Package 'speff2trial'

May 9, 2026

Type Package

Version 1.0.5

Title Semiparametric Efficient Estimation for a Two-Sample Treatment Effect

Author Michal Juraska <mjuraska@fredhutch.org>, with contributions from Peter B. Gilbert <pgilbert@scharp.org>, Xiaomin Lu <xlu2@phhp.ufl.edu>, Min Zhang <mzhangst@umich.edu>, Marie Davidian <davidian@stat.ncsu.edu>, and Anastasios A. Tsiatis <tsiatis@stat.ncsu.edu>

Maintainer Michal Juraska <mjuraska@fredhutch.org>

Description Performs estimation and testing of the treatment effect in a 2-group randomized clinical trial with a quantitative, dichotomous, or right-censored time-to-event endpoint. The method improves efficiency by leveraging baseline predictors of the endpoint. The inverse probability weighting technique of Robins, Rotnitzky, and Zhao (JASA, 1994) is used to provide unbiased estimation when the endpoint is missing at random.

License GPL-2

URL https://github.com/mjuraska/speff2trial

BugReports https://github.com/mjuraska/speff2trial/issues

Encoding UTF-8

LazyData true

Depends stats, leaps, survival

RoxygenNote 7.1.2

NeedsCompilation no

Repository CRAN

Date/Publication 2022-05-31 16:20:02 UTC

Contents

ACTG175 *AIDS Clinical Trials Group Study 175*

Description

ACTG 175 was a randomized clinical trial to compare monotherapy with zidovudine or didanosine with combination therapy with zidovudine and didanosine or zidovudine and zalcitabine in adults infected with the human immunodeficiency virus type I whose CD4 T cell counts were between 200 and 500 per cubic millimeter.

Usage

```
data(ACTG175)
```

Format

A data frame with 2139 observations on the following 27 variables:

`pidnum` patient's ID number

`age` age in years at baseline

`wtkg` weight in kg at baseline

`hemo` hemophilia (0=no, 1=yes)

`homo` homosexual activity (0=no, 1=yes)

`drugs` history of intravenous drug use (0=no, 1=yes)

`karnof` Karnofsky score (on a scale of 0-100)

`oprior` non-zidovudine antiretroviral therapy prior to initiation of study treatment (0=no, 1=yes)

`z30` zidovudine use in the 30 days prior to treatment initiation (0=no, 1=yes)

`zprior` zidovudine use prior to treatment initiation (0=no, 1=yes)

`preanti` number of days of previously received antiretroviral therapy

`race` race (0=white, 1=non-white)

`gender` gender (0=female, 1=male)

`str2` antiretroviral history (0=naive, 1=experienced)

`strat` antiretroviral history stratification (1='antiretroviral naive', 2='> 1 but $\leq$ 52 weeks of prior antiretroviral therapy', 3='> 52 weeks')

`symptom` symptomatic indicator (0=asymptomatic, 1=symptomatic)

`treat` treatment indicator (0=zidovudine only, 1=other therapies)

`offtrt` indicator of off-treatment before 96$\pm$5 weeks (0=no,1=yes)

`cd40` CD4 T cell count at baseline

`cd420` CD4 T cell count at 20±5 weeks

`cd496` CD4 T cell count at 96±5 weeks (=NA if missing)

`r` missing CD4 T cell count at 96±5 weeks (0=missing, 1=observed)

`cd80` CD8 T cell count at baseline

`cd820` CD8 T cell count at 20±5 weeks

`cens` indicator of observing the event in `days`

`days` number of days until the first occurrence of: (i) a decline in CD4 T cell count of at least 50 (ii) an event indicating progression to AIDS, or (iii) death.

`arms` treatment arm (0=zidovudine, 1=zidovudine and didanosine, 2=zidovudine and zalcitabine, 3=didanosine).

Details

The variable `days` contains right-censored time-to-event observations. The data set includes the following post-randomization covariates: CD4 and CD8 T cell count at 20±5 weeks and the indicator of whether or not the patient was taken off-treatment before 96±5 weeks.

References

Hammer SM, et al. (1996), "A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter.", *New England Journal of Medicine*, 335:1081–1090.

`modSearch` *Internal function for speff()*

Description

`modSearch` is used internally by `speff` to construct an optimal model for prediction of the study endpoint or estimation of the missingness mechanism.

