Package 'TrialEmulation'

August 25, 2023

Title Causal Analysis of Observational Time-to-Event Data

Version 0.0.3.2

Description Implements target trial emulation methods to apply randomized clinical trial design and analysis in an observational setting. Using marginal structural models, it can estimate intention-to-treat and per-protocol effects in emulated trials using electronic health records. A description and application of the method can be found in Danaei et al (2013) <doi:10.1177/0962280211403603>.

License Apache License (>= 2)

URL https://causal-lda.github.io/TrialEmulation/, https://github.com/Causal-LDA/TrialEmulation

BugReports https://github.com/Causal-LDA/TrialEmulation/issues

Depends R (>= 3.5.0)

Imports broom (>= 0.7.10), checkmate, data.table (>= 1.9.8), formula.tools, mvtnorm, parglm, Rcpp, sandwich

Suggests knitr, rmarkdown, testthat (>= 3.0.0), withr

LinkingTo Rcpp

VignetteBuilder knitr

Config/testthat/edition 3

Encoding UTF-8

Language en-GB

LazyData true

LazyDataCompression xz

RoxygenNote 7.2.3

NeedsCompilation yes

Author Roonak Rezvani [aut] (<https://orcid.org/0000-0001-5580-5058>, Original package author),
Isaac Gravestock [aut, cre] (<https://orcid.org/0000-0003-0283-2065>),
Li Su [aut],
Medical Research Council (MRC) [fnd],
F. Hoffmann-La Roche AG [cph, fnd]

Maintainer Isaac Gravestock <isaac.gravestock@roche.com>

Repository CRAN

Date/Publication 2023-08-25 07:40:02 UTC

R topics documented:

`case_control_sampling_trials` *Case-control sampling from extended data*

Description

Case-control sampling from extended data

Usage

```
case_control_sampling_trials(
  data_prep,
  p_control = NULL,
  subset_condition,
  sort = FALSE
)
```

Arguments

`data_prep`	Result from `data_preparation()`.
`p_control`	Proportion of controls to select at each follow-up time of each trial.
`subset_condition`	Expression used to `subset()` the trial data before sampling.
`sort`	Sort data before sampling. This ensures results are identical between data prepared with `separate_files` TRUE and FALSE.

Value

A `data.frame` or a `split()` `data.frame` if `length(p_control) > 1`. An additional column containing sample weights will be added to the result. These can be included in the models fit with `trial_msm()`.

Examples

```
# If necessary reduce the number of threads for data.table
data.table::setDTthreads(2)

data("te_data_ex")
samples <- case_control_sampling_trials(te_data_ex, p_control = 0.01)
```

---

data_preparation *Prepare Sequence of Trial Data*

Description

This function takes the one row per time period per patient data and constructs a dataset with the records for each period in each trial. This considerably expands the size of the data. It takes into account the eligibility for each trial and calculates the weight models and the weights for each time period in the expanded data.

Usage

```
data_preparation(
  data,
  id = "id",
  period = "period",
  treatment = "treatment",
  outcome = "outcome",
  eligible = "eligible",
  outcome_cov = ~1,
  model_var = NULL,
  switch_n_cov = ~1,
  switch_d_cov = ~1,
  first_period = NA,
  last_period = NA,
  use_weight = FALSE,
  use_censor = FALSE,
  cense = NA,
  pool_cense = FALSE,
  cense_d_cov = ~1,
  cense_n_cov = ~1,
  eligible_wts_0 = NA,
  eligible_wts_1 = NA,
  where_var = NULL,
```

```
  data_dir,
  save_weight_models = FALSE,
  glm_function = "glm",
  chunk_size = 500,
  separate_files = FALSE,
  quiet = FALSE,
  ...
)
```

