Package: svyROC (via r-universe)

October 26, 2024

Title Estimation of the ROC Curve and the AUC for Complex Survey Data

Version 1.0.0

Maintainer Amaia Iparragirre <amaia.iparragirre@ehu.eus>

Description Estimate the receiver operating characteristic (ROC) curve, area under the curve (AUC) and optimal cut-off points for individual classification taking into account complex sampling designs when working with complex survey data. Methods implemented in this package are described in: A. Iparragirre, I. Barrio, I. Arostegui (2024) <doi:10.1002/sta4.635>; A. Iparragirre, I. Barrio, J. Aramendi, I. Arostegui (2022) <doi:10.2436/20.8080.02.121>; A. Iparragirre, I. Barrio (2024) <doi:10.1007/978-3-031-65723-8_7>.

License GPL (>= 3)

Encoding UTF-8

RoxygenNote 7.3.2

Depends R (>= 2.10)

LazyData true

Imports survey, svyVarSel

NeedsCompilation no

Author Amaia Iparragirre [aut, cre, cph] (<https://orcid.org/0000-0002-0660-6535>), Irantzu Barrio [aut], Inmaculada Arostegui [aut]

Repository CRAN

Date/Publication 2024-10-25 07:40:02 UTC

Contents

`corrected.wauc` *Corrected estimate of the AUC based on replicate weights.*

Description

Optimism correction of the AUC of logistic regression models with complex survey data based on replicate weights methods.

Usage

```
corrected.wauc(
  data = NULL,
  formula,
  tag.event = NULL,
  tag.nonevent = NULL,
  weights.var = NULL,
  strata.var = NULL,
  cluster.var = NULL,
  design = NULL,
  method = c("dCV", "JKn", "RB"),
  dCV.method = c("average", "pooling"),
  RB.method = c("subbootstrap", "bootstrap"),
  k = 10,
  R = 1,
  B = 200
)
```

Arguments

`data`	A data frame which, at least, must incorporate information on the columns `response.var`, `phat.var` and `weights.var`. If `data=NULL`, the sampling design must be indicated in the argument `design`.
`formula`	Formula of the model for which the AUC needs to be corrected. The models are fitted by means of `survey::svyglm()` function.
`tag.event`	A character string indicating the label used to indicate the event of interest in `response.var`. The default option is `tag.event = NULL`, which selects the class with the lowest number of units as event.
`tag.nonevent`	A character string indicating the label used for non-event in `response.var`. The default option is `tag.nonevent = NULL`, which selects the class with the greatest number of units as non-event.

weights.var	A character string indicating the name of the column with sampling weights. It could be NULL if the sampling design is indicated in the design argument.
strata.var	A character string indicating the name of the column with strata identifiers. It could be NULL if the sampling design is indicated in the design argument.
cluster.var	A character string indicating the name of the column with cluster identifiers. It could be NULL if the sampling design is indicated in the design argument or the sampling design does not have considered clustering.
design	An object of class survey.design generated by survey::svydesign(). It could be NULL if information about cluster.var, strata.var, weights.var and data are given.
method	A character string indicating the method to be applied to define replicate weights and correct the AUC. Choose between: JKn (for the Jackknife Repeated Replication), dCV (for the design-based cross-validation), RB (for the Rescaling Bootstrap).
dCV.method	Only applies for the dCV method. Choose between: average (for the averaging cross-validation) or pooling (for the pooling cross-validation). Note: pooling is recommended over average (see, Iparragirre and Barrio (2024))
RB.method	Only applies for the RB method. Choose between: subbootstrap or bootstrap (see the documentation of svyVarSel::replicate.weights() for help).
k	A numeric value indicating the number of folds to be defined. Default is k=10. Only applies for the dCV method.
R	A numeric value indicating the number of times the sample is partitioned. Default is R=1. Only applies for dCV, split or extrapolation methods.
B	A numeric value indicating the number of bootstrap resamples. Default is B=200. Only applies for bootstrap and subbootstrap methods.

Details

See Iparragirre and Barrio (2024) for more information on the AUC correction methods and their performance.

Value

The output object of this function is a list of 5 elements containing the following information:

- corrected.AUCw: the corrected estimate of the weighted AUC.
- correction.method: the selected correction method.
- formula: formula of the model that has been fitted.
- tags: a list containing two elements with the following information:
  - tag.event: a character string indicating the event of interest.
  - tag.nonevent: a character string indicating the non-event.
- call: an object saving the information about the way in which the function has been run.

