\left|\mathbf{X}^t \mathbf{V} \mathbf{X}\right|} \hat{\boldsymbol{\beta}}_h = \sum_h w_h^* \hat{\boldsymbol{\beta}}_h \tag{3}$$

In practice, the reciprocal of the variances of error terms is usually employed for weight v_i to deal with unequal error variances (Kutner et al., 2005), so a lesser weight is assigned to an observation with a larger variance than another observation with a smaller variance. Many weight functions were suggested for dampening the influence of outlying observations, including the Huber weight function given below (Kutner et al., 2005):

$$v_i = \begin{cases} 1 & \left|u_i\right| \le 1.345 \\ \dfrac{1.345}{\left|u_i\right|} & \left|u_i\right| > 1.345 \end{cases}$$

where u_i denotes the scaled residual for which a definition can be found in Kutner et al. (2005). It does not reduce the weight of a case from 1 until the absolute scaled residual is greater than 1.345. It is usually suggested to re-estimate the scaled residual using the process of IRLS to obtain revised weights when the initial estimated coefficients are substantially different from the ones obtained by OLS (Kutner et al., 2005).

TE Estimator

The TE estimator is a special case of a class of estimators called leverage-residual weighted elemental (LRWE) estimators developed by Mayo and Gray (1997). The LRWE class consists of all estimators that can be expressed in the form

$$\hat{\boldsymbol{\beta}}(\lambda, \rho) = \frac{\sum_h w\left[\lambda(h), \rho(h)\right] \hat{\boldsymbol{\beta}}_h}{\sum_h w\left[\lambda(h), \rho(h)\right]}$$

where the factor $\lambda(h)$ represents the leverage information related to the elemental subsets *h* and the factor $\rho(h)$ represents the information of degree of fit related to elemental subsets. The OLS estimator defined in formula (2) belongs to the class of LRWE estimators with $\lambda(h) = \left|\mathbf{X}_h^t \mathbf{X}_h\right|$, $\rho(h) = 1$, and $w[\lambda(h), \rho(h)] = \lambda(h)\rho(h)$. This reveals that the OLS approach only considers the information of leverage but does not take the information of degree of fit for each elemental subset h into account; the resulting estimates can be easily affected by the influential points. Moreover, it can be seen from formula (3) that the WLS estimator is a member of the LRWE class with $\lambda(h) = \left|\mathbf{X}_h^t \mathbf{X}_h\right|$, $\rho(h) = |\mathbf{V}_h|$, and $w[\lambda(h), \rho(h)] = \lambda(h)\rho(h)$. This is because $\mathbf{X}_h$ and $\mathbf{V}_h$ are square matrices and $\left|\mathbf{X}_h^t \mathbf{V}_h \mathbf{X}_h\right| = \left|\mathbf{X}_h^t \mathbf{X}_h\right| \left|\mathbf{V}_h\right|$. This explains why the WLS approach is robust to violations of model assumptions and influential observations because it considers the information of both leverage and degree of fit.

Mayo and Gray (1997) developed a robust TE estimator based on the LRWE class. Unlike the OLS method where the same weight of degree of fit is assigned to all elemental regressions regardless of whether they are influenced by outlying cases, the TE method removes or trims out those elemental regressions that poorly fit the data due to extreme observations from calculations. With $\lambda(h)$ and $\omega[\lambda(h), \rho(h)]$ remaining the same as those in formula (2), the TE estimator alters $\rho(h)$ to have the form

$$\rho(h) = \begin{cases} 1, & \text{if } rank\left(\sum_{i=1}^{n} \left|e_{hi}\right|\right) \le \left(1 - \alpha_p\right)\binom{n}{p} \\ 0, & \text{otherwise} \end{cases}$$

where α_p represents the trimming proportion that ranges from 0 to 1 and $\sum_{i=1}^{n} \left|e_{hi}\right|$ is the sum of absolute residuals based on the elemental estimates $\hat{\boldsymbol{\beta}}_h$. By ruling out those elemental regressions adversely affected by extreme cases, the TE approach produces estimators robust to outliers and violations of model assumptions. Notice that the degree of robustness of the presented approach depends on the values selected for trimming proportion α_p. A bigger α_p means a greater robustness because it removes more elemental regressions with large sums of absolute residuals than a lower α_p does. Depending on the proportion of regressions one would like to remove from consideration, α_p can be adjusted accordingly. Taking this into account, the TE estimator is denoted as TEE(α_p).