Usage

```
modSearch(formula, x, y, endpoint, method, optimal, force.in,
          nvmax)
```

Arguments

`formula`	a formula object with the response on the left of the ~ operator, and the linear predictor on the right.
`x`	a matrix of at least two predictors
`y`	a response vector

endpoint	a character string specifying the type of the response variable; possible values are "quantitative" or "dichotomous".
method	a character string specifying the type of search technique used in the model selection procedure; possible values are "exhaustive", "forward", or "backward".
optimal	specifies the optimization criterion for model selection; possible values are "cp" for Mallow's Cp, "bic" for BIC, and "rsq" for R-squared.
force.in	a vector of indices to columns of the design matrix that should be included in each regression model.
nvmax	the maximum number of covariates considered for inclusion in a model.

See Also

speff

---

speff *Semiparametric efficient estimation and testing for a two-sample treatment effect with a quantitative or dichotomous endpoint*

Description

speff conducts estimation and testing of the treatment effect in a 2-group randomized clinical trial with a quantitative or dichotomous endpoint. The method is a special case of Robins, Rotnitzky, and Zhao (1994, JASA). It improves efficiency by leveraging baseline predictors of the endpoint. The method uses inverse probability weighting to provide unbiased estimation when the endpoint is missing at random.

Usage

```
speff(formula, endpoint=c("quantitative", "dichotomous"), data,
      postrandom=NULL, force.in=NULL, nvmax=9,
      method=c("exhaustive", "forward", "backward"),
      optimal=c("cp", "bic", "rsq"), trt.id, conf.level=0.95,
      missCtrl=NULL, missTreat=NULL, endCtrlPre=NULL,
      endTreatPre=NULL, endCtrlPost=NULL, endTreatPost=NULL)
```

Arguments

formula	a formula object with the response on the left of the ~ operator, and the linear predictor on the right. The linear predictor specifies baseline and postrandomization variables that are considered for inclusion by the automated procedure for selecting the best models predicting the endpoint, separately for each treatment group. Interactions and variable transformations might also be considered. If predicted values for the endpoint are entered explicitly by the user, the formula can be of the form response ~ 1.

endpoint	a character string specifying the type of the response variable. The option "quantitative" (default) classifies the response as quantitative, and the mean difference is the measure of the treatment effect, whereas "dichotomous" specifies a dichotomous response, and the log odds ratio is the measure of the treatment effect. Only the first character is necessary.
data	a data frame in which to interpret the variables named in the formula, postrandom, and trt.id.
postrandom	a character vector designating postrandomization covariates included in the formula (this argument allows to distinguish baseline from postrandomiation covariates).
force.in	a vector of indices to columns of the design matrix that should be included in each regression model.
nvmax	the maximum number of covariates considered for inclusion in a model. The default is 9.
method	specifies the type of search technique used in the model selection procedure carried out by the regsubsets function. "exhaustive" (default) performs the all-subsets selection, whereas "forward" and "backward" execute a forward or backward step-wise selection, respectively.
optimal	specifies the optimization criterion for model selection. The default is "cp", Mallow's Cp, which is equivalent to AIC. The other options are "bic" for BIC and "rsq" for R-squared.
trt.id	a character string specifying the name of the treatment indicator which can be a character or a numeric vector. The control and treatment group is defined by the alphanumeric order of labels used in the treatment indicator.
conf.level	the confidence level to be used for confidence intervals reported by summary.speff.
missCtrl	estimated probabilities of observing the endpoint based on pre- and postrandomization information in the control group
missTreat	estimated probabilities of observing the endpoint based on pre- and postrandomization information in the treatment group
endCtrlPre	predicted values of the endpoint using baseline information in the control group only
endTreatPre	predicted values of the endpoint using baseline information in the treatment group only
endCtrlPost	predicted values of the endpoint using baseline and postrandomization information in the control group
endTreatPost	predicted values of the endpoint using baseline and postrandomization information in the treatment group

Details

The treatment effect is represented by the mean difference or the log odds ratio for a quantitative or dichotomous endpoint, respectively. Estimates of the treatment effect that ignore baseline covariates (naive) are included in the output.

