Package 'EScvtmle'

July 21, 2025

Type Package

Title Experiment-Selector CV-TMLE for Integration of Observational and RCT Data

Version 0.0.2

Maintainer Lauren Eyler Dang <lauren.eyler@berkeley.edu>

Description The experiment selector cross-validated targeted maximum likelihood estimator (ES-CVTMLE) aims to select the experiment that optimizes the bias-variance tradeoff for estimating a causal average treatment effect (ATE) where different experiments may include a randomized controlled trial (RCT) alone or an RCT combined with real-world data. Using cross-validation, the ES-CVTMLE separates the selection of the optimal experiment from the estimation of the ATE for the chosen experiment. The estimated bias term in the selector is a function of the difference in conditional mean outcome under control for the RCT compared to the combined experiment. In order to help include truly unbiased external data in the analysis, the estimated average treatment effect on a negative control outcome may be added to the bias term in the selector. For more details about this method, please see Dang et al. (2022) <doi:10.48550/arXiv.2210.05802>.

License GPL-3

URL https://github.com/Lauren-EylerDang/EScvtmle/tree/main

BugReports https://github.com/Lauren-EylerDang/EScvtmle/issues

Depends R (>= 4.2), SuperLearner (>= 2.0.28)

Imports origami (>= 1.0.5), dplyr (>= 1.0.8), tidyselect (>= 1.2.0), MASS (>= 7.3.54), stringr (>= 1.4.0), ggplot2 (>= 3.3.6), gridExtra (>= 2.3)

Suggests testthat (>= 3.0.0), knitr

Encoding UTF-8

LazyData true

RoxygenNote 7.2.1

Config/testthat/edition 3

NeedsCompilation no

Author Lauren Eyler Dang [cre, aut],
Maya Petersen [aut],
Mark van der Laan [aut]

Repository CRAN

Date/Publication 2023-01-05 18:30:02 UTC

Contents

`ES.cvtmle` *ES.cvtmle*

Description

This function runs the experiment-selector cross-validated targeted maximum likelihood estimator (ES-CVTMLE) (Dang et al. 2022) for selecting and analyzing an optimal experiment, where candidate experiments include a randomized controlled trial (RCT) with or without various real-world datasets (RWD).

Usage

```
ES.cvtmle(
  txinrwd,
  data,
  study,
  covariates,
  treatment_var,
  treatment,
  outcome,
  NCO = NULL,
  Delta = NULL,
  Delta_NCO = NULL,
  id = NULL,
  pRCT,
  V = 10,
  Q.SL.library,
  d.SL.library.RCT,
  d.SL.library.RWD,
  g.SL.library,
  Q.discreteSL,
  d.discreteSL,
```

```
  g.discreteSL,
  family,
  family_nco,
  fluctuation = "logistic",
  comparisons = list(c(1), c(1, 0)),
  adjustnco = FALSE,
  target.gwt = TRUE,
  bounds = NULL,
  cvControl = list(),
  MCsamp = 1000
)
```

Arguments

txinrwd	Whether active treatment is available in RWD (TRUE/FALSE). If FALSE, only the control arm of the RCT will be augmented with external data.
data	The dataset
study	Character name of variable indicating study participation (e.g. "S"). This variable should take a value of 1 for the RCT and should take a value of 0 for the RWD. Note that the code is currently set up only to handle two studies, but may be expanded to handle multiple studies in the future.
covariates	Vector of character names of covariates to be adjusted for (e.g. c("W1", "W2"))
treatment_var	Character name of treatment variable (e.g. "A")
treatment	Value of treatment variable that corresponds to the active treatment (e.g. "DrugName" or 1). All other values of the treatment variable are assumed to be control.
outcome	Character name of outcome variable (e.g. "Y"). If the outcome is a binary variable subject to censoring, censored observations should either be coded as NA or should be coded as 0 and a missingness indicator should be included (see parameter Delta below).
NCO	Character name of negative control outcome variable (e.g. "nco") or NULL if no NCO available. If the NCO is a binary variable subject to censoring, censored observations should either be coded as NA or should be coded as 0 and a missingness indicator should be included (see parameter Delta_NCO below).
Delta	Character name of a variable that is 0 if an observation was censored (missing outcome) and 1 otherwise. Missing outcomes may also be coded as NA, in which case a Delta variable will be added internally. If no missing outcomes, set Delta=NULL.
Delta_NCO	Character name of a variable that is 0 if the value of NCO is missing and 1 otherwise. Missing NCOs may also be coded as NA, in which case a Delta_NCO variable will be added internally. If no missing NCO or no NCO, set Delta_NCO=NULL.
id	ID variable for the independent unit
pRCT	The probability of randomization to treatment in the RCT
V	Number of cross-validation folds (default 10)
Q.SL.library	Candidate algorithms for SuperLearner estimation of outcome regressions

