Package 'survivalsurrogate'

May 10, 2025

Type Package

Title Evaluate a Longitudinal Surrogate with a Censored Outcome

Version 1.0

Date 2025-05-05

Description Provides influence function-based methods to evaluate a longitudinal surrogate marker in a censored time-to-event outcome setting, with plug-in and targeted maximum likelihood estimation options. More details will be available in the future in: Agniel D and Parast L (2025+). ``Robust Evaluation of Longitudinal Surrogate Markers with Censored Data." Journal of the Royal Statistical Society: Series B, In press. A tutorial for this package can be found at <https://www.laylaparast.com/survivalsurrogate>.

License GPL

Imports stats, dplyr, magrittr, glue, mlr3, purrr, SparseM, rBeta2009, data.table, utils, rpart

Suggests mlr3learners

NeedsCompilation no

Author Denis Agniel [aut], Layla Parast [aut, cre]

Maintainer Layla Parast <parast@austin.utexas.edu>

Depends R (>= 3.5.0)

Contents

`estimate_R` *Estimates the proportion of the treatment effect explained*

Description

Estimates the proportion of the treatment effect on the censored primary outcome that is explained by the longitudinal surrogate marker, given the estimated influence functions for delta and delta.s.

Usage

```
estimate_R(delta_if, delta_s_if, delta = NULL, delta_s = NULL, se_type = "asymptotic",
n_boot = NULL, alpha = 0.05)
```

Arguments

`delta_if`	Influence function estimates for delta.
`delta_s_if`	Influence function estimates for delta.s.
`delta`	(Optional) Defaults to the mean of the delta influence function estimates.
`delta_s`	(Optional) Defaults to the mean of the delta.s influence function estimates.
`se_type`	Type of standard error estimation, choices are "asymptotic" or "bootstrap".
`n_boot`	(Optional unless se_type = "bootstrap") Number of boostrap samples.
`alpha`	(Optional) Alpha level used for confidence interval. If not provided, this is set to 0.05.

Value

A dataframe with the following components:

`estimate`	Estimates of the treatment effect (delta), residual treatment effect (delta.s), and proportion of treatment effect explained (R) by the longitudinal surrogate marker.
`se`	Estimated standard errors.
`ci_l`	Lower bound of the confidence intervals.
`ci_h`	Upper bound of the confidence intervals.

References

Agniel D and Parast L (2025+). "Robust Evaluation of Longitudinal Surrogate Markers with Censored Data." Journal of the Royal Statistical Society: Series B, In press.

Examples

```
data(exampledata)
names(exampledata)
library(glue)
library(rpart)
library(mlr3)
library(dplyr)
library(mlr3learners)

tt <- 5
```

```
t0 <- 4
yvars <- paste0('Y_', 0:tt)
lrnc <- glue('regr.rpart')
lrnb <- glue('classif.log_reg')


p_deltahat <- plugin_delta(
    data = exampledata,
    folds = 'ff',
    id = 'ID',
    x = 'X_0',
    g = 'G_0',
    a = paste0('A_', 0:tt),
    y = yvars,
    s = paste0('S_', 0:t0),
    binary_lrnr = lrn(lrnb, predict_type = 'prob'),
    cont_lrnr = lrn(lrnc),
    truncate_e = 0.005,
    verbose = FALSE
  )

p_deltahat_s <- plugin_delta_s(
    data = exampledata,
    folds = 'ff',
    id = 'ID',
    x = 'X_0',
    g = 'G_0',
    a = paste0('A_', 0:tt),
    y = yvars,
    s = paste0('S_', 0:t0),
    binary_lrnr = lrn(lrnb, predict_type = 'prob'),
    cont_lrnr = lrn(lrnc),
    t0=t0,
    truncate_pi = 0.005,
    truncate_e = 0.005,
    verbose = FALSE
  )


estimate_R(p_deltahat$if_data[[1]]$eif,
p_deltahat_s$if_data[[1]]$eif)
```

`exampledata` *Example dataset*

Description

Example dataset with longitudinal surrogate marker and censored primary outcome

Usage

```
data("exampledata")
```

Format

A data frame with 1000 observations on the following 21 variables. Here, the landmark time is 4 and the final time point is 5. The surrogate is measured at the following time points: 0,1,2,3,4.