Using the automated model selection procedure performed by `regsubsets`, four optimal regression models are developed for the study endpoint. Initially, all baseline and postrandomization covariates specified in the formula are considered for inclusion by the model selection procedure carried out separately in each treatment group. The optimal models are used to construct predicted values of the endpoint. Subsequently, in each treatment group, another regression model is fitted that includes only baseline covariates that were selected in the previous optimization. Then predicted values of the endpoint are computed based on these models. If missingness occurs in the endpoint variable, the model selection procedure is additionally used to determine the optimal models for predicting whether a subject has an observed endpoint, separately in each treatment group.

The function `regsubsets` conducts optimization of linear regression models only. The following modification in the model selection is adopted for a dichotomous variable: initially, a logistic regression model is fitted with all baseline and postrandomization covariates included in the formula. Subsequently, an optimal model is selected by using a weighted linear regression with weights from the last iteration of the IWLS algorithm. The optimal model is then refitted by logistic regression.

Besides using the built-in model selection algorithms, the user has the option to explicitly enter predicted values of the endpoint as well as estimated probabilities of observing the endpoint if it is missing at random.

Value

`speff` returns an object of class `"speff"` which can be processed by `summary.speff` to obtain or print a summary of the results. An object of class `"speff"` is a list containing the following components:

`coef`	a matrix with estimates of treatment-specific mean responses and the treatment effect.
`cov`	a list with components `naive` and `speff`, each storing the covariance matrix of the estimated treatment-specific mean responses.
`varbeta`	a numeric vector of variance estimates of the naive and semiparametric treatment effect estimates.
`formula`	a list with components `control` and `treatment` containing formula objects for the optimal selected regression models. Set to `NULL` if predicted values are entered explicitly.
`rsq`	a numeric vector of the R-squared statistics for the optimal selected regression models predicting the study endpoint. Set to `NULL` if predicted values are entered by the user.
`endpoint`	`"quantitative"` for a quantitative and `"dichotomous"` for a dichotomous response.
`postrandom`	a character vector of postrandomization covariates considered for selection.
`predicted`	a logical vector; if `TRUE`, the built-in model selection procedure was employed for prediction of the study endpoint in the control and treatment group, respectively.
`conf.level`	confidence level of the confidence intervals reported by `summary.speff`.
`method`	search technique employed in the model selection procedure.
`n`	number of subjects in each treatment group.

References

Robins JM, Rotnitzky A, Zhao LP. (1994), "Estimation of regression coefficients when some regressors are not always observed.", *Journal of the American Statistical Association*, 89:846–66.

Tsiatis AA, Davidian M, Zhang M, Lu X. (2007), "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach.", *Statistics in Medicine*, 27:4658–4677.

Zhang M, Tsiatis AA, Davidian M. (2008), "Improving efficiency of inferences in randomized clinical trials using auxiliary covariates.", *Biometrics*, 64:707–715.

Davidian M, Tsiatis AA, Leon S. (2005), "Semiparametric estimation of treatment effect in a pretest-posttest study with missing data.", *Statistical Science*, 20:261–301.

Zhang M, Gilbert P. (2009), "Increasing the efficiency of prevential trials by incorporating baseline covariates.", manuscript.

See Also

summary.speff

Examples

```
str(ACTG175)

### treatment effect estimation with a quantitative endpoint missing
### at random
fit1 <- speff(cd496 ~ age+wtkg+hemo+homo+drugs+karnof+oprior+preanti+
race+gender+str2+strat+symptom+cd40+cd420+cd80+cd820+offtrt,
postrandom=c("cd420","cd820","offtrt"), data=ACTG175, trt.id="treat")

### 'fit2' adds quadratic effects of CD420 and CD820 and their
### two-way interaction
fit2 <- speff(cd496 ~ age+wtkg+hemo+homo+drugs+karnof+oprior+preanti+
race+gender+str2+strat+symptom+cd40+cd420+I(cd420^2)+cd80+cd820+
I(cd820^2)+cd420:cd820+offtrt, postrandom=c("cd420","I(cd420^2)",
"cd820","I(cd820^2)","cd420:cd820","offtrt"), data=ACTG175,
trt.id="treat")

### 'fit3' uses R-squared as the optimization criterion
fit3 <- speff(cd496 ~ age+wtkg+hemo+homo+drugs+karnof+oprior+preanti+
race+gender+str2+strat+symptom+cd40+cd420+cd80+cd820+offtrt,
postrandom=c("cd420","cd820","offtrt"), data=ACTG175, trt.id="treat",
optimal="rsq")

### a dichotomous response is created with missing values maintained
ACTG175$cd496bin <- ifelse(ACTG175$cd496 > 250, 1, 0)

### treatment effect estimation with a dichotomous endpoint missing
### at random
fit4 <- speff(cd496bin ~ age+wtkg+hemo+homo+drugs+karnof+oprior+preanti+
race+gender+str2+strat+symptom+cd40+cd420+cd80+cd820+offtrt,
postrandom=c("cd420","cd820","offtrt"), data=ACTG175, trt.id="treat",
endpoint="dichotomous")
```