`d.SL.library.RCT`
: Candidate algorithms for SuperLearner estimation of missingness mechanism for RCT-only

`d.SL.library.RWD`
: Candidate algorithms for SuperLearner estimation of missingness mechanism for RCT+RWD

`g.SL.library`
: Candidate algorithms for SuperLearner estimation of treatment mechanism for combined RCT/RWD analysis

`Q.discreteSL`
: Should a discrete SuperLearner be used for estimation of outcome regressions? (TRUE/FALSE)

`d.discreteSL`
: Should a discrete SuperLearner be used for estimation of missingness mechanism? (TRUE/FALSE)

`g.discreteSL`
: Should a discrete SuperLearner be used for estimation of treatment mechanism? (TRUE/FALSE)

`family`
: Either "binomial" for binary outcomes or "gaussian" for continuous outcomes

`family_nco`
: Family for negative control outcome

`fluctuation`
: 'logistic' (default for binary and continuous outcomes), or 'linear' describing fluctuation for targeted maximum likelihood estimation (TMLE) updating. If 'logistic' with a continuous outcome, outcomes are scaled to (0,1) for TMLE targeting and then returned to the original scale for parameter estimation.

`comparisons`
: A list of the values of the study variable that you would like to compare. For example, if you have an RCT labeled S=1 and RWD labeled S=0, you would use comparisons = list(c(1),c(1,0)) to compare RCT only to RCT + RWD. The first element of comparisons must be c(1) for the RCT only.

`adjustnco`
: Should we adjust for the NCO as a proxy of bias in the estimation of the ATE of A on Y? (TRUE/FALSE). Default is FALSE.

`target.gwt`
: As in the tmle R package (Gruber & van der Laan, 2012), if target.gwt is TRUE, the treatment mechanism is moved from the denominator of the clever covariate to the weight when fitting the coefficient for TMLE updating. Default TRUE.

`bounds`
: Optional bounds for truncation of the denominator of the clever covariate. The default is c(5/sqrt(n)/log(n),1).

`cvControl`
: A list of parameters to control the cross-validation process for the SuperLearners. See ?SuperLearner for more details.

`MCsamp`
: Number of Monte Carlo samples from the estimated limit distribution to use to estimate quantile-based confidence intervals. Default 1000.

Details

The experiment selector cross-validated targeted maximum likelihood estimator (ES-CVTMLE) aims to select the experiment that optimizes the bias-variance tradeoff for estimating a causal average treatment effect where different experiments may include a randomized controlled trial (RCT) alone or an RCT combined with real-world data. Using cross-validation, the ES-CVTMLE separates the selection of the optimal experiment from the estimation of the ATE for the chosen experiment. In order to avoid positivity violations, the package internally trims RWD so that no baseline covariate values are not represented in the RCT if active treatment is not available in the RWD. The

estimated bias term in the selector is a function of the difference in conditional mean outcome under control for the RCT compared to the combined experiment. In order to help include truly unbiased external data in the analysis, the estimated average treatment effect on a negative control outcome may be added to the bias term in the selector by setting the parameter NCO to the character name of a negative control variable in the dataset. For more details about this method, please see Dang et al. (2022).

References:

Dang LE, Tarp JM, Abrahamsen TJ, Kvist K, Buse JB, Petersen M, van der Laan M (2022). A Cross-Validated Targeted Maximum Likelihood Estimator for Data-Adaptive Experiment Selection Applied to the Augmentation of RCT Control Arms with External Data. arXiv:2210.05802 [stat.ME]

Susan Gruber, Mark J. van der Laan (2012). tmle: An R Package for Targeted Maximum Likelihood Estimation. Journal of Statistical Software, 51(13), 1-35. URL <http://www.jstatsoft.org/v51/i13/>.

Value

Returns an object of class "EScvtmle" that is a list with the following components.

