Package 'gesttools'

December 8, 2020

Type Package

Title General Purpose G-Estimation for End of Study or Time-Varying Outcomes

Version 1.0.1

Author Daniel Tompsett, Stijn Vansteelandt, Oliver Dukes, Bianca De Stavola

Maintainer Daniel Tompsett <danieltompsettwork@gmail.com>

Description Provides a series of general purpose tools to perform g-estimation using the methods described in Sjolander and Vansteelandt (2016) <doi:10.1515/em-2015-0005> and Dukes and Vansteelandt <doi:10.1093/aje/kwx347>. The package allows for g-estimation in a wide variety of circumstances, including an end of study or time-varying outcome, and an exposure that is a binary, continuous, or a categorical variable with three or more categories. The package also supports g-estimation with time-varying causal effects and effect modification by a confounding variable.

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Imports DataCombine, tidyr, tibble, tidyselect, geeM, rsample, nnet, magrittr

URL https://github.com/danieltompsett/gesttools

BugReports https://github.com/danieltompsett/gesttools/issues

NeedsCompilation no

Repository CRAN

Date/Publication 2020-12-08 09:40:05 UTC

R topics documented:

dataexamples *Generate Simulated Example Datasets*

Description

The code simulates four datasets designed to demonstrate each of the four g-estimation functions of the package. These are used in the examples for each function in the user manual. Each dataset comprises of an outcome Y (time-varying or end of study), time-varying exposure A, time-varying confounder L, a baseline confounder U, and optionally a censoring indicator C over 3 time periods.

Usage

```
dataexamples(n = 1000, seed = seed, Censoring = FALSE)
```

Arguments

`n`	Number of individuals in the dataset.
`seed`	Random seed used for data generation.
`Censoring`	TRUE or FALSE indicator of whether to include a censoring indicator C. If `Censoring=TRUE`, data entries for A, Y, L and U are set to missing after censoring.

Value

Returns a list of four datasets labeled `datagest`, `datagestmult`, `datagestcat`, and `datagestmultcat`, designed to demonstrate the respective functions.

Examples

```
datas<-dataexamples(n=1000,seed=34567,Censoring=FALSE)
data<-datas$datagest
head(data,n=20)
#Multiple outcome data with censoring
datas<-dataexamples(n=100,seed=34567,Censoring=TRUE)
data<-datas$datagestmultcat
head(data,n=20)
```

FormatData *Formats Data Into Correct Form*

Description

Takes a dataset in long format and puts it into the required format for use with the g-estimation functions. Specifically it ensures there exists a data entry for each individual at each time period, by adding empty rows, and orders the dataset by time and identifier. It can also create variables for the exposure histories of all time-varying variables in the data.

Usage

```
FormatData(
  data,
  idvar,
  timevar,
  An,
  varying,
  Cn = NA,
  GenerateHistory = FALSE,
  GenerateHistoryMax = NA
)
```

Arguments

`data` A data frame in long format containing the data to be analysed.

`idvar` A character string specifying the name of of the variable specifying an individuals identifier.

`timevar` A character string specifying the name of the time variable. Note that time periods must be labeled as integers starting from 1 $(1, 2, \ldots)$.

`An` A character string specifying the name of the exposure variable

`varying` A vector of character strings specifying the names of the variables to be included in the analysis which are time-varying. Specifically the exposure, time-varying confounders and (if applicable) the time-varying outcome. If `Cn` is specified, it is added to `varying` automatically.

`Cn` Optional character string specifying the name of the censoring indicator if present.

`GenerateHistory` A TRUE or FALSE indicator. If set to TRUE, variables are generated corresponding to the lagged histories of all variables included in `varying`. These will be labeled as `LagVari` where `Var` is the variable name and `i` indicates how much the variable is lagged by. For example `LagAn2` is the value of `An`, 2 time periods prior. Note that `LagAn1` is not generated as this is automatically included in the g-estimation functions.

`GenerateHistoryMax` An optional positive integer specifying `GenerateHistory` to generate exposure histories up to `GenerateHistoryMax` time periods prior.

Details

Note that any variable in `varying` that is strictly categorical MUST be declared as an `as.factor()` variable. Binary or continuous variables should be declared as an `as.numeric()` variable.

Value

A data frame in long format with additional rows added as necessary. If `data` is already in the correct format then no additional rows will be added.