The TE approach is different from eliminating outliers from data. Omission of outlying observations takes away multiple elemental subsets including some good ones that could potentially exist with those observations. For example, if a dataset contains 10 observations and 2 unknown parameters are of interest, there are $\binom{10}{2} = 45$ elemental regressions total. If one outlier is removed, then the total number of elemental regressions reduces to $\binom{9}{2} = 36$. As you may expect, the number of elemental regressions eliminated from analysis increases dramatically as *n* or *p* becomes bigger. Deleting observations from data is not the best way to handle outliers unless the outlying cases are indeed resulted from mistakes or other extraneous causes.

BCa Bootstrap CI

The BCa approach, suggested by Efron (1987), seeks to correct the bias and skewness of the sampling distribution through adjusting the selected percentiles used for constructing CIs. The adjusted percentiles are

$$\delta_1 = \phi\left(\hat{z} + \frac{\hat{z} + z_{\alpha/2}}{1 - \hat{\alpha}\left(\hat{z} + z_{\alpha/2}\right)}\right) \quad \text{and} \quad \delta_2 = \phi\left(\hat{z} + \frac{\hat{z} + z_{1-\alpha/2}}{1 - \hat{\alpha}\left(\hat{z} + z_{1-\alpha/2}\right)}\right)$$

where $\phi(.)$ is the standard normal cumulative function and z_α represents the $100\alpha\%$ quantile of the standard normal distribution. The skewness and bias of the sampling distribution are respectively adjusted by $\hat{z}$ and $\hat{\alpha}$, expressions of which can be found in Efron (1987) and DiCiccio and Efron (1996). In general, the algorithm for creating the $100(1 - \alpha)\%$ BCa bootstrap CIs in terms of the TE estimation is given as follows:

- For $m = 1,\ldots, M$, do:
  - (a) Sample data with replacement from the dataset.
  - (b) Compute TE estimates $\hat{\boldsymbol{\beta}}_{\text{TEE}}$ based on the m^{th} bootstrap sample.
- Construct the $100(1 - \alpha)\%$ BCa bootstrap CIs using the adjusted percentiles given above based on the generated bootstrap sample of $\hat{\boldsymbol{\beta}}_{\text{TEE}}$

Hall and Mayo (2008) conducted simulation studies under various scenarios to compare the coverage probabilities of BCa bootstrap CIs based on the TE estimation to the ones based on other approaches. It was found that the BCa bootstrap CIs in terms of TE estimators are almost indistinguishable from those based on OLS when error terms follow the Normal, Contaminated Normal, or

Student's *t* distribution. For the Cauchy and Laplace error distributions, however, the TE estimation is preferred (See Hall and Mayo (2008) for more details). This indicates the OLS estimator is robust to small departures from normality; however, major departures from normality should be of concern.

Computation Efficiency

Even with powerful computers available today the computation time for deriving TE estimates increases tremendously as the number of regression parameters or sample size increases. For example, if there are 10 observations and the model only has two parameters, then $\binom{10}{2} = 45$ elemental subsets need to be fit; however, if the sample size and number of parameters increase to 20 and 4, respectively, we need to fit $\binom{20}{4} = 4845$ elemental regressions, which requires over 100 times more computations. In order to reduce the computation intensity, Hall and Mayo (2008) examined the appropriateness of the approach of random subsample, suggested by Hawkins (1993) for the best elemental estimator, for reducing the number of computations required for the TE estimator through simulation studies. They claimed that computing the TE estimates based on as low as 50% of the elemental subsets may be sufficient to produce reliable estimates as long as the error terms follow Normal, Cauchy, Laplace, 10% Contaminated Normal, or Student's *t* distribution.

TEEReg Package

The proposed R package TEEReg provides tools for computing the TE estimates and the corresponding BCa bootstrap CIs. In this section, the usage of the two functions TEE() and TEE.BCa() enclosed in TEEReg are explained.