References

Iparragirre, A., Barrio, I. (2024). Optimism Correction of the AUC with Complex Survey Data. In: Einbeck, J., Maeng, H., Ogundimu, E., Perrakis, K. (eds) Developments in Statistical Modelling. IWSM 2024. Contributions to Statistics. Springer, Cham. https://doi.org/10.1007/978-3-031-65723-8_7

Examples

```
data(example_variables_wroc)
mydesign <- survey::svydesign(ids = ~cluster, strata = ~strata,
                              weights = ~weights, nest = TRUE,
                              data = example_variables_wroc)
m <- survey::svyglm(y ~ x1 + x2 + x3 + x4 + x5 + x6, design = mydesign,
                    family = quasibinomial())
phat <- predict(m, newdata = example_variables_wroc, type = "response")
myaucw <- wauc(response.var = example_variables_wroc$y, phat.var = phat,
               weights.var = example_variables_wroc$weights)

# Correction of the AUCw:
set.seed(1)
res <- corrected.wauc(data = example_variables_wroc,
                      formula = y ~ x1 + x2 + x3 + x4 + x5 + x6,
                      tag.event = 1, tag.nonevent = 0,
                  weights.var = "weights", strata.var = "strata", cluster.var = "cluster",
                      method = "dCV", dCV.method = "pooling", k = 10, R = 20)
# Or equivalently:

set.seed(1)
res <- corrected.wauc(design = mydesign,
                      formula = y ~ x1 + x2 + x3 + x4 + x5 + x6,
                      tag.event = 1, tag.nonevent = 0,
                      method = "dCV", dCV.method = "pooling", k = 10, R = 20)
```

`example_data_wroc` *Simulated data*

Description

This dataset has been simulated in order to provide the users with an example dataset.

Usage

```
example_data_wroc
```

Format

`example_data_wroc`**:**

A data frame with 740 rows and 3 columns:

y Response variable

phat Predicted probabilities

weights Sampling weights ...

`example_variables_wroc` *Simulated data*

Description

This dataset has been simulated in order to provide the users with an example dataset.

Usage

```
example_variables_wroc
```

Format

`example_variables_wroc`**:**

A data frame with 1720 rows and 10 columns:

y Response variable

x1,...,x6 Covariates

strata Strata variable

cluster Cluster variable

weights Sampling weights ...

`wauc` *Estimation of the AUC of logistic regression models with complex survey data.*

Description

Calculate the AUC of a logistic regression model considering sampling weights with complex survey data

Usage

```
wauc(
  response.var,
  phat.var,
  weights.var = NULL,
  tag.event = NULL,
  tag.nonevent = NULL,
  data = NULL,
  design = NULL
)
```

Arguments

response.var	A character string with the name of the column indicating the response variable in the data set or a vector (either numeric or character string) with information of the response variable for all the units.
phat.var	A character string with the name of the column indicating the estimated probabilities in the data set or a numeric vector containing estimated probabilities for all the units.
weights.var	A character string indicating the name of the column with sampling weights or a numeric vector containing information of the sampling weights. It could be NULL if the sampling design is indicated in the design argument. For unweighted estimates, set all the sampling weight values to 1.
tag.event	A character string indicating the label used to indicate the event of interest in response.var. The default option is tag.event = NULL, which selects the class with the lowest number of units as event.
tag.nonevent	A character string indicating the label used for non-event in response.var. The default option is tag.nonevent = NULL, which selects the class with the greatest number of units as non-event.
data	A data frame which, at least, must incorporate information on the columns response.var, phat.var and weights.var. If data=NULL, then specific numerical vectors must be included in response.var, phat.var and weights.var, or the sampling design should be indicated in the argument design.
design	An object of class survey.design generated by survey::svydesign indicating the complex sampling design of the data. If design = NULL, information on the data set (argument data) and/or sampling weights (argument weights.var) must be included.