`ID` Unique individual ID

`X_0` Baseline covariate vector.

`G_0` Treatment indicator vector.

`Y_0` Primary outcome variable at time 0.

`S_0` Surrogate marker value at time 0.

`Y_1` Primary outcome variable at time 1.

`S_1` Surrogate marker value at time 1.

`Y_2` Primary outcome variable at time 2.

`S_2` Surrogate marker value at time 2.

`Y_3` Primary outcome variable at time 3.

`S_3` Surrogate marker value at time 3.

`Y_4` Primary outcome variable at time 4.

`S_4` Surrogate marker value at time 4.

`Y_5` Primary outcome variable at time 5.

`ff` Fold number.

`A_0` Observation variable at time 0.

`A_1` Observation variable at time 1.

`A_2` Observation variable at time 2.

`A_3` Observation variable at time 3.

`A_4` Observation variable at time 4.

`A_5` Observation variable at time 5.

---

plugin_delta *Treatment effect estimation using plug-in estimator*

Description

Estimates the treatment effect using the plug-in estimator.

Usage

```
plugin_delta(data, folds, id, x, g, a = NULL, y, s, binary_lrnr = NULL,
             cont_lrnr = NULL, e = NULL, gamma1 = NULL, gamma0 = NULL,
             mu1 = NULL, mu0 = NULL, Q1 = NULL, Q0 = NULL,
             truncate_e = 1e-12, verbose = FALSE)
```

Arguments

data	A dataframe containing all necessary variables for estimation. The functions in this package require the dataframe to be in a specific form. Therefore, you will need to reformat your dataset as follows. At a minimum, you will need variables in your dataframe that indicate (1) the folds for crossfitting; (2) a unique observation identifier; (3) baseline covariates, if there are none you must have a variable with all values equal to 1 to provide to the argument x below; (4) a variable indicating treatment group which should be 1 for treatment and 0 for control; (5) a set of variables that contain the surrogate marker value at each time point up to and including the landmark time, denoted t0; (6) a set of variables that indicate observation status at each time point where the surrogate marker is measured, in addition to the final time which is denoted as t; (7) a set of variables that indicate primary outcome status at each time point where the surrogate marker is measured, in addition to the final time which is denoted as t. See the exampledata for an example where t0=4 and t=5.
folds	A vector defining crossfitting folds for nuisance estimation.
id	A string giving the name of the column with unique unit IDs (e.g., individual ID).
x	A character vector of covariate names to adjust for in nuisance estimation. At a minimum this must have one covariate equal to 1 for all individuals.
g	A string indicating the column name of the treatment indicator with 1 for treatment and 0 for control.
a	(Optional) A character vector of the column names indicating observation status at each time point. Specifically, A_t = 1 if the individual is still under observation (i.e., uncensored) at time t, meaning their outcome status could in principle be measured at that time. A value of 0 means the individual was censored prior to t. If not provided, assumed to be 1 for all time points.
y	A character vector of the column names indicating primary outcome status at each time point. Specifically, Y_t = 1 if the primary event (e.g., failure, relapse, death) has not yet occurred by time t, and the individual is still at risk. A value of 0 means the primary event occurred on or before time t.
s	A character vector of the column names indicating the surrogate marker values measured at each time point, where the value is NA if the individual is no longer observable at that time point.
binary_lrnr	Learner object or specification used for estimating binary nuisance components (e.g., propensity scores, censoring).
cont_lrnr	Learner object used for estimating continuous-valued outcome regressions.
e	(Optional) Column name or vector of propensity score estimates. If not provided, these will be estimated.
gamma1	(Optional) Vector of column names for censoring probabilities under treatment. If not provided, these will be estimated.
gamma0	(Optional) Vector of column names for censoring probabilities under control. If not provided, these will be estimated.
mu1	(Optional) Names of estimated hazards under treatment. If not provided, these will be estimated.
mu0	(Optional) Names of estimated hazards under control. If not provided, these will be estimated.

