Package 'ThresholdROCsurvival'

August 11, 2022

Type Package

Title Threshold and AUC Estimation with Right-Censored Data at a Fixed Time t

Version 1.0.2

Date 2022-08-11

Description We focus on the estimation of optimal thresholds and AUCs when the outcome of interest is the status (alive or dead) of the subjects at a certain time-point t. This binary status is determined by right-censored times to event and it is missing for those subjects censored before t. Here we provide three methods (missing exclusion, imputation of censored times and using time-dependent ROC curves) to estimate optimal thresholds and AUCs in this context. Two references for the methods used here are Skaltsa et al. (2010) <doi:10.1002/bimj.200900294> and Heagerty et al. (2000) <doi:10.1111/j.0006-341x.2000.00337.x>.

License GPL (>= 2)

Depends R (>= 4.0.0)

Imports boot, InformativeCensoring, pROC, psych, survival, survivalROC, ThresholdROC

LazyData TRUE

NeedsCompilation no

Author Sara Perez-Jaume [aut, cre], Josep L Carrasco [aut]

Maintainer Sara Perez-Jaume <spjaume@gmail.com>

Repository CRAN

Date/Publication 2022-08-11 14:30:05 UTC

R topics documented:

`ThresholdROCsurvival-package`

Optimum threshold and AUC estimation based on cost function in a context of right-censored data

---

Description

We focus on the estimation of optimal thresholds and AUCs when the outcome of interest is the status (alive or dead) of the subjects at a certain time-point t. This binary status is determined by right-censored times to event and it is missing for those subjects censored before t. Here we provide three methods (missing exclusion, imputation of censored times and using time-dependent ROC curves) to estimate optimal thresholds and AUCs in this context. Two references for the methods used here are Skaltsa et al. (2010) <doi:10.1002/bimj.200900294> and Heagerty et al. (2000) <doi:10.1111/j.0006-341x.2000.00337.x>.

Details

The functions in this package are `th_ME()`, `th_ICT()` and `th_survivalROC()` for threshold estimation and `AUC_ME()` and `AUC_ICT()` for AUC estimation.

Author(s)

NA

Maintainer: NA

References

Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. Biometrics 2000; 56(2): 337-344. doi: 10.1111/j.0006-341X.2000.00337.x

Hsu CH, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. Statistics in Medicine 2006; 25(20): 3503-3517. doi: https://doi.org/10.1002/sim.2452

Perez-Jaume S, Skaltsa K, Pallares N, Carrasco JL. ThresholdROC: Optimum Threshold Estimation Tools for Continuous Diagnostic Tests in R. Journal of Statistical Software 2017; 82(4): 1-21. doi: 10.18637/jss.v082.i04

Skaltsa K, Jover L, Carrasco JL. Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. Biometrical Journal 2010; 52(5): 676-697. doi: 10.1002/bimj.200900294

AUC_ICT *AUC estimation using the imputation of censored times (ICT) method*

Description

This function estimates the AUC with survival data by using a method based on the imputation of censored times (ICT). The status of the subjects at a certain time-point of interest t (the event ocurred before or at t or not) is defined by the time-to-event variable.

Usage

```
AUC_ICT(cont.var, time, status, predict.time, m = 10,
        ci = TRUE, alpha = 0.05, range = 3)
```

Arguments

cont.var	continuous variable or biomarker to be used as predictor of the status
time	survival time
status	censoring status codified as 0=censored, 1=event
predict.time	time-point of interest to define the subjects' status as event present or absent
m	the number of data sets to impute
ci	Should a confidence interval be calculated? Default, TRUE
alpha	significance level for the confidence interval. Default, 0.05
range	this value, which is passed to boxplot function from graphics package, determines the data points that are considered to be extreme in the estimates and standard errors from the multiple imputation process. We consider extreme observations those that exceed range times the interquartile range. If extreme observations are found in the estimates or standard errors from the multiple imputation process, Winsorized estimators (Wilcox, 2012) are used for the point estimate of the threshold and the between and within variances. Default, 3

Details

First, the algorithm determines the status of the subjects at time predict.time. For those subjects whose status could not be determined because their censored time is lower than t (therefore, with missing status), we impute survival times using the method of Hsu et al (2006), that is implemented in the package InformativeCensoring (Ruau et al, 2020). The status of the subjects is then determined by these imputed times and is used to estimate the AUC using the roc function from pROC package (Robin et al, 2011).