Details

S indicate a sample of n observations of the vector of random variables $(Y, \boldsymbol{X})$, and $\forall i = 1, \ldots, n$, y_i indicate the i^{th} observation of the response variable Y, and $\boldsymbol{x}_i$ the observations of the vector covariates $\boldsymbol{X}$. Let w_i indicate the sampling weight corresponding to the unit i and $\hat{p}_i$ the estimated probability of event. Let S_0 and S_1 be subsamples of S, formed by the units without the event of interest ($y_i = 0$) and with the event of interest ($y_i = 1$), respectively. Then, the AUC is estimated

as follows:

$$\widehat{AUC}_w = \frac{\sum_{j\in S_0}\sum_{k\in S_1} w_j w_k \{I(\hat{p}_j < \hat{p}_k) + 0.5 \cdot I(\hat{p}_j = \hat{p}_k)\}}{\sum_{j\in S_0}\sum_{k\in S_1} w_j w_k}.$$

See Iparragirre et al (2023) for more information.

Value

The output object of this function is a list of 4 elements containing the following information:

- `AUCw`: the weighted estimate of the AUC.
- `tags`: a list containing two elements with the following information:
  - `tag.event`: a character string indicating the event of interest.
  - `tag.nonevent`: a character string indicating the non-event.
- `basics`: a list containing information of the following 4 elements:
  - `n.event`: number of units with the event of interest in the data set.
  - `n.nonevent`: number of units without the event of interest in the data set.
  - `hatN.event`: number of units with the event of interest represented in the population by all the event units in the data set, i.e., the sum of the sampling weights of the units with the event of interest in the data set.
  - `hatN.nonevent`: a numeric value indicating the number of non-event units in the population represented by means of the non-event units in the data set, i.e., the sum of the sampling weights of the non-event units in the data set.
- `call`: an object saving the information about the way in which the function has been run.

References

Iparragirre, A., Barrio, I. and Arostegui, I. (2023). Estimation of the ROC curve and the area under it with complex survey data. *Stat* **12**(1), e635. (https://doi.org/10.1002/sta4.635)

Examples

```
data(example_data_wroc)

auc.obj <- wauc(response.var = "y",
                phat.var = "phat",
                weights.var = "weights",
                tag.event = 1,
                tag.nonevent = 0,
                data = example_data_wroc)

# Or equivalently
auc.obj <- wauc(response.var = example_data_wroc$y,
                phat.var = example_data_wroc$phat,
                weights.var = example_data_wroc$weights,
                tag.event = 1, tag.nonevent = 0)
```

wocp *Optimal cut-off points for complex survey data*

Description

Calculate optimal cut-off points for complex survey data (Iparragirre et al., 2022). Some functions of the package OptimalCutpoints (Lopez-Raton et al, 2014) have been used and modified in order them to consider sampling weights.

Usage

```
wocp(
  response.var,
  phat.var,
  weights.var = NULL,
  tag.event = NULL,
  tag.nonevent = NULL,
  method = c("Youden", "MaxProdSpSe", "ROC01", "MaxEfficiency"),
  data = NULL,
  design = NULL
)
```

Arguments

response.var	A character string with the name of the column indicating the response variable in the data set or a vector (either numeric or character string) with information of the response variable for all the units.
phat.var	A character string with the name of the column indicating the estimated probabilities in the data set or a numeric vector containing estimated probabilities for all the units.
weights.var	A character string indicating the name of the column with sampling weights or a numeric vector containing information of the sampling weights. It could be NULL if the sampling design is indicated in the design argument. For unweighted estimates, set all the sampling weight values to 1.
tag.event	A character string indicating the label used to indicate the event of interest in response.var. The default option is tag.event = NULL, which selects the class with the lowest number of units as event.
tag.nonevent	A character string indicating the label used for non-event in response.var. The default option is tag.nonevent = NULL, which selects the class with the greatest number of units as non-event.
method	A character string indicating the method to be used to select the optimal cut-off point. Choose one of the following methods (Lopez-Raton et al, 2014): MaxProdSpSe, ROC01, Youden, MaxEfficiency.

`data` A data frame which, at least, must incorporate information on the columns `response.var`, `phat.var` and `weights.var`. If `data=NULL`, then specific numerical vectors must be included in `response.var`, `phat.var` and `weights.var`, or the sampling design should be indicated in the argument `design`.

`design` An object of class `survey.design` generated by `survey::svydesign` indicating the complex sampling design of the data. If `design = NULL`, information on the data set (argument `data`) and/or sampling weights (argument `weights.var`) must be included.