Examples

```
data<-dataexamples(n=1000,seed=3456,Censoring=TRUE)$datagest
#To demonstrate the function we
#Delete the third row, corresponding to the entry for ID 1 at time 3
data<-data[-3,]
datanew<-FormatData(data=data,idvar="id",timevar="time",An="A",
varying=c("A","L"),GenerateHistory=FALSE,GenerateHistoryMax=NA)
head(datanew)
#Note that the missing entry has been re-added,
#with missing values for A and L in the third row
#An example with lagged history of time varying variables created.
data<-dataexamples(n=1000,seed=3456,Censoring=TRUE)$datagestmultcat
datanew<-FormatData(data=data,idvar="id",timevar="time",An="A",
Cn="C",varying=c("A","L"),GenerateHistory=TRUE,GenerateHistoryMax=NA)
head(datanew)
```

---

gest *G-Estimation for an End of Study Outcome*

Description

Performs g-estimation of a structural nested mean model (SNMM), based on the outcome regression methods described in Sjolander and Vansteelandt (2016) and Dukes and Vansteelandt (2018). We expect a dataset that holds an end of study outcome that is either binary or continuous, time-varying and/or baseline confounders, and a time-varying exposure that is either binary or continuous.

Usage

```
gest(
  data,
  idvar,
  timevar,
  Yn,
  An,
  Ybin,
  Abin,
  Lny,
  Lnp,
  type = 1,
  Cn = NA,
  LnC = NA,
  ...
)
```

Arguments

`data` A data frame in long format containing the data to be analysed. See description for details.

`idvar` Character string specifying the name of the ID variable in the data.

`timevar` Character string specifying the name of the time variable in the data. Note that timevar must specify time periods as integer values starting from 1 (must not begin at 0).

`Yn` Character string specifying the name of the end of study outcome variable.

`An` Character string specifying the name of the time-varying exposure variable.

`Ybin` TRUE or FALSE indicator of whether the outcome is binary.

`Abin` TRUE or FALSE indicator of whether the exposure is binary. Note that if `Abin==TRUE` then the variable specified in `An` MUST be written as a numeric variable. taking values 0 or 1. If not use `gestcat`

`Lny` Vector of character strings specifying the names of the time-varying and/or baseline confounders to be included in the outcome model in quotations.

`Lnp` Vector of character strings specifying the names of the time-varying and/or baseline confounders to be included in the model calculating the propensity scores.

`type` Value from 1-4 specifying SNMM type to fit. See details.

`Cn` Optional character string specifying the name of the censoring indicator variable. The variable specified in Cn should be a numeric variable taking values 0 or 1, with 1 indicating censored.

`LnC` Vector of character strings specifying the names of the time-varying and/or baseline covariates to be used in the censoring score model to calculate the censoring weights. Note that any variable in `LnC` should also be in `Lnp` for the validity of the censoring and propensity weights.

`...` Additional arguments, currently not in use.

Details

Given a time-varying exposure variable, A_t and time-varying confounders, L_t measured over time periods $t = 1, \ldots, T$, and an end of study outcome Y measured at time $T + 1$, `gest` estimates the causal parameters ψ of a SNMM of the form

$$E(Y(\bar{a}_t, 0) - Y(\bar{a}_{t-1}, 0)|\bar{a}_{t-1}, \bar{l}_t) = \psi z_t a_t \; \forall \; t = 1, \ldots, T$$

if Y is continuous or

$$\frac{E(Y(\bar{a}_t, 0)|\bar{a}_{t-1}, \bar{l}_t)}{E(Y(\bar{a}_{t-1}, 0)|\bar{a}_{t-1}, \bar{l}_t)} = exp(\psi z_t a_t) \; \forall \; t = 1, \ldots, T$$

if Y is binary. The SNMMs form is defined by the parameter z_t, which can be controlled by the input `type` as follows

- `type=1` sets $z_t = 1$. This implies that ψ is the effect of exposure at any time t on Y.
- `type=2` sets $z_t = c(1, l_t)$, and adds affect modification by the first named variable in `Lny`, which we denote L_t. Now $\psi = c(\psi_0, \psi_1)$ where ψ_0 is the effect of exposure at any time t on Y when $l_t = 0$ for all t, modified by ψ_1 for each unit increase in l_t at all times t. Note that effect modification is currently only supported for binary (written as a numeric 0,1 vector) or continuous confounders.
- `type=3` allows for time-varying causal effects. It sets z_t to a vector of zeros of length T with a 1 in the t'th position. Now $\psi = c(\psi_1, \ldots, \psi_T)$ where ψ_t is the effect of A_t on Y.
- `type=4` allows for a time-varying causal effect that can be modified by the first named variable in `Lny`, that is it allows for both time-varying effects and effect modification. It sets z_t to a vector of zeros of length T with $c(1, l_t)$ in the t'th position. Now $\psi = (\underline{\psi_1}, \ldots, \underline{\psi_T})$ where $\underline{\psi_t} = c(\psi_{0t}, \psi_{1t})$. Here ψ_{0t} is the effect of exposure at time t on Y when $l_t = 0$ modified by ψ_{1t} for each unit increase in l_t. Note that effect modification is currently only supported for binary (written as a numeric 0,1 vector) or continuous confounders.