Confidence intervals are calculated using the standard error proposed by Rubin (1987).

Value

An object of class AUC_ICT, which is a list with the following components:

estimate	AUC estimate
se	Standard error of the estimate, obtained using Rubin rules (note: NULL if ci=FALSE)
CI	Confidence interval for the estimate, obtained using Rubin rules (note: NULL if ci=FALSE)
data	A data.frame containing the following columns previously provided by the user: cont.var, time and status, and a new column statusNA, which contains the status of the subjects at time predict.time (0=no event, 1=event, NA=missing)

References

Hsu CH, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. Statistics in Medicine 2006; 25(20): 3503-3517. doi: https://doi.org/10.1002/sim.2452

David Ruau, Nikolas Burkoff, Jonathan Bartlett, Dan Jackson, Edmund Jones, Martin Law and Paul Metcalfe (2020). InformativeCensoring: Multiple Imputation for Informative Censoring. R package version 0.3.5. https://CRAN.R-project.org/package=InformativeCensoring

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12. doi:10.1186/1471-2105-12-77

Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics. John Wiley & Sons (1987).

Wilcox, R. Introduction to Robust Estimation and Hypothesis Testing. 3rd Edition. Elsevier, Amsterdam (2012)

See Also

AUC_ME

Examples

```
data(NSCLC)
set.seed(2020)
res <- with(NSCLC, AUC_ICT(COL, OS, ST, 1095, m=50))
res
res$data
```

AUC_ME *AUC estimation using the missing exclusion (ME) method*

Description

This function estimates the AUC with survival data using by excluding subjects with missing status at the time-point of interest

Usage

```
AUC_ME(cont.var, time, status, predict.time, plot = FALSE,
       ci = TRUE, alpha = 0.05, ...)
```

Arguments

cont.var	continuous variable or biomarker to be used as predictor of the status
time	survival time
status	censoring status codified as 0=censored, 1=event
predict.time	time-point of interest to define the subjects' status as event present or absent
plot	Should a graph of the ROC curve be plotted? Default, FALSE
ci	Should a confidence interval be calculated? Default, TRUE
alpha	significance level for the confidence interval. Default, 0.05
...	further arguments to be passed to plot()

Details

First, the algorithm determines the status of the subjects at time predict.time. Those subjects whose status could be not be determined (therefore, with missing status) are excluded from the analysis. Then, the AUC is estimated using the roc function from pROC package (Robin et al, 2011). Confidence intervals for the AUC are calculated using the logit transformation (Kottas et al, 2014).

Value

An object of class AUC_ME, which is a list with three components:

AUC	AUC estimate
CI	2-dimensional vector containing the confidence interval
data	a data.frame containing the following columns previously provided by the user: cont.var, time and status, and a new column statusNA, which contains the status of the subjects at time predict.time (0=no event, 1=event, NA=missing)

References

Kottas M, Kuss O, Zapf A. A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies. BMC Medical Research Methodology 2014; 14(26). doi:10.1186/1471-2288-14-26

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12. doi:10.1186/1471-2105-12-77

See Also

th_ME

Examples

```
data(NSCLC)
res <- with(NSCLC, AUC_ME(COL, OS, ST, 1095, plot=FALSE))
res
res$data
```

NSCLC *Non-small cell lung cancer (NSCLC) data*

Description

Non-small cell lung cancer (NSCLC) is the most common lung cancer and comprises several subtypes of lung cancers. These data come from a study by Alcaraz *et al.*, 2019, in which the authors investigated the prognostic value of some activation markers in NSCLC.

Usage

```
data("NSCLC")
```

Format

A data frame with 203 observations on the following 4 variables.

`ID` subject's identifier

`OS` overall survival, that is, the time from surgery until death or last follow-up, in days

`ST` censoring status (0=censored, 1=dead)

`COL` percentage of collagen quantified using an imaging technique from tumour samples

Source

Alcaraz J, Carrasco JL, Millares L, et al. Stromal markers of activated tumor associated fibroblasts predict poor survival and are associated with necrosis in non-small cell lung cancer. Lung Cancer 2019; 135: 151 - 160. doi: 10.1016/j.lungcan.2019.07.020

Examples

```
data(NSCLC)
summary(NSCLC)
```

th_ICT *Threshold estimation using the imputation of censored times (ICT) method*

Description

This function estimates optimum thresholds with survival data by using a method based on the imputation of censored times (ICT). The status of the subjects at a certain time-point of interest t (the event ocurred before or at t or not) is defined by the time-to-event variable.