Details

Let S indicate a sample of n observations of the vector of random variables $(Y, \boldsymbol{X})$, and $\forall i = 1, \ldots, n$, y_i indicate the i^{th} observation of the response variable Y, and $\boldsymbol{x}_i$ the observations of the vector covariates $\boldsymbol{X}$. Let w_i indicate the sampling weight corresponding to the unit i and $\hat{p}_i$ the estimated probability of event. Let S_0 and S_1 be subsamples of S, formed by the units without the event of interest ($y_i = 0$) and with the event of interest ($y_i = 1$), respectively. Then, the optimal cut-off points are obtained as follows:

- `Youden`:
$$c_w^{\text{Youden}} = argmax_c\{\widehat{Se}_w(c) + \widehat{Sp}_w(c) - 1\},$$
- `MaxProdSpSe`:
$$c_w^{\text{MaxProdSpSe}} = argmax_c\{\widehat{Se}_w(c) * \widehat{Sp}_w(c)\},$$
- `ROC01`:
$$c_w^{\text{ROC01}} = argmax_c\{(\widehat{Se}_w(c) - 1)^2 + (\widehat{Sp}_w(c) - 1)^2\},$$
- `MaxEfficiency`:
$$c_w^{\text{MaxEfficiency}} = argmax_c\{\hat{p}_{Y,w}\widehat{Se}_w(c) + (1 - \hat{p}_{Y,w})\widehat{Sp}_w(c)\},$$

where, the sensitivity and specificity parameters for a given cut-off point c are estimated as follows:

$$\widehat{Se}_w(c) = \frac{\sum_{i \in S_1} w_i \cdot I(\hat{p}_i \geq c)}{\sum_{i \in S_1} w_i} \; ; \; \widehat{Sp}_w(c) = \frac{\sum_{i \in S_0} w_i \cdot I(\hat{p}_i < c)}{\sum_{i \in S_0} w_i},$$

and,

$$\hat{p}_{Y,w} = \frac{\sum_{i \in S} w_i \cdot I(y_i = 1)}{\sum_{i \in S} w_i}.$$

See Iparragirre et al. (2022) and Lopez-Raton et al. (2014) for more information.

Value

The output of this function is an object of class `wocp`. This object is a list that contains information about the following 4 elements:

- `tags`: a list containing two elements with the following information:
  - `tag.event`: a character string indicating the event of interest.
  - `tag.nonevent`: a character string indicating the non-event.
- `basics`: a list containing information of the following 4 elements:

  - n.event: number of units with the event of interest in the data set.
  - n.nonevent: number of units without the event of interest in the data set.
  - hatN.event: number of units with the event of interest represented in the population by all the event units in the data set, i.e., the sum of the sampling weights of the units with the event of interest in the data set.
  - hatN.nonevent: a numeric value indicating the number of non-event units in the population represented by means of the non-event units in the data set, i.e., the sum of the sampling weights of the non-event units in the data set.

- optimal.cutoff: this object is a list of three elements containing the information described below:
  - method: a character string indicating the method implemented to select the optimal cut-off point.
  - optimal: a list containing information of the following four elements:
    * cutoff: a numeric vector indicating the optimal cut-off point(s) that optimize(s) the selected criterion.
    * Sew: a numeric vector indicating the estimated sensitivity parameter(s) corresponding to the optimal cut-off point(s) that optimize(s) the selected criterion.
    * Spw: a numeric vector indicating the estimated specificity parameter(s) corresponding to the optimal cut-off point(s) that optimize(s) the selected criterion.
    * criterion: a numeric value indicating the criterion value optimized by means of the selected optimal cut-off point(s).
  - all: a list containing information on the following four elements:
    * cutoff: a numeric vector indicating all the cut-off points considered.
    * Sew: a numeric vector indicating the estimated sensitivity parameters corresponding to all the considered cut-off points.
    * Spw: a numeric vector indicating the estimated sensitivity parameters corresponding to all the considered cut-off points.
    * criterion: a numeric vector indicating the values of the selected criterion corresponding to all the considered cut-off points.
- call: an object saving the information about the way in which the function has been run.

References

Iparragirre, A., Barrio, I., Aramendi, J. and Arostegui, I. (2022). Estimation of cut-off points under complex-sampling design data. *SORT-Statistics and Operations Research Transactions* **46**(1), 137–158.

Lopez-Raton, M., Rodriguez-Alvarez, M.X, Cadarso-Suarez, C. and Gude-Sampedro, F. (2014). OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software* **61**(8), 1–36.

