Package 'SAEval'

May 16, 2022

Type Package

Title Small Area Estimation Evaluation

Description Allows users to produce diagnostic procedures and graphic tools for the evaluation of Small Area estimators.

Version 0.1.5

Depends R (>= 3.5.0), lmtest , car

NeedsCompilation no

Author Andrea Fasulo [aut, cre]

Maintainer Andrea Fasulo <fasulo@istat.it>

Imports stats,ggplot2,ggspatial,grid

License EUPL

Encoding UTF-8

BuildResaveData best

Repository CRAN

Date/Publication 2022-05-16 13:10:02 UTC

R topics documented:

SAEval-package *The R SAEval Package*

Description

SAEval is an R package for diagnostic analysis of Small Area Estimation (SAE). It provide a set of tools for the evaluation of SAE with respect to the direct estimates.

Details

Working with SAE it is good practice to compare different estimators to find the one with the best performance. This package contains functions for statistical calculation of diagnostic procedure aimed at evaluate the quality of the SAE. In detail, in the package are developed some methods originally proposed in Brown et al (2001) to check the quality of SAE.

Furthermore is possible to produce graphical tools that map the chosen indicator for a spatial analysis.

For a complete list of functions, use `library(help = "SAEval")`.

Author(s)

Developed by Andrea Fasulo

bias *Bias diagnostic*

Description

`bias` diagnostic allows to evaluate how the model-based estimates are closed to the unbiased direct estimates.

Usage

```
bias(data,dir,sae,scatterplot=FALSE,main=NULL)
```

Arguments

`data`	a data frame containing the direct estimates among with the small area estimates, e.g. `SAEval_example`.
`dir`	formula identifing the direct estimates.
`sae`	formula identifing the small area estimates.
`scatterplot`	logical scalar. Should the scatterplot of the estimates be produced (default=FALSE)?. See also 'Details'.
`main`	optionally, if a string is set in `main` it will be used as the scatterplot main title. The defualt main title is the name of the direct estimator versus the SAE names.

Details

bias tests whether the model based estimates are closed to the direct estimates. A parametric test for the slope and for the intercept is carried out to check the unbiasedness of the model predictions. A square-root of the estimates is required when the homoskedasticity assumption underpinning the OLS fitting method is not satisfied. The Goldfeld and Quandt homoscedasticity test is provided, to check such constant variances.

The use of this diagnostic is straightforward when the focus of interest is on small area totals since unbiased direct estimators of such totals are typically available.

If scatterplot=TRUE the SAE estimates (X-axis) are plotted on a cartesian plane against the direct estimates (Y-axis) to verify if there is a departure of the Y = X (red line) from the regression line between model based and direct estimates (black line).

The small area with direct estimate equal to NA value are automatically removed from the data.

Value

Object of class list. The list contains up to 2 objects:

output1	a data frame containing for the small area estimates of interest (methods), the intercept (b0), the slope (b1) and the R-squared (R2) values among with the F-test (F) and Goldfeld and Quandt test (GQ_Test).
output2	a data frame containing for the trasformed small area estimates of interest (methods) the intercept (b0), the slope (b1)and the R-squared (R2) values among with F-test (F) and Goldfeld and Quandt test (GQ_Test).

Author(s)

Developed by Andrea Fasulo

References

Brown, G., Chambers, R., Heady, P., Heasman, D. (2001), Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS, in Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada.

Mukhopadhyay, P. K., McDowell, A. (2011). Small area estimation for survey data analysis using SAS software, http://support.sas.com/rnd/app/papers/smallarea.pdf.

Srivastava, A. K., Sud, U. C., Chandra, H. (2007). Small area estimation - An application to National Sample Survey Data, Journal of the Indian Society of Agricultural Statistics, 61(2), 249-254.

Examples

```
# Load example data
data(SAEval_example)

SAEval.bias<-bias(data=SAEval_example,
dir=~y_d,
sae = ~y_syna+y_eblupa+y_spaznr+y_eblupb+y_synb+y_logis)
```

SAEval.bias

---

calibration *Calibration diagnostic*

Description

calibration diagnostic refers to the calibration property of model estimates, according to which they should not differ from the direct estimates when aggregated at appropriate large domain levels. Computing this diagnostic we obtain an accurate measure of the calibration property of the model estimates, providing also an evidence of the presence/absence of spatial bias/autocorrelation.