The data must be in long format, where we assume the convention that each row with `time=t` contains A_t, L_t and C_{t+1} and Y_{T+1}. Thus the censoring indicator for each row should indicate that a user is censored AFTER time t. The end of study outcome Y_{T+1} should be repeated on each row. If either A or Y are binary, they must be written as numeric vectors taking values either 0 or 1. The same is true for any covariate that is used for effect modification.

The data must be rectangular with a row entry for every individual for each exposure time 1 up to T. Data rows after censoring should be empty apart from the ID and time variables. This can be done using the function `FormatData`.

By default the censoring, propensity and outcome models include the exposure history at the previous time as an explanatory variable. One may consider also including all previous exposure and confounder history as variables in `Lny`,`Lnp`, and `LnC`, variables which can be generated using `FormatData`.

Censoring weights are handled as described in Sjolander and Vansteelandt (2016). Note that it is necessary that any variable included in `LnC` must also be in `Lnp`. Missing data not due to censoring are handled automatically by removing rows with missing data prior to fitting the model. If outcome models fail to fit, consider removing covariates from `Lny` but keeping them in `Lnp` to reduce collinearity issues.

Value

List of the fitted causal parameters of the posited SNMM. These are labeled as follows for each SNMM type, where `An` is set to the name of the exposure variable, i is the current time period, and `Lny[1]` is set to the name of the first confounder in `Lny`.

`type=1` `An`: The effect of exposure at any time t on outcome.

`type=2` `An`: The effect of exposure at any time t on outcome, when `Ln[1]` is set to zero.
`An:Ln[1]`: The effect modification by `Lny[1]`, the additional effect of A on Y for each unit increase in `Lny[1]`.

`type=3` `t=i.An`: The effect of exposure at time t=i on outcome.

`type=4` `t=i.An`: The effect of exposure at time t=i on outcome, when `Ln[1]` is set to zero.
`t=i.An:Ln[1]`: The effect modification by `Lny[1]`, the additional effect of A on Y at time t=i for each unit increase in `Lny[1]`.

References

Vansteelandt, S., & Sjolander, A. (2016). Revisiting g-estimation of the Effect of a Time-varying Exposure Subject to Time-varying Confounding, Epidemiologic Methods, 5(1), 37-56. <doi:10.1515/em-2015-0005>.

Dukes, O., & Vansteelandt, S. (2018). A Note on g-Estimation of Causal Risk Ratios, American Journal of Epidemiology, 187(5), 1079–1084. <doi:10.1093/aje/kwx347>.

Examples

```
datas<-dataexamples(n=1000,seed=123,Censoring=FALSE)
data=datas$datagest
idvar="id"
timevar="time"
Yn="Y"
An="A"
Ybin=FALSE
Abin=TRUE
Lny=c("L","U")
Lnp=c("L","U")
type=1
Cn=NA
LnC=NA
gest(data,idvar=idvar,timevar,Yn,An,Ybin,Abin,Lny,Lnp,type=1)

#Example with censoring
datas<-dataexamples(n=1000,seed=123,Censoring=TRUE)
data=datas$datagest
Cn="C"
LnC=c("L","U")
gest(data,idvar,timevar,Yn,An,Ybin,Abin,Lny,Lnp,Cn,LnC,type=3)
```

gestboot *Percentile Based Bootstrap Confidence Intervals*

Description

Generates percentile based confidence intervals for the causal parameters of a fitted SNMM. Bonferroni corrected confidence intervals are also reported for multiple comparisons.

Usage

```
gestboot(
  gestfunc,
  data,
  idvar,
  timevar,
  Yn,
  An,
  Ybin,
  Abin = NA,
  Lny,
  Lnp,
  type = 1,
  Cn = NA,
  LnC = NA,
  cutoff = NA,
  bn = 1000,
  alpha = 0.05,
  onesided = "twosided",
  seed = 123,
  ...
)
```

Arguments

`gestfunc` Name (without quotations) of the g-estimation function to run. One of `gest`, `gestmult`, `gestcat` or `gestmultcat`.

`data, idvar, timevar, Yn, An, Ybin, Abin, Lny, Lnp, type, Cn, LnC, cutoff`
Same arguments as in gest functions, to be input into gestfunc.