Examples

```
data(example_data_wroc)

myocp <- wocp(response.var = "y", phat.var = "phat", weights.var = "weights",
              tag.event = 1, tag.nonevent = 0,
```

```
              method = "Youden",
              data = example_data_wroc)

# Or equivalently
myocp <- wocp(example_data_wroc$y, example_data_wroc$phat, example_data_wroc$weights,
              tag.event = 1, tag.nonevent = 0, method = "Youden")
```

---

wroc *Estimation of the ROC curve of logistic regression models with complex survey data*

Description

Calculate the ROC curve of a logistic regression model considering sampling weights with complex survey data

Usage

```
wroc(
  response.var,
  phat.var,
  weights.var = NULL,
  tag.event = NULL,
  tag.nonevent = NULL,
  data = NULL,
  design = NULL,
  cutoff.method = NULL
)
```

Arguments

response.var	A character string with the name of the column indicating the response variable in the data set or a vector (either numeric or character string) with information of the response variable for all the units.
phat.var	A character string with the name of the column indicating the estimated probabilities in the data set or a numeric vector containing estimated probabilities for all the units.
weights.var	A character string indicating the name of the column with sampling weights or a numeric vector containing information of the sampling weights. It could be NULL if the sampling design is indicated in the design argument. For unweighted estimates, set all the sampling weight values to 1.
tag.event	A character string indicating the label used to indicate the event of interest in response.var. The default option is tag.event = NULL, which selects the class with the lowest number of units as event.

`tag.nonevent`	A character string indicating the label used for non-event in `response.var`. The default option is `tag.nonevent = NULL`, which selects the class with the greatest number of units as non-event.
`data`	A data frame which, at least, must incorporate information on the columns `response.var`, `phat.var` and `weights.var`. If `data=NULL`, then specific numerical vectors must be included in `response.var`, `phat.var` and `weights.var`, or the sampling design should be indicated in the argument `design`.
`design`	An object of class `survey.design` generated by `survey::svydesign` indicating the complex sampling design of the data. If `design = NULL`, information on the data set (argument `data`) and/or sampling weights (argument `weights.var`) must be included.
`cutoff.method`	A character string indicating the method to be used to select the optimal cut-off point. If `cutoff.method = NULL`, then no optimal cut-off point is calculated. If an optimal cut-off point is to be calculated, one of the following methods needs to be selected: `Youden`, `MaxProdSpSe`, `ROC01`, `MaxEfficiency`.

Details

S indicate a sample of n observations of the vector of random variables $(Y, \boldsymbol{X})$, and $\forall i = 1, \ldots, n$, y_i indicate the i^{th} observation of the response variable Y, and $\boldsymbol{x}_i$ the observations of the vector covariates $\boldsymbol{X}$. Let w_i indicate the sampling weight corresponding to the unit i and $\hat{p}_i$ the estimated probability of event. Let S_0 and S_1 be subsamples of S, formed by the units without the event of interest ($y_i = 0$) and with the event of interest ($y_i = 1$), respectively. Then, the ROC curve is estimated as follows:

$$\widehat{ROC}_w(\cdot) = \{(1 - \widehat{Sp}_w(c), \widehat{Se}_w(c)),\ c \in (-\infty, \infty)\}$$

, where, the sensitivity and specificity parameters for a given cut-off point c are estimated as follows:

$$\widehat{Se}_w(c) = \frac{\sum_{i \in S_1} w_i \cdot I(\hat{p}_i \geq c)}{\sum_{i \in S_1} w_i}\ ;\ \widehat{Sp}_w(c) = \frac{\sum_{i \in S_0} w_i \cdot I(\hat{p}_i < c)}{\sum_{i \in S_0} w_i}.$$

See Iparragirre et al (2023) for more information. More information of the rest of the elements is given in the documentation of the functions `wauc()` and `wocp()`.

Value

The output object of this function is a list of class `wroc`, which contains information about the weighted ROC curve of a logistic regression model and some of its components. In particular, this list contains a total of 5 or 6 elements (depending on the selected arguments) with the following information:

- `wroc.curve`: this element is a list that contains three numerical vectors. Specifically,
  - `Sew.values`: a vector of all the different values for the weighted estimate of the sensitivity across all the possible cut-off points.
  - `Spw.values`: a vector of all the different values for the weighted estimate of the specificity across all the possible cut-off points.
  - `cutoffs`: this vector contains all the cut-off points that have been considered to estimate sensitivity and specificity parameters.