Q1	(Optional) Names of estimated conditional means under treatment. If not provided, these will be estimated.
Q0	(Optional) Names of estimated conditional means under control.If not provided, these will be estimated.
truncate_e	Numeric truncation level for propensity scores to avoid division by near-zero values. Default is 1e-12.
verbose	Logical; if TRUE, print progress messages.

Value

A dataframe with the following components:

plugin_est	Plug-in estimate of the treatment effect
plugin_se	Estimated standard error of the plug-in estimator, based on the influence function.
if_data	A nested data frame containing the influence function contributions for each observation.

References

Agniel D and Parast L (2025+). "Robust Evaluation of Longitudinal Surrogate Markers with Censored Data." Journal of the Royal Statistical Society: Series B, In press.

Examples

```
data(exampledata)
names(exampledata)
library(glue)
library(rpart)
library(mlr3)
library(dplyr)
library(mlr3learners)

tt <- 5
t0 <- 4
yvars <- paste0('Y_', 0:tt)
lrnc <- glue('regr.rpart')
lrnb <- glue('classif.log_reg')

p_deltahat <- plugin_delta(
    data = exampledata,
    folds = 'ff',
    id = 'ID',
    x = 'X_0',
    g = 'G_0',
    a = paste0('A_', 0:tt),
    y = yvars,
    s = paste0('S_', 0:t0),
    binary_lrnr = lrn(lrnb, predict_type = 'prob'),
    cont_lrnr = lrn(lrnc),
    truncate_e = 0.005,
    verbose = FALSE
  )
p_deltahat
```

plugin_delta_s *Residual treatment effect estimation using the plug-in estimator*

Description

Estimates the residual treatment effect the plug-in estimator for evaluating a longitudinal surrogate marker.

Usage

```
plugin_delta_s(data, folds, id, x, g, a = NULL, y, s, binary_lrnr = NULL, cont_lrnr
= NULL, t0 = length(s), e = NULL, gamma1 = NULL, gamma0 = NULL, mu1 = NULL, mu0 = NULL,
pi = NULL, pistar = NULL, Q1 = NULL, Q0 = NULL, truncate_e = 1e-12, truncate_pi = 1e-12,
se_type = "asymptotic", n_boot = NULL, alpha = 0.05, verbose = FALSE, retain_data
= FALSE)
```

Arguments

data	A dataframe containing all necessary variables for estimation. The functions in this package require the dataframe to be in a specific form. Therefore, you will need to reformat your dataset as follows. At a minimum, you will need variables in your dataframe that indicate (1) the folds for crossfitting; (2) a unique observation identifier; (3) baseline covariates, if there are none you must have a variable with all values equal to 1 to provide to the argument x below; (4) a variable indicating treatment group which should be 1 for treatment and 0 for control; (5) a set of variables that contain the surrogate marker value at each time point up to and including the landmark time, denoted t0; (6) a set of variables that indicate observation status at each time point where the surrogate marker is measured, in addition to the final time which is denoted as t; (7) a set of variables that indicate primary outcome status at each time point where the surrogate marker is measured, in addition to the final time which is denoted as t. See the exampledata for an example where t0=4 and t=5.
folds	A vector defining crossfitting folds for nuisance estimation.
id	A string giving the name of the column with unique unit IDs (e.g., individual ID).
x	A character vector of covariate names to adjust for in nuisance estimation. At a minimum this must have one covariate equal to 1 for all individuals.
g	A string indicating the column name of the treatment indicator with 1 for treatment and 0 for control.
a	(Optional) A character vector of the column names indicating observation status at each time point. Specifically, A_t = 1 if the individual is still under observation (i.e., uncensored) at time t, meaning their outcome status could in principle be measured at that time. A value of 0 means the individual was censored prior to t. If not provided, assumed to be 1 for all time points.
y	A character vector of the column names indicating primary outcome status at each time point. Specifically, Y_t = 1 if the primary event (e.g., failure, relapse, death) has not yet occurred by time t, and the individual is still at risk. A value of 0 means the primary event occurred on or before time t.

