Package ‘cohetsurr’

April 9, 2025

Type Package

Title Assessing Complex Heterogeneity in Surrogacy

Version 2.0

Date 2025-04-09

Author Rebecca Knowlton

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

Description Provides functions to assess and test for complex heterogeneity in the strength of a surrogate marker with respect to multiple baseline covariates, using both a parametric model and a semiparametric two-step model. More details are available in: Knowlton, R., Tian, L., & Parast, L. (2025). ``A General Framework to Assess Complex Heterogeneity in the Strength of a Surrogate Marker,'' Statistics in Medicine, 44(5), e70001 <doi:10.1002/sim.70001>. Additionally, provides functions to assess complex heterogeneity in the strength of a surrogate marker in observational settings using meta-learners, with options for different base learners. More details will be available in the future in: Knowlton, R., Parast, L. (2025) ``Assessing Surrogate Heterogeneity in Real World Data Using Meta-Learners.'' A tutorial for this package can be found at <https://www.laylaparast.com/cohetsurr>.

License GPL

Imports stats, matrixStats, mvtnorm, mgcv, grf

NeedsCompilation no

R topics documented:

boot.var *Performs bootstrap estimation procedures for the variance of the proportion of treatment effect explained, the omnibus test, and identifying a region above a treshold*

Description

Performs bootstrap estimation procedures for the variance of the proportion of treatment effect explained, the omnibus test, and identifying a region above a treshold

Usage

```
boot.var(data.control, data.treat, W.grid.expand, type, test = FALSE,
data.all = NULL, num.cov = NULL, results.for.test = NULL, threshold = NULL)
```

Arguments

data.control	dataframe containing data from the control group, specifically the outcome Y, the surrogate S, and the baseline covariates W
data.treat	dataframe containing data from the treamtent group, specifically the outcome Y, the surrogate S, and the baseline covariates W
W.grid.expand	expanded version of the W grid of baseline covariates, where each row is a specific combination of the covariates for which the estimates should be provided
type	options are "model", "two step", or "both"; specifies the estimation method that should be used for the proportion of treatment effect explained
test	TRUE or FALSE, if test for heterogeneity is wanted
data.all	dataframe containing data from the control and treamtent groups, specifically the outcome Y, the surrogate S, and the baseline covariates W
num.cov	number of baseline covariates in the matrix W
results.for.test	the grid of estimates for the proportion of treatment effect explained obtained prior to bootstrapping, needed for the omnibus test
threshold	threshold to flag regions where the estimated proportion of the treatment effect explained is at least that high

Value

A list is returned:

return.grid	grid of variance estimates for the overall treatment effect, the residual treatment effect, and the proportion of treatment effect explained as a function of the baseline covariates, W. If requested by user, includes regions flagged above the threshold.
pval	p-value(s) from the F test and the two step omnibus test for heterogeneity, depending on type argument

`complex.heterogeneity` *Estimates the proportion of treatment effect explained by the surrogate marker as a function of multiple baseline covariates.*

Description

Assesses complex heterogeneity in the utility of a surrogate marker by estimating the proportion of treatment effect explained by the surrogate marker as a function of multiple baseline covariates. Optionally, tests for evidence of heterogeneity overall and flags regions where the proportion of treatment effect explained is above a given threshold.

Usage

```
complex.heterogeneity(y, s, a, W.mat, type = "model", variance = FALSE,
test = FALSE, W.grid = NULL, grid.size = 4, threshold = NULL)
```

Arguments

`y`	y, the outcome
`s`	s, the surrogate marker
`a`	a, the treatment assignment with 1 indicating the treatment group and 0 indicating the control group
`W.mat`	matrix of baseline covariate observations, where the first column is W1, second columns is W2, and so on.
`type`	options are "model", "two step", or "both"; specifies the estimation method that should be used for the proportion of treatment effect explained
`variance`	TRUE or FALSE, if variance/standard error estimates are wanted
`test`	TRUE or FALSE, if test for heterogeneity is wanted
`W.grid`	grid for the baseline covariates W where estimation will be provided
`grid.size`	number of measures for each baseline covariate to include in the estimation grid, if one is not provided by the user directly
`threshold`	threshold to flag regions where the estimated proportion of the treatment effect explained is at least that high

Value

A list is returned:

`return.grid`	grid of estimates for the overall treatment effect, the residual treatment effect, and the proportion of treatment effect explained as a function of the baseline covariates, W. Includes variance estimates and regions flagged above the threshold, if specified by the user.
`pval`	p-value(s) from the F test and the two step omnibus test for heterogeneity, depending on type argument

Author(s)

Rebecca Knowlton

References

Knowlton, R., Tian, L., & Parast, L. (2025). A General Framework to Assess Complex Heterogeneity in the Strength of a Surrogate Marker. Statistics in Medicine, 44(5), e70001