- wauc: a numeric value indicating the area under the weighted estimate of the ROC curve.
- optimal.cutoff: if the argument cutoff.method != NULL, this object is a list containing the 4 elements described below:
  - method: character string indicating the method implemented to calculate the optimal cut-off point.
  - cutoff.value: the optimal cut-off point value.
  - Spw: the weighted estimate of the specificity for the optimal cut-off point value (indicated in cutoff.value).
  - Sew: the weighted estimate of the sensitivity for the optimal cut-off point value (indicated in cutoff.value).
- tags: a list containing two elements with the following information:
  - tag.event: a character string indicating the event of interest.
  - tag.nonevent: a character string indicating the non-event.
- basics: a list containing information of the following 4 elements:
  - n.event: number of units with the event of interest in the data set.
  - n.nonevent: number of units without the event of interest in the data set.
  - hatN.event: number of units with the event of interest represented in the population by all the event units in the data set, i.e., the sum of the sampling weights of the units with the event of interest in the data set.
  - hatN.nonevent: a numeric value indicating the number of non-event units in the population represented by means of the non-event units in the data set, i.e., the sum of the sampling weights of the non-event units in the data set.
- call: an object saving the information about the way in which the function has been run.

References

Iparragirre, A., Barrio, I. and Arostegui, I. (2023). Estimation of the ROC curve and the area under it with complex survey data. *Stat* **12**(1), e635. (https://doi.org/10.1002/sta4.635)

Examples

```
data(example_data_wroc)

mycurve <- wroc(response.var = "y", phat.var = "phat", weights.var = "weights",
                data = example_data_wroc,
                tag.event = 1, tag.nonevent = 0,
                cutoff.method = "Youden")

# Or equivalently

mycurve <- wroc(response.var = example_data_wroc$y,
                phat.var = example_data_wroc$phat,
                weights.var = example_data_wroc$weights,
                tag.event = 1, tag.nonevent = 0,
                cutoff.method = "Youden")
```

`wroc.plot` *Estimation of the ROC curve of logistic regression models with complex survey data*

Description

Plot the ROC curve of a logistic regression model considering sampling weights with complex survey data.

Usage

```
wroc.plot(
  x,
  print.auc = TRUE,
  print.cutoff = FALSE,
  col.cutoff = "red",
  cex.text = 0.75,
  round.digits = 4
)
```

Arguments

`x`	An object of class `wroc` obtained by means of the function `wroc()`.
`print.auc`	A logical value. If `TRUE`, the value of the area under the ROCw curve (AUCw) is printed (default `print.auc = TRUE`).
`print.cutoff`	A logical value. If `TRUE`, the value of the optimal cut-off point, and the corresponding weighted estimates of the sensitivity and specificity parameters are printed (default `print.cutoff = TRUE`).
`col.cutoff`	A character string indicating the color in which the cut-off point is depicted. The default option is `col.cutoff = "red"`.
`cex.text`	A numeric value indicating the size with which the information of the AUCw and optimal cut-off point is printed. The default option is `cex.text = 0.75`.
`round.digits`	A numeric value indicating the number of digits that will be employed when printing the information about the AUCw and optimal cut-off point. The default option is `round.digits = 4`.

Details

More information is given in the documentation of the `wroc()`, `wauc{}` and `wocp()` functions.

Value

a graph

Examples

```
data(example_data_wroc)

mycurve <- wroc(response.var = "y", phat.var = "phat", weights.var = "weights",
                data = example_data_wroc,
                tag.event = 1, tag.nonevent = 0,
                cutoff.method = "Youden")
wroc.plot(x = mycurve, print.auc = TRUE, print.cutoff = TRUE)
```

`wse` *Estimation of the sensitivity with complex survey data*

Description

Estimate the sensitivity parameter for a given cut-off point considering sampling weights with complex survey data.

Usage

```
wse(
  response.var,
  phat.var,
  weights.var = NULL,
  tag.event = NULL,
  cutoff.value,
  data = NULL,
  design = NULL
)
```

Arguments

`response.var`	A character string with the name of the column indicating the response variable in the data set or a vector (either numeric or character string) with information of the response variable for all the units.
`phat.var`	A character string with the name of the column indicating the estimated probabilities in the data set or a numeric vector containing estimated probabilities for all the units.
`weights.var`	A character string indicating the name of the column with sampling weights or a numeric vector containing information of the sampling weights. It could be `NULL` if the sampling design is indicated in the `design` argument. For unweighted estimates, set all the sampling weight values to 1.
`tag.event`	A character string indicating the label used to indicate the event of interest in `response.var`. The default option is `tag.event = NULL`, which selects the class with the lowest number of units as event.