s	A character vector of the column names indicating the surrogate marker values measured at each time point, where the value is NA if the individual is no longer observable at that time point.
binary_lrnr	Learner object or specification used for estimating binary nuisance components (e.g., propensity scores, censoring).
cont_lrnr	Learner object used for estimating continuous-valued outcome regressions.
t0	Landmark time.
e	(Optional) Column name or vector of propensity score estimates. If not provided, these will be estimated.
gamma1	(Optional) Vector of column names for censoring probabilities under treatment. If not provided, these will be estimated.
gamma0	(Optional) Vector of column names for censoring probabilities under control. If not provided, these will be estimated.
mu1	(Optional) Names of estimated hazards under treatment. If not provided, these will be estimated.
mu0	(Optional) Names of estimated hazards under control. If not provided, these will be estimated.
pi	(Optional) Vector of column names giving estimated probabilities of observed surrogate values under observed treatment. If not provided, these will be estimated.
pistar	(Optional) Vector of column names giving estimated probabilities under the reference distribution. If not provided, these will be estimated.
Q1	(Optional) Names of estimated conditional means under treatment. If not provided, these will be estimated.
Q0	(Optional) Names of estimated conditional means under control.If not provided, these will be estimated.
truncate_e	Numeric truncation level for propensity scores to avoid division by near-zero values. Default is 1e-12.
truncate_pi	Truncation threshold for surrogate probabilities to avoid instability. Default is 1e-12.
se_type	Type of standard error estimation, choices are "asymptotic" or "bootstrap".
n_boot	(Optional unless se_type = "bootstrap") Number of boostrap samples.
alpha	Significance level for confidence intervals. Default is 0.05.
verbose	Logical; if TRUE, print progress messages.
retain_data	Logical; if TRUE, retains full dataset with estimated components in the output.

Value

A dataframe with the following components:

plugin_est	The plug-in estimate of the residual treatment effect at the landmark time.
plugin_se	Standard error of the estimate, based on influence function or bootstrap.
ci_l	Lower bound of the confidence interval.
ci_h	Upper bound of the confidence interval.
if_data	A list containing the dataset with the efficient influence function if retain_data = TRUE.

References

Agniel D and Parast L (2025+). "Robust Evaluation of Longitudinal Surrogate Markers with Censored Data." Journal of the Royal Statistical Society: Series B, In press.

Examples

```
data(exampledata)
names(exampledata)
library(glue)
library(rpart)
library(mlr3)
library(dplyr)
library(mlr3learners)

tt <- 5
t0 <- 4
yvars <- paste0('Y_', 0:tt)
lrnc <- glue('regr.rpart')
lrnb <- glue('classif.log_reg')


p_deltahat_s <- plugin_delta_s(
    data = exampledata,
    folds = 'ff',
    id = 'ID',
    x = 'X_0',
    g = 'G_0',
    a = paste0('A_', 0:tt),
    y = yvars,
    s = paste0('S_', 0:t0),
    binary_lrnr = lrn(lrnb, predict_type = 'prob'),
    cont_lrnr = lrn(lrnc),
    t0=t0,
    truncate_pi = 0.005,
    truncate_e = 0.005,
    verbose = FALSE
  )
p_deltahat_s
```

---

tmle_delta *Treatment effect estimation using TMLE*

Description

Estimates the treatment effect using targeted maximum likelihood estimation (TMLE).

Usage

```
tmle_delta(data, folds, id, x, g, a = NULL, y, s, binary_lrnr = NULL, cont_lrnr
= NULL, e = NULL, gamma1 = NULL, gamma0 = NULL, truncate_e = 1e-12, verbose
= FALSE)
```