Examples

```
data(exampledata)
names(exampledata)
complex.heterogeneity(y = exampledata$y,
                      s = exampledata$s,
                      a = exampledata$a,
                      W.mat = matrix(cbind(exampledata$w1, exampledata$w2), ncol = 2),
                      type = "model",
               W.grid = matrix(cbind(exampledata$w1.grid, exampledata$w2.grid),ncol=2))
```

exampledata *Example data*

Description

Example data

Usage

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

Format

A list with 7 elements representing 1000 observations from a treatment group and 1000 observations from a control group, and a grid of baseline covariate values at which to calculate estimates:

- `y` the outcome
- `s` the surrogate marker
- `a` the treatment assignment, where 1 indicates treatment and 0 indicates control
- `w1` the first baseline covariate of interest
- `w2` the second baseline covariate of interest
- `w1.grid` the grid of first baseline covariate values to provide estimates for
- `w2.grid` the grid of second baseline covariate values to provide estimates for

Examples

```
data(exampledata)
names(exampledata)
```

`obs.boot.var` *Calculate bootstrapped variance estimates in an observational setting.*

Description

Calculates bootstrapped variance estimates of delta, delta.s, and R.s, and optionally calculates p-values for identifying strong surrogates.

Usage

```
obs.boot.var(df.train, df.test, type, numeric_predictors, categorical_predictors,
  threshold, use.actual.control.S, gam.smoothers, tree.tuners)
```

Arguments

`df.train` A dataframe containing training data.

`df.test` A dataframe containing testing data.

`type` Options are "linear", "gam", "trees", or "all"; type of base learners to use.

`numeric_predictors` The column names in the dataframes that represent numeric baseline covariates.

`categorical_predictors` The column names in the dataframes that represent categorical baseline covariates.

`threshold` An optional threshold to test individuals for the null hypothesis that PTE > threshold.

`use.actual.control.S` TRUE or FALSE, if user prefers to use the actual observed values for the surrogate in the control group instead of predicting values from the base learners.

`gam.smoothers` A list of smoothing parameters to use for GAM base learners, so they are not retuned with bootstrapping iterations ("m1sp", "m0sp", "m1ssp", "m0ssp", "s0")

`tree.tuners` A list of tuning parameters to use for tree base learners, so they are not retuned with bootstrapping iterations ("m1sp", "m0sp", "m1ssp", "m0ssp", "s0")

Value

A dataframe is returned, which is the df.test argument with new columns appended for the estimated variances of delta, delta.s, and R.s, as well as p-values if a threshold is provided.

`obs.estimate.PTE` *Estimate the proportion of the treatment effect explained in an observational setting.*

Description

Fits base learners using the specified type of model on the training data, and uses those models to calculate delta, delta.s, and R.s on the testing set.

Usage

```
obs.estimate.PTE(df.train, df.test, type, numeric_predictors, categorical_predictors,
  use.actual.control.S, gam.smoothers, tree.tuners, want.smooth, want.tune)
```

Arguments

`df.train` A dataframe containing training data.

`df.test` A dataframe containing testing data.

`type` Options are "linear", "gam", "trees", or "all"; type of base learners to use.

`numeric_predictors`
The column names in the dataframes that represent numeric baseline covariates.

`categorical_predictors`
The column names in the dataframes that represent categorical baseline covariates.

`use.actual.control.S`
TRUE or FALSE, if user prefers to use the actual observed values for the surrogate in the control group instead of predicting values from the base learners.

`gam.smoothers` A list of smoothing parameters to use for GAM base learners, so they are not retuned with bootstrapping iterations ("m1sp", "m0sp", "m1ssp", "m0ssp", "s0")

`tree.tuners` A list of tuning parameters to use for tree base learners, so they are not retuned with bootstrapping iterations ("m1sp", "m0sp", "m1ssp", "m0ssp", "s0")

`want.smooth` TRUE or FALSE, if smoothing parameters for GAM should be saved

`want.tune` TRUE or FALSE, if tuning parameters for trees should be saved

Value

A list is returned:

`df.test` df.test argument with new columns appended for the estimates of delta, delta.s, and R.s

`smooth_params` A list of smoothing parameters used for GAM base learners ("m1sp", "m0sp", "m1ssp", "m0ssp", "s0")

`tuner_params` A list of tuning parameters used for tree base learners ("m1sp", "m0sp", "m1ssp", "m0ssp", "s0")

`obs.het.surr` *Estimate the proportion of the treatment effect explained by the surrogate marker as a function of multiple baseline covariates in an observational setting*

Description

Assesses surrogate heterogeneity in real world data by estimating the proportion of the treatment effect explained as a function of baseline covariates. Optionally tests individuals for strong surrogacy based on a threshold.