The function TEE() is used to compute the TE estimates. Its usage with complete arguments is given as:

```
TEE(formula, data, offset=NULL, p.trimmed=NULL, p.subsample=1,
method="tee")
```

Similar to other R functions developed for linear regressions, such as lm() and glm(), the first argument formula gives a symbolic description of the model to be fitted (e.g. formula = $y \sim x$). The second argument specifies the dataset used for performing regression analyses. Be aware that the data must be formatted as a data frame prior to using the TEE() function. The offset can be used to specify regressors with coefficients of 1. This argument can be either NULL or a numeric vector with length equal to the number of observations. The argument p.trimmed indicates the proportion of elemental subsets removed from the computation of estimates. It should be either NULL or a numeric value between 0 and 1. However, a value must be provided to p.trimmed when method = "tee" is specified. The argument p.subsample is for specifying the proportion of random selection of elemental subsets. One may improve the computation efficiency by providing a numeric value between 0 and 1 to this argument. The default value of p.subsample is 1 under which the TE estimates are calculated based on all elemental subsets. When using the TEE() function, the TE regression is carried out by default (i.e., the default value to argument method is "tee"). Another supported option for this argument is "ols" under which the OLS approach is employed for fitting linear regressions. When the value ols is given to the argument of method, the TEE() function computes the estimates based on the full data no matter what values are assigned to p.trimmed and p.subsample.

The second function TEE.BCa() is used to construct the $100(1-\alpha)\%$ BCa bootstrap CIs based on the TE estimation. It is similar in structure to TEE() and has the form with complete arguments as follows:

```
TEE.BCa(formula, data, offset=NULL, p.trimmed=NULL, p.subsample=1,
method="tee", est.TEE, conf.level, n.boot)
```

The specifications of the first six arguments in TEE.BCa() are the same as explained above for TEE(). For the remaining three, est.TEE stands for TE regression estimates, and conf.level and n.boot represent the confidence level and the number of bootstrap samples, respectively. Detailed descriptions of the arguments enclosed in these two functions can also be viewed using the command ??TEE.

Sometimes, the elemental regression $\hat{\boldsymbol{\beta}}_h$ is not estimable because $\mathbf{X}_h$ is singular and the inverse matrix $\mathbf{X}_h^{-1}$ does not exist. This could happen, for example, when several subjects have the same covariates values and so the matrix $\mathbf{X}_\text{h}$ is not full-rank. The TEEReg package handles such situations using the Moore-Penrose generalized inverse, which is defined and unique for all matrices whose entries are real or complex numbers. It is computed using the singular value decomposition. For a review of the Moore-Penrose generalized inverse, see Campbell and Meyer (2009).

Examples

To evaluate the robustness of the presented TE approach, the first example is based on the telephone data (Rousseeuw & Leroy, 1987) with several outlying observations and the second example is simulated data based on a Cauchy distribution. For both examples, the 95% BCa bootstrap CIs are created based on 1000 bootstrap samples.

Example 1: Data with Outliers

In this example, the telephone data (Rousseeuw & Leroy, 1987) are used to demonstrate the usage of the TEEReg package. In the data, the number of telephone calls (tens of millions) made in Belgium was recorded from 1959 to 1973. It contains several extreme observations resulted from mistakes in recording units over the years 1964-1969 (see Figure 1), which is useful in order to examine the robustness of the TE method to outliers. The response variable of the telephone data is the number of telephone calls and the independent variable is the year. For illustration purposes, the TE estimates and the corresponding 95% BCa bootstrap CIs are computed based on both 30% and 42% trimming proportions. The results in terms of all elemental subsets and those based on 70% random subsample are also compared in this example.

The TEEReg package can be loaded into R by the command library(TEEReg). The telephone data are stored inside the package and can be accessed by the command data(telephone). As explained above, the TE estimates and the corresponding 95% BCa bootstrap CIs in terms of the subsample proportion of 100% and trimming proportion of 42% can be computed by typing the following:

```
R> fitTEE1 <- TEE(formula=Y~X, data=telephone, p.trimmed=0.42,
p.subsample=1, method="tee")
R> CITEE1 <- TEE.BCa(formula=Y~X, data=telephone, p.trimmed=0.42,
p.subsample=1, + method="tee", est.TEE=fitTEE1$coefficients,
conf.level=0.05, n.boot=1000)
```