`cutoff.value`	A numeric value indicating the cut-off point to be used. No default value is set for this argument, and a numeric value must be indicated necessarily.
`data`	A data frame which, at least, must incorporate information on the columns `response.var`, `phat.var` and `weights.var`. If `data=NULL`, then specific numerical vectors must be included in `response.var`, `phat.var` and `weights.var`, or the sampling design should be indicated in the argument `design`.
`design`	An object of class `survey.design` generated by `survey::svydesign` indicating the complex sampling design of the data. If `design = NULL`, information on the data set (argument `data`) and/or sampling weights (argument `weights.var`) must be included.

Details

Let S indicate a sample of n observations of the vector of random variables $(Y, \boldsymbol{X})$, and $\forall i = 1, \ldots, n$, y_i indicate the i^{th} observation of the response variable Y, and $\boldsymbol{x}_i$ the observations of the vector covariates $\boldsymbol{X}$. Let w_i indicate the sampling weight corresponding to the unit i and $\hat{p}_i$ the estimated probability of event. Let S_0 and S_1 be subsamples of S, formed by the units without the event of interest ($y_i = 0$) and with the event of interest ($y_i = 1$), respectively. Then, the sensitivity parameter for a given cut-off point c is estimated as follows:

$$\widehat{Se}_w(c) = \frac{\sum_{i \in S_1} w_i \cdot I(\hat{p}_i \geq c)}{\sum_{i \in S_1} w_i}.$$

See Iparragirre et al. (2022) and Iparragirre et al. (2023) for more details.

Value

The output of this function is a list of 4 elements containing the following information:

- `Sew`: a numeric value indicating the weighted estimate of the sensitivity parameter.
- `tags`: list containing one element with the following information:
  - `tag.event`: a character string indicating the label used to indicate event of interest.
- `basics`: a list containing information of the following 6 elements:
  - `n`: a numeric value indicating the number of units in the data set.
  - `n.event`: a numeric value indicating the number of units in the data set with the event of interest.
  - `n.event.class`: a numeric value indicating the number of units in the data set with the event of interest that are correctly classified as events based on the selected cut-off point.
  - `hatN`: number of units in the population, represented by all the units in the data set, i.e., the sum of the sampling weights of the units in the data set.
  - `hatN.event`: number of units with the event of interest represented in the population by all the event units in the data set, i.e., the sum of the sampling weights of the units with the event of interest in the data set.
  - `hatN.event.class`: number of event units represented in the population by the event units in the data set that have been correctly classified as events based on the selected cut-off point, i.e., the sum of the sampling weights of the correctly classified event units in the data set.
- `call`: an object saving the information about the way in which the function has been run.

References

Iparragirre, A., Barrio, I., Aramendi, J. and Arostegui, I. (2022). Estimation of cut-off points under complex-sampling design data. *SORT-Statistics and Operations Research Transactions* **46**(1), 137–158. (https://doi.org/10.2436/20.8080.02.121)

Iparragirre, A., Barrio, I. and Arostegui, I. (2023). Estimation of the ROC curve and the area under it with complex survey data. *Stat* **12**(1), e635. (https://doi.org/10.1002/sta4.635)

Examples

```
data(example_data_wroc)

se.obj <- wse(response.var = "y", phat.var = "phat", weights.var = "weights",
              tag.event = 1, cutoff.value = 0.5, data = example_data_wroc)

# Or equivalently
se.obj <- wse(response.var = example_data_wroc$y,
              phat.var = example_data_wroc$phat,
              weights.var = example_data_wroc$weights,
              tag.event = 1, cutoff.value = 0.5)
```

---

wsp *Estimation of the specificity with complex survey data*

Description

Estimate the specificity parameter for a given cut-off point considering sampling weights with complex survey data.

Usage

```
wsp(
  response.var,
  phat.var,
  weights.var = NULL,
  tag.nonevent = NULL,
  cutoff.value,
  data = NULL,
  design = NULL
)
```

Arguments

response.var	A character string with the name of the column indicating the response variable in the data set or a vector (either numeric or character string) with information of the response variable for all the units.
phat.var	A character string with the name of the column indicating the estimated probabilities in the data set or a numeric vector containing estimated probabilities for all the units.