speffSurv *Semiparametric efficient estimation and testing for a two-sample treatment effect with a right-censored time-to-event endpoint*

Description

speffSurv conducts estimation and testing of the treatment effect in a two-group randomized clinical trial with a right-censored time-to-event endpoint. It improves efficiency by leveraging baseline predictors of the endpoint.

Usage

```
speffSurv(formula, data, force.in=NULL, nvmax=9,
          method=c("exhaustive", "forward", "backward"),
          optimal=c("cp", "bic", "rsq"), trt.id,
          conf.level=0.95, fixed=FALSE)
```

Arguments

formula	a formula object with the response variable on the left of the ~ operator and the linear predictor on the right. The response is a survival object of class Surv. The linear predictor specifies baseline variables that are considered for inclusion by the automated procedure for selecting the best models predicting the endpoint. Interactions and variable transformations might also be considered.
data	a data frame in which to interpret the variables named in the formula and trt.id.
force.in	a vector of indices to columns of the design matrix that should be included in each regression model.
nvmax	the maximum number of covariates considered for inclusion in a model. The default is 9.
method	specifies the type of search technique used in the model selection procedure carried out by the regsubsets function. "exhaustive" (default) performs the all-subsets selection, whereas "forward" and "backward" execute a forward or backward step-wise selection, respectively.
optimal	specifies the optimization criterion for model selection. The default is "cp", Mallow's Cp, which is equivalent to AIC. The other options are "bic" for BIC and "rsq" for R-squared.
trt.id	a character string specifying the name of the treatment indicator which can be a character or a numeric vector. The control and treatment group is defined by the alphanumeric order of labels used in the treatment indicator.
conf.level	the confidence level to be used for confidence intervals reported by summary.speffSurv.
fixed	logical value; if FALSE (default), automated selection procedure is used for predicting the endpoint. Otherwise, *all* baseline variables specified in the formula are used.

Details

The treatment effect is represented by the (unadjusted) log hazard ratio for the treatment versus control group. The estimate of the treatment effect using the (unadjusted) proportional hazards model is included in the output.

Using the automated model selection procedure performed by `regsubsets`, two optimal linear regression models are developed to characterize the influence function of an estimator that is more efficient than the maximum partial likelihood estimator. The "efficient" influence function is searched in the space of influence functions that determine all regular and asymptotically linear estimators for the treatment effect (for definitions see, for example, Tsiatis, 2006). The space of influence functions has three components: the estimation space that characterizes all regular and asymptotically linear estimators that do not use baseline covariates. The other two subspaces, the randomization and censoring space, use baseline covariates to improve the efficiency in the estimation of the treatment effect (Lu, 2008). The automated model selection procedure is used to identify functions in the randomization and censoring space that satisfy a prespecified optimality criterion and that lead to efficiency gain by using baseline predictors of the outcome.

The user has the option to avoid the automated variable selection and, instead, use all variables specified in the formula for the estimation of the treatment effect. This is achieved by setting `fixed=TRUE`.

`speffSurv` does not allow missing values in the data.

Value

`speffSurv` returns an object of class "`speffSurv`" which can be processed by `summary.speffSurv` to obtain or print a summary of the results. An object of class "`speffSurv`" is a list containing the following components:

`beta`	a numeric vector with estimates of the treatment effect from the unadjusted proportional hazards model and the semiparametric efficient model using baseline covariates, respectively.
`varbeta`	a numeric vector of variance estimates for the treatment effect estimates in `beta`.
`formula`	a list with components `rndSpace` and `censSpace` containing formula objects for the optimal selected linear regression models that characterize the optimal elements in the randomization and censoring space, respectively. Set to `NULL` if `fixed=TRUE`.
`fixed`	a logical value; if `TRUE`, the efficient estimator utilizes all baseline covariates specified in the formula. Otherwise, the automated selection procedure is used to identify covariates that ensure optimality.
`conf.level`	confidence level of the confidence intervals reported by `summary.speffSurv`.
`method`	search technique employed in the model selection procedure.
`n`	number of subjects in each treatment group.