Usage

```
th_ICT(cont.var, time, status, predict.time, costs = NULL,
       R = NULL, method = c("normal", "empirical"),
       var.equal = FALSE, m = 10, ci = TRUE, alpha = 0.05,
       B = 1000, range = 3)
```

Arguments

cont.var	continuous variable or biomarker to be used as predictor of the status
time	survival time
status	censoring status codified as 0=censored, 1=event
predict.time	time-point of interest to define the subjects' status as event present or absent
costs	cost matrix. Costs should be entered as a 2x2 matrix, where the first row corresponds to the true positive and true negative costs and the second row to the false positive and false negative costs. Default cost values (costs=NULL, when also R=NULL) are a combination of costs that yields R=1, which is equivalent to the Youden index method (for details about this concept, see **Details** and **References**)
R	if the cost matrix costs is not set (that is, costs=NULL), R desired (the algorithm will choose a suitable combination of costs that leads to R). Default, NULL (which leads to R=1 using the default costs). For details about this concept, see **Details** and **References**
method	method used in the estimation: "normal" (default) or "empirical". The user can specify just the initial letters. See **Details** for more information about the methods available
var.equal	when method="normal", assume equal variances? Default, FALSE. When method="empirical", var.equal is ignored
m	the number of data sets to impute
ci	Should a confidence interval be calculated? Default, TRUE
alpha	significance level for the confidence interval. Default, 0.05
B	number of bootstrap resamples for the confidence interval when method="empirical". Otherwise, ignored. Default, 1000

range — this value, which is passed to boxplot function from graphics package, determines the data points that are considered to be extreme in the estimates and standard errors from the multiple imputation process. We consider extreme observations those that exceed range times the interquartile range. If extreme observations are found in the estimates or standard errors from the multiple imputation process, Winsorized estimators (Wilcox, 2012) are used for the point estimate of the threshold and the between and within variances. Default, 3

Details

First, the algorithm determines the status of the subjects at time predict.time. For those subjects whose status could not be determined because their censored time is lower than t (therefore, with missing status), we impute survival times using the method of Hsu et al (2006), that is implemented in the package InformativeCensoring (Ruau et al, 2020). The status of the subjects is then determined by these imputed times and is used to estimate the optimum threshold by minimizing the cost function using the thres2 function in the ThresholdROC package (Perez-Jaume et al, 2017).

For parameter method the user can choose between "normal" (assumes binormality) or "empirical". When method="normal", the user can specify if the algorithm should assume equal or different variances using the parameter var.equal. For further details see the thres2 function in the ThresholdROC package.

Confidence intervals are calculated using the standard error proposed by Rubin (1987).

R, mentioned in parameters costs and R, is the product of the non-disease odds and the cost ratio:

$$R = ((1-p)/p)((C_{TN} - C_{FP})/(C_{TP} - C_{FN})),$$

where p is the disease prevalence (estimated using Kaplan-Meier) and C_i are the classification costs.

Value

An object of class th_ICT, which is a list with the following components:

T — A list of five elements:
- thres threshold estimate.
- prev disease prevalence used.
- costs cost matrix.
- R R term, the product of the non-disease odds and the cost ratio (for further details about this concept, see **References**).
- method method used in the estimation.

CI — A list of five elements (or NULL if ci=FALSE):
- lower the lower limit of the confidence interval.
- upper the upper limit of the confidence interval.
- se the standard error.
- alpha significance level provided by the user.
- ci.method method used for the confidence intervals calculation.

sens	A list of three elements: est estimate of sensitivity at thres. lower the lower limit of the confidence interval for the estimate of sensitivity at thres (NULL if ci=FALSE). upper the upper limit of the confidence interval for the estimate of sensitivity at thres (NULL if ci=FALSE).
spec	A list of three elements: est estimate of specificity at thres. lower the lower limit of the confidence interval for the estimate of specificity at thres (NULL if ci=FALSE). upper the upper limit of the confidence interval for the estimate of specificity at thres (NULL if ci=FALSE).
data	A data.frame containing the following columns previously provided by the user: cont.var, time and status, and a new column statusNA, which contains the status of the subjects at time predict.time (0=no event, 1=event, NA=missing)

References

Hsu CH, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. Statistics in Medicine 2006; 25(20): 3503-3517. doi: https://doi.org/10.1002/sim.2452

David Ruau, Nikolas Burkoff, Jonathan Bartlett, Dan Jackson, Edmund Jones, Martin Law and Paul Metcalfe (2020). InformativeCensoring: Multiple Imputation for Informative Censoring. R package version 0.3.5. https://CRAN.R-project.org/package=InformativeCensoring

Perez-Jaume S, Skaltsa K, Pallares N, Carrasco JL. ThresholdROC: Optimum Threshold Estimation Tools for Continuous Diagnostic Tests in R. Journal of Statistical Software 2017; 82(4): 1-21. doi: 10.18637/jss.v082.i04

Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics. John Wiley & Sons (1987).