Their outputs are displayed as below:

```
R> fitTEE1
$call
TEE(formula = Y ~ X, data = telephone, p.trimmed = 0.42, p.subsample = 1,
method = "tee")
$formula
Y ~ X
$coefficients
(Intercept)          X
-100.0543   1.991974
$residuals
        1          2          3          4          5          6          7
 4.855597   3.163623   1.171649  0.3796743    -0.9123  -2.204274  -3.396248
        8          9         10         11         12         13         14
-4.688223  -4.880197  -5.472171  -5.964145   -6.55612  -7.348094  -4.240068
       15         16         17         18         19         20         21
 91.56796   94.57598    110.584    125.592   146.6001   174.6081   3.616112
       22         23         24
-17.37586  -16.36784  -16.35981
$fitted.values
        1         2          3         4         5         6         7
-0.455597  1.536377   3.528351  5.520326    7.5123  9.504274  11.49625
        8         9         10        11        12        13        14
 13.48822   15.4802   17.47217  19.46415  21.45612  23.44809  25.44007
       15        16         17        18        19        20        21
 27.43204  29.42402   31.41599  33.40797  35.39994  37.39191  39.38389
       22        23        24
 41.37586  43.36784  45.35981

R> CITEE1
$call
TEE.BCa(formula = Y ~ X, data = telephone, p.trimmed = 0.42, p.subsample = 1,
method  =  "tee",  est.TEE  =  fitTEE1$coefficients,  conf.level  =  0.05,  n.boot  =
1000)
$ci
             estimates(TEE)  Lower limit  Upper limit
(Intercept)       -100.0543  -452.481442   -49.220453
X                  1.991974     1.045627     8.588198
```

Note the output yielded by the function TEE() contains the model formula, estimates of coefficients, residuals, and fitted values, and the output of the TEE.BCa() function consists of the model formula and BCa bootstrap CIs for regression parameters. In the case that one only wants to extract, for example, the coefficient estimates from the output of TEE() function, the command fit1$coefficients can be used. The TE estimates and the corresponding 95% BCa bootstrap CIs based on other scenarios planned to be investigated in this example can be computed following a similar manner by specifying p.trimmed = 0.30 and p.subsample = 1 or 0.7. The key results are summarized in Table 1. For comparison purposes, the results based on the OLS approach and the IRLS using Huber weight function are also presented in this table.

The estimated regression function using the TE approach with p.subsample = 1 and 42% trimming suggests that the mean number of telephone calls are expected to increase by 1.992 (in tens of millions) when the year increases by 1. The corresponding 95% BCa bootstrap CI for the slope is (1.046, 8.588) which does not include 0. Based on this scenario, it can be concluded that year is significantly linearly related to the number of telephone calls. As expected, the outlying observations are more influential in the fitted TE regression function with p.subsample = 1 and 30% trimming proportion. The estimated slope is dragged up by outliers to 3.940 (BCa CI: 1.114, 8.424) due to the fact that more elemental regressions with large sums of absolute residuals are used in calculations. The same trend can be observed in the case of p.subsample = 0.7.

Moreover, it can be seen in Table 1 that the TE estimates based on 70% random subsample of elemental subsets are similar to those based on all elemental subsets for both cases of TEE(30%) and TEE(42%). The 95% BCa bootstrap CIs in terms of 70% subsample are wider than the ones based on all elemental subsets, but both lead to the same conclusion of statistical inference. It seems that using the 70% subsampling provides fairly accurate estimates and works almost equally well as utilizing the full data for the given telephone data.

Table 1. Estimates of coefficients and 95% BCa bootstrap Cis based on various approaches using telephone data

Methods	Intercept est.	95% CI (intercept)	Slope est.	95% CI (slope)
TEE(30%): p.subsample = 1	-204.034	(-452.688, -52.983)	3.940	(1.114, 8.424)
TEE(30%): p.subsample = 0.7	-217.143	(-516.187, -54.649)	4.193	(1.145, 9.520)
TEE(42%): p.subsample = 1	-100.054	(-452.481, -49.220)	1.992	(1.046, 8.588)
TEE(42%): p.subsample = 0.7	-112.678	(-540.452, -50.289)	2.235	(1.062, 10.069)
OLS	-260.059	(-523.136, -118.906)	5.041	(2.475, 9.549)
IRLS	-99.904	(-590.294, -52.987)	1.987	(1.113, 10.873)