Arguments

data	A dataframe containing all necessary variables for estimation. The functions in this package require the dataframe to be in a specific form. Therefore, you will need to reformat your dataset as follows. At a minimum, you will need variables in your dataframe that indicate (1) the folds for crossfitting; (2) a unique observation identifier; (3) baseline covariates, if there are none you must have a variable with all values equal to 1 to provide to the argument x below; (4) a variable indicating treatment group which should be 1 for treatment and 0 for control; (5) a set of variables that contain the surrogate marker value at each time point up to and including the landmark time, denoted t0; (6) a set of variables that indicate observation status at each time point where the surrogate marker is measured, in addition to the final time which is denoted as t; (7) a set of variables that indicate primary outcome status at each time point where the surrogate marker is measured, in addition to the final time which is denoted as t. See the exampledata for an example where t0=4 and t=5.
folds	A vector defining crossfitting folds for nuisance estimation.
id	A string giving the name of the column with unique unit IDs (e.g., individual ID).
x	A character vector of covariate names to adjust for in nuisance estimation. At a minimum this must have one covariate equal to 1 for all individuals.
g	A string indicating the column name of the treatment indicator with 1 for treatment and 0 for control.
a	(Optional) A character vector of the column names indicating observation status at each time point. Specifically, A_t = 1 if the individual is still under observation (i.e., uncensored) at time t, meaning their outcome status could in principle be measured at that time. A value of 0 means the individual was censored prior to t. If not provided, assumed to be 1 for all time points.
y	A character vector of the column names indicating primary outcome status at each time point. Specifically, Y_t = 1 if the primary event (e.g., failure, relapse, death) has not yet occurred by time t, and the individual is still at risk. A value of 0 means the primary event occurred on or before time t.
s	A character vector of the column names indicating the surrogate marker values measured at each time point, where the value is NA if the individual is no longer observable at that time point.
binary_lrnr	Learner object or specification used for estimating binary nuisance components (e.g., propensity scores, censoring).
cont_lrnr	Learner object used for estimating continuous-valued outcome regressions.
e	(Optional) Column name or vector of propensity score estimates. If not provided, these will be estimated.
gamma1	(Optional) Vector of column names for censoring probabilities under treatment. If not provided, these will be estimated.
gamma0	(Optional) Vector of column names for censoring probabilities under control. If not provided, these will be estimated.
truncate_e	Numeric truncation level for propensity scores to avoid division by near-zero values. Default is 1e-12.
verbose	Logical; if TRUE, print progress messages.

Value

A dataframe with the following components:

tmle_est	TMLE estimate of the treatment effect.
tmle_se	Estimated standard error of the TMLE estimator, based on the influence function.
if_data	A nested data frame containing the influence function contributions for each observation.

References

Agniel D and Parast L (2025+). "Robust Evaluation of Longitudinal Surrogate Markers with Censored Data." Journal of the Royal Statistical Society: Series B, In press.

Examples

```
data(exampledata)
names(exampledata)
library(glue)
library(rpart)
library(mlr3)
library(dplyr)
library(mlr3learners)

tt <- 5
t0 <- 4
yvars <- paste0('Y_', 0:tt)
lrnc <- glue('regr.rpart')
lrnb <- glue('classif.log_reg')


tml_deltahat <- tmle_delta(data = exampledata,
                           folds = 'ff',
                           id = 'ID',
                           x = 'X_0',
                           g = 'G_0',
                           a = paste0('A_', 0:tt),
                           y = yvars,
                           s = paste0('S_', 0:t0),
                           binary_lrnr = lrn(lrnb, predict_type = 'prob'),
                           cont_lrnr = lrn(lrnc),
                           truncate_e = 0.005,
                           verbose = FALSE)
tml_deltahat
```

---

tmle_delta_s *Residual treatment effect estimation using TMLE*

Description

Estimates the residual treatment effect using targeted maximum likelihood estimation (TMLE) for evaluating a longitudinal surrogate marker.