`bn` Number of bootstrapped datasets.

`alpha` Confidence level of confidence intervals.

`onesided` Controls the type of confidence interval generated. Takes one of three inputs, `"upper"` for upper one-sided confidence intervals, `"lower"` for lower one-sided confidence intervals, and `"twosided"` for two-sided confidence intervals. Defaults to `"twosided"`.

`seed` Integer specifying the random seed for generation of bootstrap samples.

`...` additional arguments.

Value

Returns a list of the following four elements.

`original` The value of the causal parameters estimated on the original data `data`.

`mean.boot` The average values of the causal parameters estimated on the bootstrapped datasets.

`conf` The upper and/or lower bounds of $1 - \alpha$ confidence intervals for each element of ψ. For example, if `type=2`, and $\psi = (\psi_0, \psi_1)$, a separate confidence interval is fitted for ψ_0 and ψ_1.

`conf.Bonferroni`
The upper and/or lower bounds of Bonferroni corrected confidence intervals for ψ, used for multiple comparisons.

Examples

```
datas<-dataexamples(n=500,seed=123,Censoring=FALSE)
data=datas$datagest
idvar="id"
timevar="time"
Yn="Y"
An="A"
Ybin=FALSE
Abin=TRUE
Lny=c("L","U")
Lnp=c("L","U")
gestfunc<-gest
type=2
bn=5
alpha=0.05
Cn<-NA
LnC<-NA
gestboot(gest,data,idvar,timevar,Yn,An,Ybin,Abin,Lny,
Lnp,type=1,bn=bn,alpha=alpha,onesided="twosided",seed=123)
```

gestcat *G-Estimation for an End of Study Outcome and Categorical Exposure Variable*

Description

Performs g-estimation of a structural nested mean model (SNMM), based on the outcome regression methods described in Sjolander and Vansteelandt (2016) and Dukes and Vansteelandt (2018). We expect a dataset with an end of study outcome that is either binary or continuous, time-varying and/or baseline confounders, and a categorical time-varying exposure of two of more categories.

Usage

```
gestcat(
  data,
  idvar,
  timevar,
  Yn,
  An,
  Ybin,
  Lny,
  Lnp,
  type = 1,
  Cn = NA,
  LnC = NA,
  ...
)
```

Arguments

`data` A data frame in long format containing the data to be analysed. See description for details.

`idvar` Character string specifying the name of the ID variable in data.

`timevar` Character string specifying the name of the time variable in the data. Note that timevar must specify time periods as integer values starting from 1 (must not begin at 0).

`Yn` Character string specifying the name of the end of study outcome variable.

`An` Character string specifying the name of the time-varying exposure variable.

`Ybin` TRUE or FALSE indicator of whether the outcome is binary.

`Lny` Vector of character strings specifying the names of the variables to be included in the outcome model in quotations.

`Lnp` Vector of character strings specifying the names of the variables to be included in the model calculating the propensity scores.

`type` Value from 1-4 specifying SNMM type to fit. See details.

`Cn` Optional character string specifying the name of the censoring indicator variable. The variable specified in Cn should be a numeric vector taking values 0 or 1, with 1 indicating censored.

`LnC` Vector of character strings specifying the names of the variables to be used in the censoring score model to calculate the censoring weights. Note that any variable in `LnC` should also be in `Lnp` for the validity of the censoring and propensity weights.

`...` Additional arguments, currently not in use.

Details

Suppose a set of time varying confounders L_t, and a time-varying categorical exposure variable A_t, measured over time periods $t = 1, \ldots, T$, with an end of study outcome Y measured at time $T + 1$. Also suppose that A_t holds data consisting of $k \geq 2$ categories. These categories may take any arbitrary list of names, but we assume for theory purposes they are labeled as $j = 0, 1 \ldots, k$, where $j = 0$ denotes no exposure, or some reference category. Define binary variables A_t^j $j = 0, 1, \ldots, k$ where $A_t^j = 1$ if $A_t = j$ and 0 otherwise. Then `gestcat` fits a SNMM of the form

$$E(Y(\bar{a}_t, a^0) - Y(\bar{a}_{t-1}, a^0)|\bar{a}_{t-1}, \bar{l}_t) = \sum_{j=1}^{k} \psi^j z_t a_t^j \ \forall \ t = 1, \ldots, T$$

if Y is continuous or

$$\frac{E(Y(\bar{a}_t, a^0)|\bar{a}_{t-1}, \bar{l}_t)}{E(Y(\bar{a}_{t-1}, a^0)|\bar{a}_{t-1}, \bar{l}_t)} = exp(\sum_{j=1}^{k} \psi^j z_t a_t^j) \ \forall \ t = 1, \ldots, T$$

if Y is binary. The SNNM fits a separate set of causal parameters ψ^j, for the effect of exposure at category j on outcome, compared to exposure at the reference category 0, for each non-reference category. The models form is defined by the parameter z_t, which can be controlled by the input `type` as follows