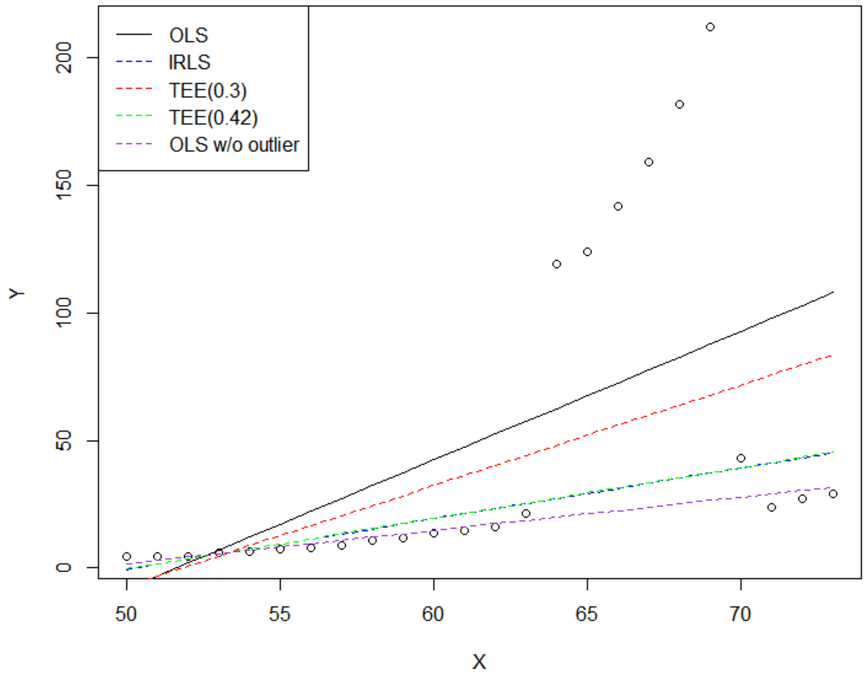

Figure 1. Fitted regression lines using different regression approaches for telephone data

Figure 1 displays the fitted regression lines for a variety of regression approaches. The overlaid TE regression lines are obtained in terms of all elemental subsets (i.e., p.subsample = 1). In addition, a regression line fitted using the OLS approach based on the telephone data with outliers removed is also included in this figure for comparison purposes. It is obvious that the OLS approach performs the worst with its estimates dramatically affected by outliers. The regression lines based on IRLS and TEE(42%) are overlapped with each other because they lead to almost identical estimates of unknown parameters (see Table 1). This is not surprising because the IRLS approach is also robust to outlying cases. The 95% BCa bootstrap CIs for IRLS are wider than the ones for TEE(42%) (see Table 1). As explained in the previous paragraph, due to the fact that relatively more elemental regressions having large sums of absolute residuals are employed in calculations, the TEE(30%) is affected more by the outliers than the TEE(42%) and IRLS. Both fitted regression lines of TEE(30%) and TEE(42%) are above the one based on the OLS approach with outliers removed. The reason is that deleting outlying observations takes away all of their corresponding elemental subsets.

Example 2: Cauchy Data

In this example, a simulated dataset consisting of 50 observations and one independent variable is used to clarify the usage of TEEReg package and to illustrate the robustness to non-normal data of the presented TE estimator. The values of the independent variable X are generated from a Poisson distribution with mean equal to 10 and the values of the dependent variable Y are computed as $Y = 0.5 + 1X + \epsilon$, where the error term ϵ is assumed to follow a Cauchy distribution with location 0 and scale 1. We call this artificial dataset the data.sim. In this example, the TE estimates and the corresponding 95% BCa bootstrap CIs are computed based on all elemental subsets and both 50% and 75% trimming proportions. As demonstrated in Hall and Mayo (2008), these two trimming proportions provide high coverage probabilities (at least 95%) to the 95% BCa bootstrap CIs when the error term follows Cauchy distribution.