Usage

```
calibration(data,dir,sae,area)
```

Arguments

data	a data frame containing the direct and small area estimates among with their variance, e.g. SAEval_example.
dir	formula identifing the direct estimates.
sae	formula identifing the small area estimates.
area	formula identifing the area for which the calibration diagnostic is computed.

Details

calibration compute the relative difference between the aggregated model-based estimates and the aggregated direct estimates.

The small area with both direct estimate and variance of the direct estimates equal to NA value are automatically removed from the data.

Value

Object of class list. The list contains objects equal to the number of larger domain specified in area. Each object will contains the calibration diagnostic for all the modes of the area.

Author(s)

Developed by Andrea Fasulo

References

Brown, G., Chambers, R., Heady, P., Heasman, D. (2001), Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS, in Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada.

Mukhopadhyay, P. K., McDowell, A. (2011). Small area estimation for survey data analysis using SAS software, http://support.sas.com/rnd/app/papers/smallarea.pdf.

Srivastava, A. K., Sud, U. C., Chandra, H. (2007). Small area estimation - An application to National Sample Survey Data, Journal of the Indian Society of Agricultural Statistics, 61(2), 249-254.

Examples

```
# Load example data
data(SAEval_example)

SAEval.calibration<-calibration(data=SAEval_example,
       dir=~y_d,
       sae=~y_syna+y_eblupa+y_spaznr+y_eblupb+y_synb+y_logis,area=~nuts0+nuts1+nuts2)

SAEval.calibration
```

`cinterval` *Confident interval analysis*

Description

`cinterval` analize the SAE estimates with respect to the confident inteval of the direct estimate.

Usage

```
cinterval(data,dir,sae,v.dir,mse.sae,level=0.95,plot=F)
```

Arguments

`data`	a data frame containing the direct and small area estimates among with their variance, e.g. `SAEval_example`.
`dir`	formula identifing the direct estimates.
`sae`	formula identifing the small area estimates.
`v.dir`	formula identifing the direct estimates variance.
`mse.sae`	formula identifing the small area estimates mean squared error.
`level`	double number. The confidence level represents the proportion of correspondingly confedent inteval that end up containing the true value of the parameter (default=0.95).
`plot`	logical scalar. Should the plot be produced (default=FALSE)?. See also 'Details'.

Details

This diagnostic measures for each SAE estimators the number of estimates that fall between the upper and lower bound of the direct estimates confidence intervals.

If `plot=TRUE` the SAE estimates are plotted with the direct estimates confindent interval to analyze the distributions.

The small area with both direct estimate and variance of the direct estimates equal to NA value are automatically removed from the data.

Value

Object of class `data.frame`. The data frame contains information for the small area estimators (`methods`) about the number of SAE estimates included in the confident interval (`included`) and the number of overlapping confident intervals (`overlap`).

Author(s)

Developed by Andrea Fasulo

References

Brown, G., Chambers, R., Heady, P., Heasman, D. (2001), Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS, in Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada.

Mukhopadhyay, P. K., McDowell, A. (2011). Small area estimation for survey data analysis using SAS software, http://support.sas.com/rnd/app/papers/smallarea.pdf.

Srivastava, A. K., Sud, U. C., Chandra, H. (2007). Small area estimation - An application to National Sample Survey Data, Journal of the Indian Society of Agricultural Statistics, 61(2), 249-254.

Examples

```
# Load example data
data(SAEval_example)

SAEval.cinterval<-cinterval(data=SAEval_example,
       dir=~y_d,
       sae=~y_syna+y_eblupa+y_spaznr+y_eblupb+y_synb+y_logis,
       v.dir=~mse_d,
       mse.sae=~mse_sa+mse_eba2+mse_spaznr+mse_ebb+mse_sb+mse_log)

SAEval.cinterval
```

coverage *Coverage diagnostic*

Description

coverage diagnostic tests the validity between the 95% adjusted confidence intervals of the model based estimates making comparison with the corresponding adjusted confidence intervals for the direct estimates.

Usage

```
coverage(data,dir,sae,v.dir,mse.sae,alfa=0.05)
```

Arguments

data	a data frame containing the direct and small area estimates among with their variance, e.g. SAEval_example.
dir	formula identifing the direct estimates.
sae	formula identifing the small area estimates.
v.dir	formula identifing the direct estimates variance.
mse.sae	formula identifing the small area estimates mean squared error.
alfa	double number. The significance level of the non-parametric Binomial test (default=0.05).

Details

This diagnostic measures the overlap between the confidence intervals, which is expected to be not significantly different from the 95% of the numbers of small areas.

The small area with both direct estimate and variance of the direct estimates equal to NA value are automatically removed from the data.