Arguments

data	A data.frame containing all the required columns.
id	Name of the data column for id feature Defaults to id
period	Name of the data column for period feature Defaults to period
treatment	Name of the data column for treatment feature Defaults to treatment
outcome	Name of the data column for outcome feature Defaults to outcome
eligible	Indicator of whether or not an observation is eligible to be expanded about Defaults to eligible
outcome_cov	A RHS formula with baseline covariates to adjust in final model
model_var	List of Variables of interest to be used in final model. Derived variables to use in outcome models. Typically assigned_treatment for ITT and per-protocol, and dose + dose^2 for as-treated.
switch_n_cov	A RHS formula for modelling probability of switching treatment. Used in the numerator of weight calculation. May use time_on_regime to include treatment duration.
switch_d_cov	A RHS formula for modelling probability of switching treatment. Used in the denominator of weight calculation. May use time_on_regime to include treatment duration.
first_period	First time period to include as trial baseline in expanded data
last_period	Last time period to include as trial baseline in expanded data
use_weight	Use weights in analysis. If FALSE then no weights will be calculated.
use_censor	Use censoring for per-protocol analysis - censor person-times once a person-trial stops taking the initial treatment value
cense	Censoring variable
pool_cense	Fit pooled or separate censoring models for those treated and those untreated at the immediately previous visit. (default is FALSE, separate numerator and denominator models for treatment groups)
cense_d_cov	A RHS formula for modelling probability of being censored. Used in the numerator of weight calculation.
cense_n_cov	A RHS formula for modelling probability of being censored. Used in the denominator of weight calculation.
eligible_wts_0	Eligibility criteria used in weights for model condition Am1 = 0
eligible_wts_1	Eligibility criteria used in weights for model condition Am1 = 1

where_var	List of variables used in where conditions used in subsetting the data used in final analysis (where_case), the variables not included in the final model
data_dir	Directory to model objects when save_weight_models=TRUE and expanded data as trial_i.csvs if separate_files = TRUE. If the specified directory does not exist it will be created. If the directory already contains trial files an error will occur, other files may be overwritten.
save_weight_models	Save weight models objects in data_dir.
glm_function	Which glm function to use for the final model from stats or parglm packages
chunk_size	Number of patients to process in one chunk when separate_files = TRUE
separate_files	Save expanded data in separate CSV files for each trial.
quiet	Don't print progress messages.
...	Additional arguments passed to glm_function. This may be used to specify initial parameter estimates or arguments to control. See stats::glm, parglm::parglm and parglm::parglm.control() for more information.

Details

The arguments chunk_size and separate_files allow for processing of large datasets that would not fit in memory once expanded. When separate_files = TRUE, the input data are processed in chunks of patients and saved into separate files for each trial starting period. These separate files can be sampled to create the dataset for the modelling.

Value

An object of class TE_data_prep, which can either be sampled from (case_control_sampling_trials) or directly used in a model (trial_msm). It contains the elements

data the expanded trial dataset for all trial periods. If separate=FALSE a data.table, if separate=TRUE a character vector with the file path of the expanded data as csv.

min_period the first trial period in the expanded data

max_period the last trial period in the expanded data

N the total number of observations in the expanded data

data_template a zero-row data.frame in the with the columns and attributes of the expanded data

switch_models a list of summaries of the models fitted for probability of switching treatment, if use_weight=TRUE

censor_models a list of summaries of the models fitted for probability of censoring treatment, if use_weight=TRUE

extract_baseline *Extract Baseline Observations*

Description

Extract Baseline Observations

Usage

```
extract_baseline(trial_file, baseline_file, quiet = TRUE)
```

Arguments

trial_file	Path to an expanded trial csv file
baseline_file	Path to csv to save baseline observations
quiet	Don't print progress messages.

Details

Reads `trial_file` and saves the observations with `followup_time == 0` to `baseline_file` csv.

Value

The file path of the csv if successful.

initiators *Initiators Analysis*

Description

An all-in-one analysis using a sequence of target trials. This provides a simplified interface to the main working functions `data_preparation()` and `trial_msm()`.

Usage

```
initiators(
  data,
  id = "id",
  period = "period",
  treatment = "treatment",
  outcome = "outcome",
  eligible = "eligible",
  outcome_cov = ~1,
  model_var = NULL,
  switch_n_cov = ~1,
```

```
  switch_d_cov = ~1,
  first_period = NA,
  last_period = NA,
  first_followup = NA,
  last_followup = NA,
  use_weight = FALSE,
  save_weight_models = FALSE,
  analysis_weights = c("asis", "unweighted", "p99", "weight_limits"),
  weight_limits = c(0, Inf),
  use_censor = FALSE,
  cense = NA,
  pool_cense = FALSE,
  cense_d_cov = ~1,
  cense_n_cov = ~1,
  include_followup_time = ~followup_time + I(followup_time^2),
  include_trial_period = ~trial_period + I(trial_period^2),
  eligible_wts_0 = NA,
  eligible_wts_1 = NA,
  where_var = NULL,
  where_case = NA,
  data_dir,
  glm_function = "glm",
  quiet = FALSE,
  ...
)
```