References

Lu X, Tsiatis AA. (2008), "Improving the efficiency of the log-rank test using auxiliary covariates.", *Biometrika*, 95:679–694.

Tsiatis AA. (2006), *Semiparametric Theory and Missing Data.*, New York: Springer.

See Also

summary.speffSurv

Examples

```
str(ACTG175)

data <- na.omit(ACTG175[ACTG175$arms==0 | ACTG175$arms==1, ])
data <- data[1:100, ]

### efficiency-improved estimation of log hazard ratio using
### baseline covariates
### 'fit1' coerces the use of all specified baseline covariates;
### automated selection procedure is skipped
fit1 <- speffSurv(Surv(days,cens) ~ cd40+cd80+age,
                  data=data, trt.id="arms", fixed=TRUE)


fit2 <- speffSurv(Surv(days,cens) ~ cd40+cd80+age+wtkg+drugs+karnof+z30+
preanti+symptom, data=data, trt.id="arms")
```

---

summary.speff *Summarizing results for semiparametric efficient estimation and testing for a 2-sample treatment effect*

Description

summary method for an object of class "speff".

Usage

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

Arguments

object	an object of class "speff".
...	for other methods.

Details

print.summary.speff prints a formatted summary of results. In the initial section, formulas for the optimal selected regression models are printed with pertaining R-squared statistics for each treatment group. Further, an inferential table is produced with point and interval estimates of the treatment effect, standard error estimates, and Wald test p-values using both the naive and covariate-adjusted estimation methods. At least five significant digits are printed.

Value

A list with the following components:

tab	inferential table for the treatment effect.
method	search technique employed in the model selection procedure.
rsq	a numeric vector of the R-squared statistics for the optimal selected regression models predicting the study endpoint.
formula	a list with components control and treatment containing formula objects for the optimal selected regression models.
predicted	a logical vector; if TRUE, the built-in model selection procedure was employed for prediction of the study endpoint in the control and treatment group, respectively.

See Also

speff

Examples

```
### from the example for 'speff':
fit1 <- speff(cd496 ~ age+wtkg+hemo+homo+drugs+karnof+oprior+preanti+
race+gender+str2+strat+symptom+cd40+cd420+cd80+cd820+offtrt,
postrandom=c("cd420","cd820","offtrt"), data=ACTG175, trt.id="treat")

summary(fit1)
```

---

summary.speffSurv *Summarizing results for semiparametric efficient estimation and testing for a two-sample treatment effect with a right-censored time-to-event endpoint*

Description

summary method for an object of class "speffSurv".

Usage

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

Arguments

object	an object of class "speffSurv".
...	for other methods.

Details

`print.summary.speffSurv` prints a formatted summary of results. In the initial section, right-sided formulas defining the optimal selected functions in the randomization and censoring space are printed. Further, an inferential table is generated with point and interval estimates of the log hazard ratio, standard error estimates, and Wald test p-values using both the proportional hazards and covariate-adjusted estimation methods. At least five significant digits are printed.

Value

A list with the following components:

`tab`	inferential table for the treatment effect.
`method`	search technique employed in the model selection procedure.
`formula`	a list with components `rndSpace` and `censSpace` containing formula objects for the optimal selected linear regression models that characterize the optimal elements in the randomization and censoring space, respectively. Set to `NULL` if `fixed=TRUE`.
`fixed`	a logical value; if `TRUE`, the efficient estimator utilized all baseline covariates specified in the formula. Otherwise, the automated selection procedure was used to identify covariates that ensure optimality.

See Also

`speffSurv`

Examples

```
### from the example for 'speffSurv':
data <- na.omit(ACTG175[ACTG175$arms==0 | ACTG175$arms==1, ])
data <- data[1:100, ]

fit1 <- speffSurv(Surv(days,cens) ~ cd40+cd80+age,
                  data=data, trt.id="arms", fixed=TRUE)

summary(fit1)
```

Index