The TE estimates and the corresponding 95% BCa bootstrap CIs in terms of the subsample proportion of 100% and trimming proportion of 50% can be computed by typing the following:

```
R> fitTEE3 <- TEE(formula=Y~X, data=data.sim, p.trimmed=0.5,
p.subsample=1,method = "tee")
R> CITEE3 <- TEE.BCa(formula=Y~X, data=data.sim, p.trimmed=0.5,
p.subsample=1, + method="tee", est.TEE=fitTEE3$coefficients,
conf.level=0.05, n.boot=1000)
```

The TE estimates and their BCa CIs based on 75% trimming can be computed similarly by specifying p.trimmed = 0.75. The key outputs of both scenarios are summarized in Table 2. For comparison purposes, the results based on the OLS method and the IRLS using Huber weight function are also given in this table.

Table 2. Estimates of coefficients and 95% BCa bootstrap Cis based on various regression approaches using simulated data

Methods	Intercept est.	95% CI (intercept)	Slope est.	95% CI (slope)
TEE(50%)	1.341	(-0.542 , 6.602)	0.899	(0.305, 1.120)
TEE(75%)	0.919	(-1.026, 3.734)	0.967	(0.634, 1.170)
OLS	6.639	(1.858, 12.516)	0.471	(-0.096, 0.934)
IRLS	2.100	(0.0728, 7.281)	0.832	(0.240, 1.055)

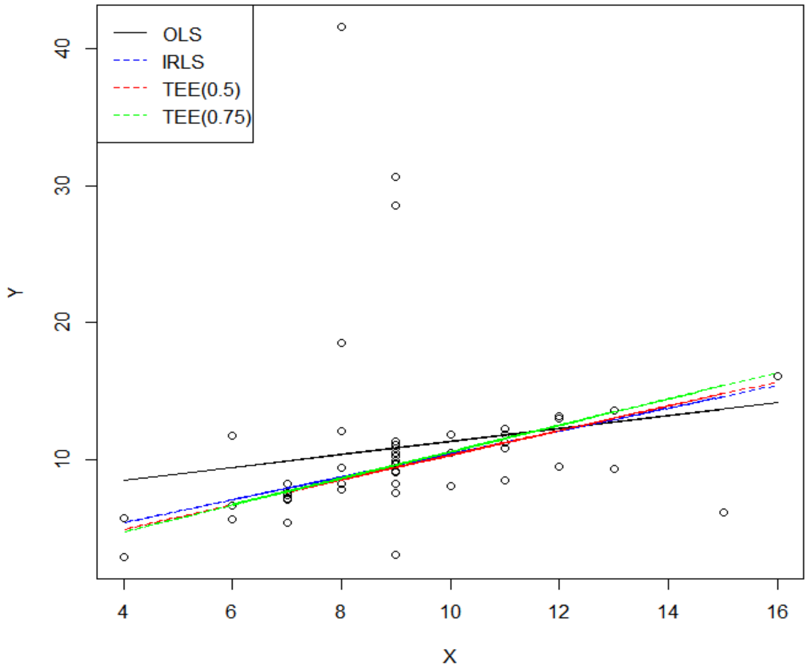

Figure 2. Fitted regression lines using different regression approaches for simulated data

As expected, the OLS approach performs the worst in terms of handling the simulated Cauchy data. The corresponding 95% BCa bootstrap CIs for intercept and slope are, respectively, (1.858, 12.516) and (-0.096, 0.934), none of which captures the true values of 0.5 and 1. The OLS estimates of both intercept and slope are significantly different from the true values as well. In contrast, it appears that the TEE(75%) performs the best for the given dataset. The resulting TE estimates for slope and intercept are, respectively, 0.919 and 0.967, both of which are very close to the true intercept and slope used for generating data. The estimates produced by TEE(50%) seems to be slightly worse than ones based on TEE(75%), but it is closer to the true values than the ones resulting from IRLS. The 95% BCa bootstrap CIs of both TEE(50%) and IRLS contain the true intercept and slope of 0.5 and 1. It appears that the TE approach is robust to the simulated Cauchy data that severely depart from normality. A scatterplot of the simulated data along with fitted regression lines using different approaches is shown in Figure 2.