`weights.var` A character string indicating the name of the column with sampling weights or a numeric vector containing information of the sampling weights. It could be `NULL` if the sampling design is indicated in the `design` argument. For unweighted estimates, set all the sampling weight values to 1.

`tag.nonevent` A character string indicating the label used for non-event in `response.var`. The default option is `tag.nonevent = NULL`, which selects the class with the greatest number of units as non-event.

`cutoff.value` A numeric value indicating the cut-off point to be used. No default value is set for this argument, and a numeric value must be indicated necessarily.

`data` A data frame which, at least, must incorporate information on the columns `response.var`, `phat.var` and `weights.var`. If `data=NULL`, then specific numerical vectors must be included in `response.var`, `phat.var` and `weights.var`, or the sampling design should be indicated in the argument `design`.

`design` An object of class `survey.design` generated by `survey::svydesign` indicating the complex sampling design of the data. If `design = NULL`, information on the data set (argument `data`) and/or sampling weights (argument `weights.var`) must be included.

Details

Let S indicate a sample of n observations of the vector of random variables $(Y, \boldsymbol{X})$, and $\forall i = 1, \ldots, n$, y_i indicate the i^{th} observation of the response variable Y, and $\boldsymbol{x}_i$ the observations of the vector covariates $\boldsymbol{X}$. Let w_i indicate the sampling weight corresponding to the unit i and $\hat{p}_i$ the estimated probability of event. Let S_0 and S_1 be subsamples of S, formed by the units without the event of interest ($y_i = 0$) and with the event of interest ($y_i = 1$), respectively. Then, the specificity parameter for a given cut-off point c is estimated as follows:

$$\widehat{Sp}_w(c) = \frac{\sum_{i \in S_0} w_i \cdot I(\hat{p}_i < c)}{\sum_{i \in S_0} w_i}.$$

See Iparragirre et al. (2022) and Iparragirre et al. (2023) for more details.

Value

The output of this function is a list of 4 elements containing the following information:

- `Spw`: a numeric value indicating the weighted estimate of the specificity parameter.
- `tags`: a list containing one element with the following information:
  - `tag.nonevent`: a character string indicating the label used for non-events.
- `basics`: a list containing information of the following 6 elements:
  - `n`: a numeric value indicating the number of units in the data set.
  - `n.nonevent`: a numeric value indicating the number of units in the data set without the event of interest.
  - `n.nonevent.class`: a numeric value indicating the number of units in the data set without the event of interest that are correctly classified as non-events based on the selected cut-off point.

  - `hatN`: a numeric value indicating the number of units in the population that are represented by means of the units in the data set, i.e., the sum of the sampling weights of all the units in the data set.
  - `hatN.nonevent`: a numeric value indicating the number of non-event units in the population represented by means of the non-event units in the data set, i.e., the sum of the sampling weights of the non-event units in the data set.
  - `hatN.nonevent.class`: number of non-event units represented in the population by the non-event units in the data set that have been correctly classified as non-events based on the selected cut-off point, i.e., the sum of the sampling weights of the correctly classified non-event units in the data set.

- `call`: an object saving the information about the way in which the function has been run.

References

Iparragirre, A., Barrio, I., Aramendi, J. and Arostegui, I. (2022). Estimation of cut-off points under complex-sampling design data. *SORT-Statistics and Operations Research Transactions* **46**(1), 137–158. (https://doi.org/10.2436/20.8080.02.121)

Iparragirre, A., Barrio, I. and Arostegui, I. (2023). Estimation of the ROC curve and the area under it with complex survey data. *Stat* **12**(1), e635. (https://doi.org/10.1002/sta4.635)

Examples

```
data(example_data_wroc)

sp.obj <- wsp(response.var = "y",
              phat.var = "phat",
              weights.var = "weights",
              tag.nonevent = 0,
              cutoff.value = 0.5,
              data = example_data_wroc)

# Or equivalently
sp.obj <- wsp(response.var = example_data_wroc$y,
              phat.var = example_data_wroc$phat,
              weights.var = example_data_wroc$weights,
              tag.nonevent = 0,
              cutoff.value = 0.5)
sp.obj
```

Index