Package 'surveyoutliers'

December 21, 2015

Type Package

Title Helps Manage Outliers in Sample Surveys

Version 0.0

Date 2015-12-14

Author Robert Clark

Maintainer ``Robert Clark" <rclark@uow.edu.au>

Description At present, the only functionality is the calculation of optimal one-sided winsorizing cutoffs. The main function is optimal.onesided.cutoff.bygroup. It calculates the optimal tuning parameter for one-sided winsorisation, and so calculates winsorised values for a variable of interest. See the help file for this function for more details and an example.

License GPL-2 | GPL-3

LazyData TRUE

RoxygenNote 5.0.1

R topics documented:

`optimal.onesided.cutoff`

Optimal one-sided winsorization for survey outliers

Description

This function calculates optimal tuning parameter, cutoffs, and winsorized values for one-sided winsorization.

Usage

```
optimal.onesided.cutoff(formula, surveydata, historical.reweight = 1,
  estimated.means.name = "", stop = F)
```

Arguments

formula	The regression formula (e.g. income ~ employment + old.turnover if income is survey variable and employment and old.turnover are auxiliary variables).
surveydata	A data frame of the survey data including the variables in formula, piwt (inverse probability of selection), gregwt (generalized regression estimator weight) and regwt (weight to be used in regression - will be set to 1 if missing).
historical.reweight	A set of reweighting factors for use when a historical dataset is being used. It reweights from the historical sample to the sample of interest. The default value of 1 should be used if the sample being used for optimising Q is the same sample (or at least the same design) as the sample to which the winsorizing cutoffs are to be applied.
estimated.means.name	The variable of this name in surveydata should contain an estimator of the expected values for each sample value of the variable of interest. If set to "", the regression model is estimated using IRLS.
stop	Set to T to open a browser window (for debugging purposes)

Details

This function calculates optimal one-sided cutoffs for winsorization where regression residuals are truncated at Q / (generalized_regression_estimator_weight-1) and Q satisfies the optimality result in Kokic and Bell (1994) and Clark (1995).

Value

A list consisting of Q.opt (the optimal Q), rlm.coef (the robust regression coefficients), windata which is a dataset containing the same observations and variables as surveydata in the same order, with additional variables cutoffs (the winsorizing cutoffs for each unit in sample), y (the values of the variable of interest), win1.values (the type 1 winsorized values of interest, i.e. the minimums of the cutoff and y) and win2.values (the type 2 winsorized values of interest, so that sum(surveydata$gregwt*win2.values) is the winsorized estimator.

References

Clark, R. G. (1995), "Winsorisation methods in sample surveys," Masters thesis, Australian National University, http://hdl.handle.net/10440/1031.

Kokic, P. and Bell, P. (1994), "Optimal winsorizing cutoffs for a stratified finite population estimator," J. Off. Stat., 10, 419-435.

Examples

```
test <- optimal.onesided.cutoff(formula=y~x1+x2,surveydata=survdat.example)
plot(test$windata$y,test$windata$win1.values)
```

`optimal.onesided.cutoff.bygroup` *Optimal one-sided winsorization for survey outliers by group*

Description

This function calculates optimal tuning parameter, cutoffs, and winsorized values for one-sided winsorization, by group.

Usage

```
optimal.onesided.cutoff.bygroup(formula, surveydata, historical.reweight = 1,
  groupname, estimated.means.name = "", stop = F)
```

Arguments

`formula`	The regression formula (e.g. income ~ employment + old.turnover if income is survey variable and employment and old.turnover are auxiliary variables).
`surveydata`	A data frame of the survey data including the variables in formula, piwt (inverse probability of selection) and gregwt (generalized regression estimator weight).
`historical.reweight`	A set of reweighting factors for use when a historical dataset is being used. It reweights from the historical sample to the sample of interest. The default value of 1 should be used if the sample being used for optimising Q is the same sample (or at least the same design) as the sample to which the winsorizing cutoffs are to be applied.
`groupname`	The variable of this name in surveydata defines the groups for which Q is to be optimised. If groupname is missing, it is assumed that cutoffs are to be optimised for the overall mean or total.
`estimated.means.name`	The variable of this name in surveydata should contain an estimator of the expected values for each sample value of the variable of interest. If set to "", the regression model is estimated using IRLS.
`stop`	Set to T to open a browser window (for debugging purposes)

Details

This function calculates optimal one-sided cutoffs for winsorization where regression residuals are truncated at Q / (weight-1) and Q satisfies the optimality result in Kokic and Bell (1994) and Clark (1995).

Value

A list consisting of Q.opt (the optimal Q), rlm.coef (the robust regression coefficients), cutoffs (the winsorizing cutoffs for each unit in sample), y (the values of the variable of interest), win1.values (the type 1 winsorized values of interest, i.e. the minimums of the cutoff and y) win2.values (the type 2 winsorized values of interest, so that sum(surveydata$gregwt*win2.values) is the winsorized estimator)

References

Clark, R. G. (1995), "Winsorisation methods in sample surveys," Masters thesis, Australian National University, http://hdl.handle.net/10440/1031.

Kokic, P. and Bell, P. (1994), "Optimal winsorizing cutoffs for a stratified finite population estimator," J. Off. Stat., 10, 419-435.

Examples

```
test <- optimal.onesided.cutoff.bygroup(formula=y~x1+x2,
surveydata=survdat.example,groupname="industry")
plot(test$windata$y,test$windata$win1.values)
```

`robust.lm.onesided` *Robust Regression using One-Sided Huber Function*

Description

This function performs robust regression using M-estimation using the one-sided Huber function, with residuals truncated at Q / (data$gregwt-1) where data$gregwt is the generalized regression weight.

Usage

```
robust.lm.onesided(formula, data, Q, Qname, maxit = 100, stop = F)
```

Arguments

`formula`	The regression formula (e.g. income ~ employment + old.turnover if income is survey variable and employment and old.turnover are auxiliary variables).
`data`	A data frame including the variables in formula, and gregwt (generalized regression estimator weight), and regwt (weight to be used in regression - will be set to 1 if missing).
`Q`	The tuning parameter where large Q corresponds to no outlier treatment, and small Q corresponds to many outliers being flagged.
`Qname`	Gives a variable name on data which contains a separate tuning parameter Q for every observation (either Q or Qname should be specified but not both).
`maxit`	The maximum number of iterations.
`stop`	Set to T to open a browser window (for debugging purposes)

Details

Uses iteratively reweighted least squares.

Value

The final linear model fit (an object of class "lm").

References

Clark, R. G. (1995), "Winsorisation methods in sample surveys," Masters thesis, Australian National University, http://hdl.handle.net/10440/1031.

Kokic, P. and Bell, P. (1994), "Optimal winsorizing cutoffs for a stratified finite population estimator," J. Off. Stat., 10, 419-435.

Examples

```
robust.lm.onesided(formula=y~x1+x2,data=survdat.example,Q=250)
```

`survdat.example` *An example sample data file of 5000 respondents.*

Description

A dataset containing the values of y (a variable of interest which is outlier-prone), two auxiliary variables x1 and x2, piwt (inverse probability of selection weight), and gregwt (generalized regression estimation weight, assumed to have been calculated using x1 and x2)

Usage

```
survdat.example
```

Format

An object of class `data.frame` with 500 rows and 6 columns.

Index