Usage

```
obs.het.surr(df.train, df.test, type, var.want = FALSE, threshold = NULL,
  use.actual.control.S = FALSE)
```

Arguments

`df.train`	A dataframe containing training data. Must have columns G (treatment assignment), S (surrogate marker), and Y (primary outcome), in addition to the baseline covariates of interest.
`df.test`	A dataframe containing testing data. Must contain the same baseline covariate columns as the training data.
`type`	Options are "linear", "gam", "trees", or "all"; type of base learners to use.
`var.want`	TRUE or FALSE, if variance estimates are wanted.
`threshold`	An optional threshold to test individuals for the null hypothesis that PTE > threshold; must have var.want = TRUE to return p-values.
`use.actual.control.S`	TRUE or FALSE, if user prefers to use the actual observed values for the surrogate in the control group instead of predicting values from the base learners.

Value

A dataframe is returned, which is the df.test argument with new columns appended for the estimates and corresponding variances of delta, delta.s, and R.s. If a threshold is specified, returns a p-value for the null hypothesis that PTE > threshold.

Author(s)

Rebecca Knowlton

References

Knowlton, R. and Parast, L. (2025) “Assessing Surrogate Heterogeneity in Real World Data Using Meta-Learners." Under Review.

Examples

```
data(obs_exampledata_train)
data(obs_exampledata_test)
obs.het.surr(df.train = obs_exampledata_train, df.test = obs_exampledata_test,
type = "linear", var.want = FALSE)
```

`obs_exampledata_test` *Example testing data for observational setting*

Description

Example testing data for observational setting

Usage

```
data("obs_exampledata_test")
```

Format

A data frame with 200 observations on the following 9 variables.

- `X1` a numeric baseline covariate of interest
- `X2` a numeric baseline covariate of interest
- `X3` a numeric baseline covariate of interest
- `X4` a numeric baseline covariate of interest
- `X5` a numeric baseline covariate of interest
- `X6` a numeric baseline covariate of interest
- `G` the treatment assignment, where 1 indicates treated and 0 indicates control
- `S` the surrogate marker
- `Y` the primary outcome

Examples

```
data(obs_exampledata_test)
names(obs_exampledata_test)
```

`obs_exampledata_train` *Example training data for observational setting*

Description

Example training data for observational setting

Usage

```
data("obs_exampledata_train")
```

Format

A data frame with 1800 observations on the following 9 variables.

- `X1` a numeric baseline covariate of interest
- `X2` a numeric baseline covariate of interest
- `X3` a numeric baseline covariate of interest
- `X4` a numeric baseline covariate of interest
- `X5` a numeric baseline covariate of interest
- `X6` a numeric baseline covariate of interest
- `G` the treatment assignment, where 1 indicates treated and 0 indicates control
- `S` the surrogate marker
- `Y` the primary outcome

Examples

```
data(obs_exampledata_train)
names(obs_exampledata_train)
```

`parametric.est` *Estimates the proportion of treatment effect explained as a function of multiple baseline covariates, W, using a parametric model*

Description

Estimates the proportion of treatment effect explained as a function of multiple baseline covariates, W, using a parametric model

Usage

```
parametric.est(data.control, data.treat, W.grid.expand)
```

Arguments

`data.control`	dataframe containing data from the control group, specifically the outcome Y, the surrogate S, and the baseline covariates W
`data.treat`	dataframe containing data from the treamtent group, specifically the outcome Y, the surrogate S, and the baseline covariates W
`W.grid.expand`	expanded version of the W grid of baseline covariates, where each row is a specific combination of the covariates for which the estimates should be provided

Value

grid of estimates of the proportion of treatment effect explained, the overall treatment effect, and the residual treatment effect for the given baseline covariate combinations

`two.step.est` *Estimates the proportion of treatment effect explained as a function of multiple baseline covariates, W, using a two step, semiparametric model*

Description

Estimates the proportion of treatment effect explained as a function of multiple baseline covariates, W, using a two step, semiparametric model

Usage

```
two.step.est(data.control, data.treat, W.grid.expand.function)
```

Arguments

`data.control`	dataframe containing data from the control group, specifically the outcome Y, the surrogate S, and the baseline covariates W
`data.treat`	dataframe containing data from the treamtent group, specifically the outcome Y, the surrogate S, and the baseline covariates W
`W.grid.expand.function`	expanded version of the W grid of baseline covariates, where each row is a specific combination of the covariates for which the estimates should be provided

Value

grid of estimates of the proportion of treatment effect explained, the overall treatment effect, and the residual treatment effect for the given baseline covariate combinations

Index