Value

Object of class data.frame. The data frame contains information for the small area estimators (methods), non-coverage total (non_coverage), number of small area domains (domains), non-overlap ratio (non_overlap), p-value for Binomial statistic (p_value) and the test result (results).

Author(s)

Developed by Andrea Fasulo

References

Brown, G., Chambers, R., Heady, P., Heasman, D. (2001), Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS, in Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada.

Mukhopadhyay, P. K., McDowell, A. (2011). Small area estimation for survey data analysis using SAS software, http://support.sas.com/rnd/app/papers/smallarea.pdf.

Srivastava, A. K., Sud, U. C., Chandra, H. (2007). Small area estimation - An application to National Sample Survey Data, Journal of the Indian Society of Agricultural Statistics, 61(2), 249-254.

Examples

```
# Load example data
data(SAEval_example)

SAEval.coverage<-coverage(data=SAEval_example,
       dir=~y_d,
       sae=~y_syna+y_eblupa+y_spaznr+y_eblupb+y_synb+y_logis,
       v.dir=~mse_d,
       mse.sae=~mse_sa+mse_eba2+mse_spaznr+mse_ebb+mse_sb+mse_log)

SAEval.coverage
```

---

cv_table *Coefficent of variation's table*

Description

`cv_table` is used to analyse the coefficent of variation distribution of the chosen indicators.

Usage

```
cv_table(data,cv,boxplot=FALSE)
```

Arguments

`data`	a data frame containg the coefficient of variation for the direct and small area estimators
`cv`	formula identifing the coefficient of variation.
`boxplot`	logical scalar. Should the boxplot of the coefficient of variation be produced (default=FALSE)?.

Details

cv_table allows to evaluate the cv of the different estimators with respect to some well-known thresholds given by Statistics Canada (2009). For cv below 0.165 there are no rescrictions to the dissemination, for cv in the range 0.166-0.333 is suggested a publication with a warning, for cv above 0.333 the dissemination is not recommendent.

Value

Object of class data.frame. The data frame contains informations about the number of cvs that fall within each class.

Author(s)

Developed by Andrea Fasulo

References

Statistics Canada, 2009, "Quality Guideline", Fifth edition, October 2009

Examples

```
# Load example data
data(SAEval_example)

# cv for the direct estimates
SAEval_example$cvd<-sqrt(SAEval_example$mse_d)/SAEval_example$y_d
#cv for the synthetic estimates
SAEval_example$cvsae<-sqrt(SAEval_example$mse_sa)/SAEval_example$y_syna

cv_data<-SAEval_example[,c("cvd","cvsae")]

SAEval_cvtable<-cv_table(data=cv_data,
cv=~cvd+cvsae)

SAEval_cvtable
```

---

gof *Goodness of fit diagnostic*

Description

The goodness of fit diagnostic allows to evaluate how close the model-based estimates are to the direct estimates when they are good.

Usage

```
gof(data,dir,sae,v.dir,mse.sae,alfa=0.05)
```

Arguments

data	a data frame containing the direct and small area estimates among with their variance, e.g. SAEval_example.
dir	formula identifing the direct estimates.
sae	formula identifing the small area estimates.
v.dir	formula identifing the direct estimates variance.
mse.sae	formula identifing the small area estimates mean squared error.
alfa	double number. The significance level of the Chi-squared test (default=0.05).

Details

As in the bias diagnostic, even with this procedure we want to know if the model estimates are close to the direct estimates. To evaluate this we compute the squared difference between the model estimates and the direct estimate which are weighted inversely by their variance and summed over all the domains. As a check for the lack of bias of the model estimates this statistic is compared with the quantiles of Chi-squared distribution. Finally results are provided using a Wald goodness of fit statistic.

The small area with both direct estimate and variance of the direct estimates equal to NA value are automatically removed from the data.

Value

Object of class data.frame. The data frame contains information for the small area estimators (methods), Wald statistic (W), Chi-squared statistic (c2), p-value for Wald statistic (p_value) and the test result (results).

Author(s)

Developed by Andrea Fasulo

References

Brown, G., Chambers, R., Heady, P., Heasman, D. (2001), Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS, in Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada.

Mukhopadhyay, P. K., McDowell, A. (2011). Small area estimation for survey data analysis using SAS software, http://support.sas.com/rnd/app/papers/smallarea.pdf.

Srivastava, A. K., Sud, U. C., Chandra, H. (2007). Small area estimation - An application to National Sample Survey Data, Journal of the Indian Society of Agricultural Statistics, 61(2), 249-254.