Skaltsa K, Jover L, Carrasco JL. Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. Biometrical Journal 2010; 52(5): 676-697. doi: 10.1002/bimj.200900294

Wilcox, R. Introduction to Robust Estimation and Hypothesis Testing. 3rd Edition. Elsevier, Amsterdam (2012)

See Also

thres2

Examples

```
data(NSCLC)
set.seed(2020)
res <- with(NSCLC, th_ICT(log(COL), OS, ST, 1095, method="normal", var.equal=FALSE, m=50))
res
exp(res$T$thres)
exp(res$CI$lower)
exp(res$CI$upper)
res$data
```

```
res$sens
res$spec
```

---

th_ME *Threshold estimation using the missing exclusion (ME) method*

Description

This function estimates optimum thresholds with survival data by excluding subjects with missing status at the time-point of interest

Usage

```
th_ME(cont.var, time, status, predict.time, costs = NULL,
      R = NULL, method = c("normal", "empirical"),
      var.equal = FALSE, plot = FALSE, ci = TRUE,
      alpha = 0.05, B = 1000, ...)
```

Arguments

`cont.var`	continuous variable or biomarker to be used as predictor of the status
`time`	survival time
`status`	censoring status codified as 0=censored, 1=event
`predict.time`	time-point of interest to define the subjects' status as event present or absent
`costs`	cost matrix. Costs should be entered as a 2x2 matrix, where the first row corresponds to the true positive and true negative costs and the second row to the false positive and false negative costs. Default cost values (`costs=NULL`, when also `R=NULL`) are a combination of costs that yields R=1, which is equivalent to the Youden index method (for details about this concept, see **Details** and **References**)
`R`	if the cost matrix `costs` is not set (that is, `costs=NULL`), R desired (the algorithm will choose a suitable combination of costs that leads to `R`). Default, `NULL` (which leads to R=1 using the default `costs`). For details about this concept, see **Details** and **References**
`method`	method used in the estimation: `"normal"` (default) or `"empirical"`. The user can specify just the initial letters. See **Details** for more information about the methods available
`var.equal`	When `method="normal"`, assume equal variances? Default, `FALSE`. When `method="empirical"`, `var.equal` is ignored
`plot`	Should some graphs about the estimation be plotted? Default, `FALSE`
`ci`	Should a confidence interval be calculated? Default, `TRUE`
`alpha`	significance level for the confidence interval. Default, 0.05
`B`	number of bootstrap resamples for the confidence interval when `method="empirical"`. Otherwise, ignored. Default, 1000
`...`	further arguments to be passed to `plot()`

Details

First, the algorithm determines the status of the subjects at time `predict.time`. Those censored subjects whose status could be not be determined (therefore, with missing status) are excluded from the analysis. Then, the optimum threshold is estimated by minimizing the cost function using the `thres2` function in the `ThresholdROC` package (Perez-Jaume et al, 2017).

For parameter `method` the user can choose between `"normal"` (assumes binormality) or `"empirical"` (leaves out any distributional assumption). When `method="normal"`, the user can specify if the algorithm should assume equal or different variances using the parameter `var.equal`. For further details see the `thres2` function in the `ThresholdROC` package.

R, mentioned in parameters `costs` and `R`, is the product of the non-disease odds and the cost ratio:

$$R = ((1-p)/p)((C_{TN} - C_{FP})/(C_{TP} - C_{FN})),$$

where p is the disease prevalence (estimated using Kaplan-Meier) and C_i are the classification costs.

To calculate sensitivity, specificity and predictive values corresponding to the estimated threshold, we suggest to use the `diagnostic` function in the `ThresholdROC` package.