Arguments

`data`	A `data.frame` containing all the required columns.
`id`	Name of the data column for id feature Defaults to id
`period`	Name of the data column for period feature Defaults to period
`treatment`	Name of the data column for treatment feature Defaults to treatment
`outcome`	Name of the data column for outcome feature Defaults to outcome
`eligible`	Indicator of whether or not an observation is eligible to be expanded about Defaults to eligible
`outcome_cov`	A RHS formula with baseline covariates to adjust in final model
`model_var`	List of Variables of interest to be used in final model. Derived variables to use in outcome models. Typically `assigned_treatment` for ITT and per-protocol, and `dose + dose^2` for as-treated.
`switch_n_cov`	A RHS formula for modelling probability of switching treatment. Used in the numerator of weight calculation. May use `time_on_regime` to include treatment duration.
`switch_d_cov`	A RHS formula for modelling probability of switching treatment. Used in the denominator of weight calculation. May use `time_on_regime` to include treatment duration.
`first_period`	First time period to include as trial baseline in expanded data

`last_period` Last time period to include as trial baseline in expanded data

`first_followup` First follow-up period

`last_followup` Last follow-up period

`use_weight` Use weights in analysis. If `FALSE` then no weights will be calculated.

`save_weight_models`
Save weight models objects in `data_dir`.

`analysis_weights`
One of

- `"asis"`: use the weights as calculated
- `"p99"`: truncate weights at the 1st and 99th percentiles
- `"weight_limits"`: truncate weights at the values specified in `weight_limits`
- `"unweighted"`: set all analysis weights to 1, even with `use_weight = TRUE`

`weight_limits` Lower and upper limits to truncate weights, given as `c(lower, upper)`

`use_censor` Use censoring for per-protocol analysis - censor person-times once a person-trial stops taking the initial treatment value

`cense` Censoring variable

`pool_cense` Fit pooled or separate censoring models for those treated and those untreated at the immediately previous visit. (default is `FALSE`, separate numerator and denominator models for treatment groups)

`cense_d_cov` A RHS formula for modelling probability of being censored. Used in the numerator of weight calculation.

`cense_n_cov` A RHS formula for modelling probability of being censored. Used in the denominator of weight calculation.

`include_followup_time`
The model to include the follow up time of the trial (`followup_time`) in outcome model, specified as a RHS formula.

`include_trial_period`
The model to include the trial period (`trial_period`) in outcome model, specified as a RHS formula.

`eligible_wts_0` Eligibility criteria used in weights for model condition Am1 = 0

`eligible_wts_1` Eligibility criteria used in weights for model condition Am1 = 1

`where_var` List of variables used in where conditions used in subsetting the data used in final analysis (where_case), the variables not included in the final model

`where_case` List of where conditions used in subsetting the data used in final analysis

`data_dir` Directory to save model objects in.

`glm_function` Which glm function to use for the final model from `stats` or `parglm` packages

`quiet` Don't print progress messages.

`...` Additional arguments passed to `glm_function`. This may be used to specify initial parameter estimates or arguments to `control`. See stats::glm, parglm::parglm and `parglm::parglm.control()` for more information.

Details

If `model_var = NULL` the package will add some terms to the outcome model:

- if `use_censor = FALSE` and `use_weight = FALSE`, an as-treated analysis will be done the outcome model will have `~ dose + I(dose^2)` terms added
- if `use_censor = TRUE`, a per-protocol analysis will be done with an `~assigned_treatment` term added
- if `use_censor = FALSE` and `use_weight = TRUE`, an intention to treat analysis will be done with an `~assigned_treatment` term added

Value

Returns the result of `trial_msm()` on the expanded data. An object of class `TE_msm` containing

model a `glm` object

robust a list containing a coefficient summary table and the robust covariance `matrix`

`predict.TE_msm` *Predict Cumulative Incidence with Confidence Intervals*

Description

Predict Cumulative Incidence with Confidence Intervals

Usage

```
## S3 method for class 'TE_msm'
predict(
  object,
  newdata,
  predict_times,
  conf_int = TRUE,
  samples = 100,
  type = c("cum_inc", "survival"),
  ...
)
```

Arguments

`object`	Object from `trial_msm()` or `initiators()`.
`newdata`	Baseline trial data to predict cumulative incidence or survival for. If `newdata` contains rows with `followup_time > 0` these will be removed.
`predict_times`	Follow-up times to predict. Any times given in newdata will be ignored.
`conf_int`	Calculate a confidence interval using coefficient samples from a multivariate normal distribution based on the robust covariance matrix.

samples	The number of samples of the coefficients for prediction models.
type	Type of values to calculate. Either cumulative incidence ("cum_inc") or survival ("survival").
...	Further arguments passed to or from other methods.