Summary

The usage of a new R package TEEReg was explicated for computing the TE estimates and creating the BCa bootstrap CIs. This package includes two functions: TEE() for the TE regression and TEE.BCa() for the BCa bootstrap CIs. Two examples were provided in this paper to demonstrate the usage of the TEEReg package. In the first example, the telephone data with several influential observations were used to examine the robustness of the TE method to outliers. It was found that the TEE(42%) and IRLS approaches work equally well for the given dataset. The TEE(30%) was affected more by the outliers because, compared to $\alpha_p = 42\%$, relatively more elemental regressions with large sums of absolute residuals are involved in calculations. The random subsample approach, suggested by Hawkins (1993), was employed in this example as well. It appeared that, for the telephone dataset, using the 70% subsampling provides fairly accurate estimates and works almost equally well as utilizing the full data. This is consistent with the conclusions of Hall and Mayo (2008), that the random subsample approach is appropriate for reducing computation intensity when the error terms follow certain distributions. In the second example, a simulated data set with Cauchy error terms was used to assess the robustness of the TE approach to non-normal data. It appeared that the TE estimator is robust and efficient to the simulated data with Cauchy error terms. This is also consistent with the findings based on simulation studies from Hall and Mayo (2008). The new TEEReg package can be readily used to conduct TE regression analysis which is a useful and robust alternative to OLS in the presence of outliers and violations of model assumptions.

References

Campbell, S. L., & Meyer, C. D. (2009). *Generalized inverses of linear transformations*. Philadelphia, PA: Society for Industrial and Applied Mathematics. doi: 10.1137/1.9780898719048

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York, NY: Chapman and Hall.

DiCiccio, T., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189-212. Available from: http://www.jstor.org/stable/2246110

Dielman, T. E. (2005). Least absolute value regression: Recent contributions. *Journal of Statistical Computation and Simulation, 75*(4), 263-286. doi: 10.1080/0094965042000223680

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171-185. doi: 10.2307/2289144

Gutenbrunner, C., & Jureckova, J. (1992). Regression rank scores and regression quantiles. *The Annals of Statistics, 20*(1), 305-330. doi: 10.1214/aos/1176348524

Hall, M., & Mayo, M. S. (2008). Bootstrap confidence intervals and coverage probabilities of regression parameter estimates using trimmed elemental estimation. *Journal of Modern Applied Statistical Methods, 7*(2), 514-525. Retrieved from: http://digitalcommons.wayne.edu/jmasm/vol7/iss2/17/

Hawkins, D. M. (1993). The accuracy of elemental set approximations for regression. *Journal of the American Statistical Association, 88*(422), 580-589. doi: 10.2307/2290339

Koenker, R. W. (1994). Confidence intervals for regression quantiles. In P. Mandl, M. Hušková (Eds.), *Asymptotic statistics* (pp. 349-359). New York, NY: Springer-Verlag. doi: 10.1007/978-3-642-57984-4_29

Koenker, R. W. (2005). *Quantile regression*. New York, NY: Cambridge University Press. doi: 10.1017/cbo9780511754098

Koenker, R., & Bassett, G. J. (1978). Regression quantiles. *Econometrica, 46*(1), 33-50. doi: 10.2307/1913643

Koenker, R. W., & D'Orey, V. (1987). Algorithm AS 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 36*(3), 383-393. doi: 10.2307/2347802

Krasker, W. S., & Welsch, R. E. (1983). The use of bounded-influence regression in data analysis: theory, computation, and graphics. *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*. New York, NY: Springer-Verlag.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw-Hill.

Mayo, M. S., & Gray, J. B. (1997). Elemental subsets: The building blocks of regression. *The American Statistician, 51*(2), 122-129. doi: 10.2307/2685402

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871-880. doi: 10.2307/2288718

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: John Wiley and Sons. doi: 10.1002/0471725382

Schlossmacher, E. J. (1973). An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association*, *68*(344), 857-859. doi: 10.2307/2284512

Sposito, V. A., Kennedy, W. J., & Gentle, J. E. (1977). Algorithm AS 110: L_p norm fit of a straight line. *Journal of the Royal Statistical Society*. *Series C (Applied Statistics)*, *26*(1), 114-118. doi: 10.2307/2346888

Street, J. O., Carroll, R. J., & Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, *42*(2), 152-154. doi: 10.2307/2684491