Value

An object of class `thres2`, which is a list of two components (see the help on the `thres2` function). Here we add a third component, `data`: a data.frame containing the following columns previously provided by the user: `cont.var`, `time` and `status`, and a new column `statusNA`, which contains the status of the subjects at time `predict.time` (0=no event, 1=event, `NA`=missing)

References

Perez-Jaume S, Skaltsa K, Pallares N, Carrasco JL. ThresholdROC: Optimum Threshold Estimation Tools for Continuous Diagnostic Tests in R. Journal of Statistical Software 2017; 82(4): 1-21. doi: 10.18637/jss.v082.i04

Skaltsa K, Jover L, Carrasco JL. Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. Biometrical Journal 2010; 52(5): 676-697. doi: 10.1002/bimj.200900294

See Also

`thres2`

Examples

```
data(NSCLC)
res <- with(NSCLC, th_ME(log(COL), OS, ST, 1095, method="normal",
            var.equal=FALSE, plot=TRUE, xlab="Collagen"))
res
exp(res$T$thres)
exp(res$CI$lower)
exp(res$CI$upper)
res$data
```

th_survivalROC *Threshold estimation using the method based on time-dependent ROC curves (survivalROC)*

Description

This function estimates optimum thresholds with survival data by using a method based on time-dependent ROC curves

Usage

```
th_survivalROC(cont.var, time, status, predict.time, costs = NULL,
               R = NULL, lambda = 0.05, plot = FALSE, ci = FALSE,
               alpha = 0.05, B = 1000, ...)
```

Arguments

cont.var	continuous variable or biomarker to be used as predictor of the status
time	survival time
status	censoring status codified as 0=censored, 1=event
predict.time	time-point of interest to define the subjects' status as event present or absent
costs	cost matrix. Costs should be entered as a 2x2 matrix, where the first row corresponds to the true positive and true negative costs and the second row to the false positive and false negative costs. Default cost values (costs=NULL, when also R=NULL) are a combination of costs that yields R=1, which is equivalent to the Youden index method (for details about this concept, see **References**)
R	if the cost matrix costs is not set (that is, costs=NULL), R desired (the algorithm will choose a suitable combination of costs that leads to R). Default, NULL (which leads to R=1 using the default costs).
lambda	smoothing parameter for the NNE algorithm used in survivalROC() function
plot	Should the cost function be plotted? Default, FALSE
ci	Should a confidence interval be calculated? Default, FALSE
alpha	significance level for the confidence interval. Default, 0.05
B	number of bootstrap resamples for the confidence interval. Default, 1000
...	further arguments to be passed to plot()

Details

This function estimates optimal thresholds by constructing the ROC curve at time t through time-dependent ROC curves (Heagerty et al, 2000).

Confidence intervals are obtained using normal and percentile bootstrap. In normal boostrap, the bootstrap is used to obtain an estimate of the standard error of the threshold estimate, and then the standard normal distribution is used for the confidence interval calculation. In percentile boostrap, B bootstrap resamples are generated and the threshold is estimated in all of them. Then, the confidence interval is calculated as the empirical 1-alpha/2 and 1+alpha/2 percentiles of the B bootstrap estimates.

Value

An object of class `th_survivalROC`, which is a list with the following components:

`T` A list of four elements:
`thres` threshold estimate.
`prev` disease prevalence used.
`costs` cost matrix.
`R` R term, the product of the non-disease odds and the cost ratio (for further details about this concept, see **References**).

`CI` A list of five elements (or `NULL` if `ci=FALSE`):
`lower.norm` the lower limit of the confidence interval using normal bootstrap.
`upper.norm` the upper limit of the confidence interval using normal bootstrap.
`se` the standard error.
`lower.perc` the lower limit of the confidence interval using percentile bootstrap.
`upper.perc` the upper limit of the confidence interval using percentile bootstrap.
`alpha` significance level provided by the user.
`ci.method` method used for the confidence intervals calculation.
`B` number of bootstrap resamples.

References

Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. Biometrics 2000; 56(2): 337-344. doi: 10.1111/j.0006-341X.2000.00337.x

Skaltsa K, Jover L, Carrasco JL. Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. Biometrical Journal 2010; 52(5): 676-697. doi: 10.1002/bimj.200900294

Heagerty PJ, Saha-Chaudhuri P (2013). survivalROC: Time-dependent ROC curve estimation from censored survival data. R package version 1.0.3. https://CRAN.R-project.org/package=survivalROC

Examples

```
with(NSCLC, th_survivalROC(COL, OS, ST, 1095,
            plot=TRUE, ci=TRUE, B=500, xlab="Collagen"))
```

Index