Usage

```
tmle_delta_s(data, folds, id, x, g, a = NULL, y, s, binary_lrnr = NULL, cont_lrnr
= NULL, t0 = length(s), e = NULL, gamma1 = NULL, gamma0 = NULL, pi = NULL,
pistar = NULL, truncate_e = 1e-12, verbose = FALSE, retain_data = FALSE)
```

Arguments

data	A dataframe containing all necessary variables for estimation. The functions in this package require the dataframe to be in a specific form. Therefore, you will need to reformat your dataset as follows. At a minimum, you will need variables in your dataframe that indicate (1) the folds for crossfitting; (2) a unique observation identifier; (3) baseline covariates, if there are none you must have a variable with all values equal to 1 to provide to the argument x below; (4) a variable indicating treatment group which should be 1 for treatment and 0 for control; (5) a set of variables that contain the surrogate marker value at each time point up to and including the landmark time, denoted t0; (6) a set of variables that indicate observation status at each time point where the surrogate marker is measured, in addition to the final time which is denoted as t; (7) a set of variables that indicate primary outcome status at each time point where the surrogate marker is measured, in addition to the final time which is denoted as t. See the exampledata for an example where t0=4 and t=5.
folds	A vector defining crossfitting folds for nuisance estimation.
id	A string giving the name of the column with unique unit IDs (e.g., individual ID).
x	A character vector of covariate names to adjust for in nuisance estimation. At a minimum this must have one covariate equal to 1 for all individuals.
g	A string indicating the column name of the treatment indicator with 1 for treatment and 0 for control.
a	(Optional) A character vector of the column names indicating observation status at each time point. Specifically, A_t = 1 if the individual is still under observation (i.e., uncensored) at time t, meaning their outcome status could in principle be measured at that time. A value of 0 means the individual was censored prior to t. If not provided, assumed to be 1 for all time points.
y	A character vector of the column names indicating primary outcome status at each time point. Specifically, Y_t = 1 if the primary event (e.g., failure, relapse, death) has not yet occurred by time t, and the individual is still at risk. A value of 0 means the primary event occurred on or before time t.
s	A character vector of the column names indicating the surrogate marker values measured at each time point, where the value is NA if the individual is no longer observable at that time point.
binary_lrnr	Learner object or specification used for estimating binary nuisance components (e.g., propensity scores, censoring).
cont_lrnr	Learner object used for estimating continuous-valued outcome regressions.
t0	Landmark time.
e	(Optional) Column name or vector of propensity score estimates. If not provided, these will be estimated.
gamma1	(Optional) Vector of column names for censoring probabilities under treatment. If not provided, these will be estimated.

gamma0	(Optional) Vector of column names for censoring probabilities under control. If not provided, these will be estimated.
pi	(Optional) Vector of column names giving estimated probabilities of observed surrogate values under observed treatment. If not provided, these will be estimated.
pistar	(Optional) Vector of column names giving estimated probabilities under the reference distribution. If not provided, these will be estimated.
truncate_e	Numeric truncation level for propensity scores to avoid division by near-zero values. Default is 1e-12.
verbose	Logical; if TRUE, print progress messages.
retain_data	Logical; if TRUE, the function return the full data and influence function values.

Value

A dataframe with the following components:

tmle_est	TMLE estimate of the residual treatment effect.
tmle_se	Estimated standard error using the empirical standard deviation of the influence function.
if_data	(Optional) Full dataset joined with estimated influence function contributions.

References

Agniel D and Parast L (2025+). "Robust Evaluation of Longitudinal Surrogate Markers with Censored Data." Journal of the Royal Statistical Society: Series B, In press.

Examples

```
data(exampledata)
names(exampledata)
library(glue)
library(rpart)
library(mlr3)
library(dplyr)
library(mlr3learners)

tt <- 5
t0 <- 4
yvars <- paste0('Y_', 0:tt)
lrnc <- glue('regr.rpart')
lrnb <- glue('classif.log_reg')

tml_deltahat_s <- tmle_delta_s(data = exampledata,
                               folds = 'ff',
                               id = 'ID',
                               x = 'X_0',
                               g = 'G_0',
                               a = paste0('A_', 0:tt),
                               y = yvars,
                               s = paste0('S_', 0:t0),
                               binary_lrnr = lrn(lrnb, predict_type = 'prob'),
                               cont_lrnr = lrn(lrnc),
                               t0=t0,
                               truncate_e = 0.005,
```

```
                                    verbose = FALSE)
    tml_deltahat_s
```

Index