ATE Average treatment effect (ATE) point estimates for the ES-CVTMLE estimator using the estimated bias squared plus variance selector ("b2v") and for the selector that includes an estimate of the ATE on a negative control outcome (NCO) in the bias term of the selector ("ncobias") if an NCO is available.

foldATEs Average treatment effect (ATE) point estimates for each cross-validation fold of the ES-CVTMLE estimator using the estimated bias squared plus variance selector ("b2v") and for the selector that includes an estimate of the ATE on a negative control outcome (NCO) in the bias term of the selector ("ncobias") if an NCO is available.

g g is a list of the same length as comparisons where each element of the list is a vector of the denominator of the covariate in front of the residual in the efficient influence curve for all observations in the experiment described by that element of comparisons. Values of g close to 0 or 1 indicate practical near-positivity violations.

CI Estimated 95% confidence intervals for the average treatment effect estimates of the ES-CVTMLE estimator using the estimated bias squared plus variance selector ("b2v") and for the selector that includes an estimate of the ATE on a negative control outcome (NCO) in the bias term of the selector ("ncobias") if an NCO is available.

limitdistributionsample Monte Carlo samples for the average treatment effect estimates of the ES-CVTMLE estimator that are used to construct confidence intervals for the estimated bias squared plus variance selector ("b2v") and for the selector that includes an estimate of the ATE on a negative control outcome (NCO) in the bias term of the selector ("ncobias") if an NCO is available.

Var Estimated variance of the ES-CVTMLE average treatment effect estimator using the estimated bias squared plus variance selector ("b2v") and for the selector that includes an estimate of the ATE on a negative control outcome (NCO) in the bias term of the selector ("ncobias") if an NCO is available.

selected_byfold Vector noting which experiment from the list of comparisons was selected in each cross-validation fold of the ES-CVTMLE estimator using the estimated bias squared plus variance selector ("b2v") and for the selector that includes an estimate of the ATE on a negative control outcome (NCO) in the bias term of the selector ("ncobias") if an NCO is available.

proportionselected Proportion of all cross-validation folds in which real-world (external) data were included in the analysis for the ES-CVTMLE estimator using the estimated bias squared plus variance selector ("b2v") and for the selector that includes an estimate of the ATE on a negative control outcome (NCO) in the bias term of the selector ("ncobias") if an NCO is available.

Examples

```
data(wash)
#For unbiased external controls, use:
dat <- wash[which(wash$study %in% c(1,2)),]
dat$study[which(dat$study==2)]<-0
set.seed(2022)
results_rwd1 <- ES.cvtmle(txinrwd=TRUE,
                          data=dat, study="study",
                          covariates=c("aged", "sex", "momedu", "hfiacat"),
                          treatment_var="intervention", treatment=1,
                          outcome="laz", NCO="Nlt18scale",
                          Delta=NULL, Delta_NCO=NULL,
                          pRCT=0.5, V=5, Q.SL.library=c("SL.glm"),
                         g.SL.library=c("SL.glm"), Q.discreteSL=TRUE, g.discreteSL=TRUE,
                       family="gaussian", family_nco="gaussian", fluctuation = "logistic",
                      comparisons = list(c(1),c(1,0)), adjustnco = FALSE, target.gwt = TRUE)
print.EScvtmle(results_rwd1)
```

`plot.EScvtmle` *plot.EScvtmle*

Description

Plots fold-specific average treatment effect (ATE) estimates and a histogram of Monte Carlo sample ATE estimates used to construct confidence intervals.

Usage

```
## S3 method for class 'EScvtmle'
plot(x, ...)
```

Arguments

`x`	An object of class "EScvtmle"
`...`	Other arguments to plot

Value

Returns a graphical object of class "grob" that contains two side-by-side plots: one of the fold-specific average treatment effect estimates for all cross-validation folds (including information regarding which experiment was selected in each fold), and the other of a histogram of the Monte Carlo samples that are used to construct confidence intervals. If a negative control outcome (NCO) is available, the plots are for the selector that includes the estimated average treatment effect on the NCO in the bias estimate. If not, the plots are for the selector that uses the estimated bias squared plus the variance selector, without information from an NCO. For more information about the different selectors, the use of cross-validation, or the construction of confidence intervals for this method, please see Dang et al. (2022) <arXiv:2210.05802>.

`print.EScvtmle` *print.EScvtmle*

Description

Prints output from object produced by ES.cvtmle function

Usage

```
## S3 method for class 'EScvtmle'
print(x, ...)
```

Arguments

`x`	An object of class "EScvtmle"
`...`	Other arguments to print

Details

Prints the average treatment effect (ATE) point estimate and 95% confidence interval for the ES-CVTMLE estimator (object of class "EScvtmle") using the estimated bias squared plus variance experiment-selector. If a negative control outcome (NCO) is available, this function also prints the ATE point estimate and 95% confidence interval for the selector that includes the estimated ATE on the NCO in the bias term. See Dang et al. (2022) <arXiv:2210.05802> for more details.