Examples

```
# Load example data
data(SAEval_example)

SAEval.gof<-gof(data=SAEval_example,
```

```
        dir=~y_d,
        sae=~y_syna+y_eblupa+y_spaznr+y_eblupb+y_synb+y_logis,
        v.dir=~mse_d,
        mse.sae=~mse_sa+mse_eba2+mse_spaznr+mse_ebb+mse_sb+mse_log)

SAEval.gof
```

`map_sae` *Map the disagregated estimates and the coefficients of variation.*

Description

`map_sae` produces geographical maps for the small area estimates or the direct estimaes among with their CVs.

Usage

```
map_sae(shapefile,data,area,indicators,breaks=FALSE,main=FALSE,output_data=FALSE)
```

Arguments

`shapefile`	object of class `sf` and `data.frame` as defined by the `sf` package containing shapefile informations, e.g. `sa_shp`. See also 'Details'.
`data`	data frame containing for the area of interest the information to be visualized, e.g. `SAEval_example`.
`area`	formula identifing the area of interest.
`indicators`	formula identifing the variables to be visualized.
`breaks`	list containing the end points for each indcator of interest (default=FALSE).
`main`	logical scalar. Should the maps include a main title (default=FALSE)?. See also 'Details'.
`output_data`	logical scalar. Should the funtion returns a data frame including the map data among with the indicators of interest (default=FALSE)?. See also 'Details'.

Details

`shapefile` object can be created with the `sf` package using the function `st_read`. If `main` is equal to `TRUE` the name of the indicator will be used as main title of the map. When `output_data` is equal to `TRUE` a map data object is returned so can be easaly maneged using `ggplot` for a better graphical personalizzation.

Value

Returns maps, and, if selected, a data.frame containing the mapdata enriched with the indicators of interest.

Author(s)

Developed by Andrea Fasulo

References

Pebesma E., et al.,2021, "sf: Simple Features for R", CRAN repository https://CRAN.R-project.org/package=sf

Examples

```
# Load example data and shape file
data(SAEval_example);data(sa_shp)

SAEval_example$cv_d<-sqrt(SAEval_example$mse_d)/SAEval_example$y_d

SAEval_example$cv_sa<-sqrt(SAEval_example$mse_sa)/SAEval_example$y_syna

# Without using breaks
map_sae(shapefile=sa_shp,data=SAEval_example,area=~sa,indicators=~y_d+cv_d+y_syna+cv_sa,main=TRUE)

# Using breaks
map_sae(shapefile=sa_shp,data=SAEval_example,area=~sa,indicators=~y_d+cv_d+y_syna+cv_sa,
     breaks=list(seq(0,31000,3000),seq(0,1.5,0.15),seq(0,31000,3000),seq(0,1.5,0.15)),main=TRUE)
```

SAEval_example *Example dataset for the evaluation of Small Area Estimates*

Description

`SAEval_example` contains a `data.frame` with direct and indirect estimates for unplunned domain among with their variance.

Usage

```
data(SAEval_example)
```

Format

`SAEval_example` is a data frame with 107 domains and 18 variables:

`sa` domain of interest codes

`nuts1` NUTS1 codes

`nuts2` NUTS2 codes

`nuts0` NUTS0 codes

`y_d` direct estimated

mse_d variance of direct estimates

y_syna unit level synthetic estimates

mse_sa MSE of unit level synthetic estimates

y_eblupa unit level EBLUP estimates

mse_eba2 MSE of unit level EBLUP estimates

y_spaznr unit level EBLUP estimates with spatial correlation of random effects

mse_spaznr MSE of unit level EBLUP estimates with spatial correlation of random effects

y_eblupb area level EBLUP estimates

mse_ebb MSE of area level EBLUP estimates

y_synb area level synthetic estimates

mse_sb MSE of area level synthetic estimates

y_logis unit level EBLUP type logit estimates

mse_log MSE of unit level EBLUP type logit estimates

Examples

```
# Load example data
data(SAEval_example)
summary(SAEval_example)
# being the domain unplunned there are 7 areas without direct estimates
dim(SAEval_example[SAEval_example$y_d==0,])
```

---

sa_shp *Example dataset to map Small Area Estimates*

Description

sa_shp contains a sf object to map the small area estimates.

Usage

```
data(sa_shp)
```

Format

sa_shp is a sf object with the shapefile for the sa domain.

Examples

```
# Load example data
data(sa_shp)

summary(sa_shp)
```

Index