Value

A list of three data frames containing the cumulative incidences for each of the assigned treatment options and the difference between them.

Examples

```
# If necessary set the number of `data.table` threads
data.table::setDTthreads(2)

data("te_model_ex")
predicted_ci <- predict(te_model_ex, predict_times = 0:30, samples = 10)

# Plot the cumulative incidence curves for each treatment
plot(predicted_ci[[1]]$followup_time, predicted_ci[[1]]$cum_inc,
  type = "l",
  xlab = "Follow-up Time", ylab = "Cumulative Incidence"
)
lines(predicted_ci[[1]]$followup_time, predicted_ci[[1]]$`2.5%`, lty = 2)
lines(predicted_ci[[1]]$followup_time, predicted_ci[[1]]$`97.5%`, lty = 2)

lines(predicted_ci[[2]]$followup_time, predicted_ci[[2]]$cum_inc, type = "l", col = 2)
lines(predicted_ci[[2]]$followup_time, predicted_ci[[2]]$`2.5%`, lty = 2, col = 2)
lines(predicted_ci[[2]]$followup_time, predicted_ci[[2]]$`97.5%`, lty = 2, col = 2)
legend("topleft", title = "Assigned Treatment", legend = c("0", "1"), col = 1:2, lty = 1)

# Plot the difference in cumulative incidence over follow up
plot(predicted_ci[[3]]$followup_time, predicted_ci[[3]]$cum_inc_diff,
  type = "l",
  xlab = "Follow-up Time", ylab = "Difference in Cumulative Incidence",
  ylim = c(-0.1, 0.1)
)
lines(predicted_ci[[3]]$followup_time, predicted_ci[[3]]$`2.5%`, lty = 2)
lines(predicted_ci[[3]]$followup_time, predicted_ci[[3]]$`97.5%`, lty = 2)
```

---

print.TE_weight_summary

Print a Weight Summary Object

Description

Print a Weight Summary Object

Usage

```
## S3 method for class 'TE_weight_summary'
print(x, full = TRUE, ...)
```

Arguments

x	print.TE_weight_summary object.
full	Print full or short summary.
...	Arguments passed to print.data.frame.

Value

No return value, only for printing.

---

summary.TE_data_prep *Summary methods*

Description

Print summaries of data and model objects produced by TrialEmulation.

Usage

```
## S3 method for class 'TE_data_prep'
summary(object, ...)

## S3 method for class 'TE_data_prep_sep'
summary(object, ...)

## S3 method for class 'TE_data_prep_dt'
summary(object, ...)

## S3 method for class 'TE_msm'
summary(object, ...)

## S3 method for class 'TE_robust'
summary(object, ...)
```

Arguments

object	Object to print summary
...	Additional arguments passed to print methods.

Value

No value, displays summaries of object.

te_data_ex *Example of a Prepared Data Object*

Description

A small example object from data_preparation used in examples. It is created with the following code:

Usage

```
te_data_ex
```

Format

An object of class `TE_data_prep_dt` (inherits from `TE_data_prep`) of length 5.

Details

```
dat <- trial_example[trial_example$id < 200, ]

te_data_ex <- data_preparation(
data = dat,
 outcome_cov = c("nvarA", "catvarA"),
 first_period = 260,
 last_period = 280
)
```

See Also

te_model_ex

te_model_ex *Example of a Fitted MSM Model Object*

Description

A small example object from trial_msm used in examples. It is created with the following code:

Usage

```
te_model_ex
```

Format

An object of class `TE_msm` of length 2.

Details

```
te_model_ex <- trial_msm(
 data = data_subset,
 outcome_cov = c("catvarA", "nvarA"),
 last_followup = 40,
 model_var = "assigned_treatment",
 include_followup_time = ~followup_time,
 include_trial_period = ~trial_period,
 use_sample_weights = FALSE,
 quiet = TRUE,
 glm_function = "glm"
 )
```

See Also

te_data_ex

---

trial_example *Example of longitudinal data for randomised trial emulation*

Description

A dataset containing the treatment, outcomes and other attributes of 503 patients for randomised trial emulation. See `vignette("Getting-Started")`.