Value

No return value. Called to print a summary of the results for objects of class "EScvtmle".

wash *WASH Benefits Bangladesh Dataset*

Description

This dataset was constructed from the publicly-available WASH Benefits Bangladesh cluster randomized controlled trial (RCT) dataset. The results of this trial were originally reported by Luby et al. (2018), and the original dataset may be downloaded from https:://osf.io.wvyn4/. The trial found no evidence of an effect of an intervention to improve sanitation, including construction of improved latrines, on child length-for-age Z-scores (laz). A subsequent re-analysis by Arnold et al. (2018) of the control arm of this dataset as an observational cohort did find an effect of having an improved latrine at baseline on child laz. The authors concluded that observational analyses of water, sanitation, and hygiene (WASH) interventions may suffer from unmeasured confounding. To demonstrate how conducting a small RCT combined with unbiased or biased observational data could prevent unmeasured confounding from influencing results, we construct the dataset for this software package, as follows. Study 1: A random sample of 150 "Sanitation" arm observations and 150 "Control" arm observations with complete information was taken from the overall RCT, with "study" variable set to 1. Study 2: A second random sample of 150 "Sanitation" arm observations and 150 "Control" arm observations with complete information was taken from the remaining RCT observations, with "study" variable set to 2 to mimic an unbiased external dataset. Study 3: From the "Control" arm observations not included in the study=1 sample, 150 observations who had improved latrines at baseline and 150 observations who did not have improved latrines at baseline were sampled, with "study" variable set to 3 to mimic a biased external dataset. The data contained in this file consist of all three "studies". Because this study was not set up to have a negative control outcome for length-for-age Z-score, the options were limited. We would like a variable that is associated with socioeconomic status (SES) because that is a likely cause of the unmeasured confounding highlighted by Arnold et al. (2018). We chose number of household members <=18 years old as an NCO, because prior studies have shown this variable to be associated with SES in Bangladesh (The World Bank, 2013), but it is unlikely to be affected by having an improved latrine. We scaled this variable to match the scale of the true outcome (length-for-age Z-score).

Usage

```
data(wash)
```

Format

An object of class "data.frame"

intervention For studies 1 and 2: 1 if participant was in the "Sanitation" arm and 0 if participant was in the "Control" arm. For study 3: 1 if participant's household had an improved latrine at baseline and 0 otherwise.

study Study variable indicating RCT sample or external dataset as described above.

laz Child length-for-age Z-score at 2 years post-baseline.

aged Child's age in days.

sex Child's sex.

momedu Mother's education level.

hfiacat Category of household food insecurity. Levels are "Food Secure", "Mildly Food Insecure", "Moderately to Severely Food Insecure".

Nlt18scale Scaled number of household members <= 18 years old.

Source

https://osf.io/wvyn4/

References

Luby SP, Rahman M, Arnold BF, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomised controlled trial. The Lancet Global Health. 2018;6(3):e302-e315. doi:10.1016/S2214-109X(17)30490-4

Arnold BF, Null C, Luby SP, Colford JM. Implications of WASH Benefits trials for water and sanitation – Authors' reply. The Lancet Global Health. 2018;6(6):e616-e617. doi:10.1016/S2214-109X(18)30229-8

The World Bank. (2013). Bangladesh Poverty Assessment: Assessing a Decade of Progress in Reducing Poverty, 2000-2010. Bangladesh Development Series. Paper No. 31. https://documents1.worldbank.org/curated/en/1(

Examples

```
data(wash)
#For unbiased external controls, use:
dat <- wash[which(wash$study %in% c(1,2)),]
dat$study[which(dat$study==2)]<-0
set.seed(2022)
results_rwd1 <- ES.cvtmle(txinrwd=TRUE,
                          data=dat, study="study",
                          covariates=c("aged", "sex", "momedu", "hfiacat"),
                          treatment_var="intervention", treatment=1,
                          outcome="laz", NCO="Nlt18scale",
                          Delta=NULL, Delta_NCO=NULL,
                          pRCT=0.5, V=5, Q.SL.library=c("SL.glm"),
                         g.SL.library=c("SL.glm"), Q.discreteSL=TRUE, g.discreteSL=TRUE,
                    family="gaussian", family_nco="gaussian", fluctuation = "logistic",
                  comparisons = list(c(1),c(1,0)), adjustnco = FALSE, target.gwt = TRUE)
print.EScvtmle(results_rwd1)
```

Index