Usage

```
trial_example
```

Format

A data frame with 48400 rows and 11 variables:

id patient identifier

eligible eligible for trial start in this period, 1=yes, 0=no

period time period

outcome indicator for outcome in this period, 1=event occurred, 0=no event

treatment indicator for receiving treatment in this period, 1=treatment, 0=no treatment

catvarA A categorical variable relating to treatment and the outcome

catvarB A categorical variable relating to treatment and the outcome

catvarC A categorical variable relating to treatment and the outcome

nvarA A numerical variable relating to treatment and the outcome

nvarB A numerical variable relating to treatment and the outcome

nvarC A numerical variable relating to treatment and the outcome

`trial_msm` *Fit the Marginal Structural Model for the Sequence of Trials*

Description

Fits a weighted pooled logistic regression model for the sequence of trials and calculates a robust covariance matrix using a sandwich estimator.

Usage

```
trial_msm(
  data,
  outcome_cov = ~1,
  model_var = NULL,
  first_followup = NA,
  last_followup = NA,
  use_weight = FALSE,
  analysis_weights = c("asis", "unweighted", "p99", "weight_limits"),
  weight_limits = c(0, Inf),
  use_censor = FALSE,
  include_followup_time = ~followup_time + I(followup_time^2),
  include_trial_period = ~trial_period + I(trial_period^2),
  where_case = NA,
  glm_function = c("glm", "parglm"),
  use_sample_weights = TRUE,
  quiet = FALSE,
  ...
)
```

Arguments

`data`	A `data.frame` containing all the required columns.
`outcome_cov`	A RHS formula with baseline covariates to adjust in final model
`model_var`	List of Variables of interest to be used in final model. Derived variables to use in outcome models. Typically `assigned_treatment` for ITT and per-protocol, and `dose + dose^2` for as-treated.
`first_followup`	First follow-up period
`last_followup`	Last follow-up period
`use_weight`	Use weights in analysis. If `FALSE` then no weights will be calculated.
`analysis_weights`	One of • `"asis"`: use the weights as calculated • `"p99"`: truncate weights at the 1st and 99th percentiles • `"weight_limits"`: truncate weights at the values specified in `weight_limits`

- "unweighted": set all analysis weights to 1, even with use_weight = TRUE

weight_limits — Lower and upper limits to truncate weights, given as c(lower, upper)

use_censor — Use censoring for per-protocol analysis - censor person-times once a person-trial stops taking the initial treatment value

include_followup_time — The model to include the follow up time of the trial (followup_time) in outcome model, specified as a RHS formula.

include_trial_period — The model to include the trial period (trial_period) in outcome model, specified as a RHS formula.

where_case — List of where conditions used in subsetting the data used in final analysis

glm_function — Which glm function to use for the final model from stats or parglm packages

use_sample_weights — Use sample weights in addition to IP weights. data must contain a column sample_weight. The weights used in the model are calculated as weight = weight * sample_weight.

quiet — Don't print progress messages.

... — Additional arguments passed to glm_function. This may be used to specify initial parameter estimates or arguments to control. See stats::glm, parglm::parglm and parglm::parglm.control() for more information.

Details

The model formula is constructed by combining the arguments outcome_cov, model_var, include_followup_time, and include_trial_period.

Value

Object of class TE_msm containing

model a glm object

robust a list containing a coefficient summary table and the robust covariance matrix

---

vignette_switch_data *Example of expanded longitudinal data for randomised trial emulation*

Description

This is the expanded dataset created in the vignette("Getting-Started") known as switch_data.

Usage

```
vignette_switch_data
```

Format

A data frame with 1939053 rows and 7 variables:

id patient identifier

trial_period trial start time period

followup_time follow up time within trial

outcome indicator for outcome in this period, 1=event occurred, 0=no event

treatment indicator for receiving treatment in this period, 1=treatment, 0=no treatment

assigned_treatment indicator for assigned treatment at baseline of the trial, 1=treatment, 0=no treatment

weight weights for use with model fitting

catvarA A categorical variable relating to treatment and the outcome

catvarB A categorical variable relating to treatment and the outcome

catvarC A categorical variable relating to treatment and the outcome

nvarA A numerical variable relating to treatment and the outcome

nvarB A numerical variable relating to treatment and the outcome

nvarC A numerical variable relating to treatment and the outcome

---

weights.TE_data_prep_dt

Extract Weights

Description

Extract Weights

Usage

```
## S3 method for class 'TE_data_prep_dt'
weights(object, ...)

## S3 method for class 'TE_data_prep_sep'
weights(object, ...)
```

Arguments

`object`	Object to extract weights from
`...`	Not used.

Value

Weights extracted from `object` as a numeric vector.

Index