- `type=1` sets $z_t = 1$. This implies that ψ^j is now the effect of exposure when set to category j, (compared the reference category) at any time t on Y.
- `type=2` sets $z_t = c(1, l_t)$, and adds affect modification by the first named variable in `Lny`, which we denote L_t. Now $\psi^j = c(\psi_0^j, \psi_1^j)$ where ψ_0^j is the effect of exposure when set to category j, (compared the reference category) at any time t on Y when $l_t = 0$ for all t, modified by ψ_1^j for each unit increase in l_t for all t. Note that effect modification is currently only supported for binary or continuous confounders.
- `type=3` sets z_t to a vector of zeros of length T with a 1 in the t'th position. Now $\psi^j = c(\psi_1^j, \ldots, \psi_T^j)$ where ψ_t^j is the effect of exposure, when set to category j, at time t on Y.
- `type=4` allows for a time-varying causal effect that can be modified by the first named variable in `Lny`, that is it allows for both time-varying effects and effect modification. It sets z_t to a vector of zeros of length T with $c(1, l_t)$ in the t'th position. Now $\psi^j = (\underline{\psi_1^j}, \ldots, \underline{\psi_T^j})$ where $\underline{\psi_t^j} = c(\psi_{0t}^j, \psi_{1t}^j)$. Here ψ_{0t}^j is the effect of exposure when set to category j at time t on Y at $l_t = 0$, modified by ψ_{1t}^j for each unit increase in l_t. Note that effect modification is currently only supported for binary or continuous confounders.

The data must be in long format, where we assume the convention that each row with `time=t` contains A_t, L_t and C_{t+1} and Y_{T+1}. That is the censoring indicator for each row should indicate that a user is censored AFTER time t. The end of study outcome Y_{T+1} should be repeated on each row. If Y is binary, it must be written as a numeric vector taking values either 0 or 1. The same is true for any covariate that is used for effect modification.

The data must be rectangular with a row entry for every individual for each exposure time 1 up to T. Data rows after censoring should be empty apart from the ID and time variables. This can be done using the function `FormatData`.

By default the censoring, propensity and outcome models include the exposure history at the previous time as an explanatory variable. One may consider also including all previous exposure and confounder history as variables in `Lny`,`Lnp`, and `LnC` if necessary.

Censoring weights are handled as described in Sjolander and Vansteelandt (2016). Note that it is necessary that any variable included in `LnC` must also be in `Lnp`. Missing data not due to censoring are handled automatically by removing rows with missing data prior to fitting the model. If outcome models fail to fit, consider removing covariates from `Lny` but keeping them in `Lnp` to reduce collinearity issues, or consider the sparseness of the data.

Value

List of the fitted causal parameters of the posited SNMM. These are labeled as follows for each SNMM type, where An is set to the name of the exposure variable, i is the current time period, j is the category level, and Lny[1] is set to the name of the first confounder in `Lny`.

`type=1` Anj: The effect of exposure at category j at any time t on outcome.

`type=2` Anj: The effect of exposure at category j, at any time t on outcome, when Ln[1] is set to zero.
Anj:Ln[1]: The effect modification by Lny[1], the additional effect of A at category j on Y for each unit increase in Lny[1].

`type=3` t=i.Anj: The effect of exposure at category j, at time t=i on outcome.

`type=4` t=i.Anj: The effect of exposure at category j, at time t=i on outcome when Ln[1] is set to zero.

t=i.Anj:Ln[1]: The effect modification by Lny[1], the additional effect of A at category j at time t=i on Y, for each unit increase in Lny[1].

References

Vansteelandt, S., & Sjolander, A. (2016). Revisiting g-estimation of the Effect of a Time-varying Exposure Subject to Time-varying Confounding, Epidemiologic Methods, 5(1), 37-56. <doi:10.1515/em-2015-0005>.

Dukes, O., & Vansteelandt, S. (2018). A Note on g-Estimation of Causal Risk Ratios, American Journal of Epidemiology, 187(5), 1079–1084. <doi:10.1093/aje/kwx347>.

Examples

```
datas<-dataexamples(n=1000,seed=123,Censoring=FALSE)
data=datas$datagestcat
#A is a categorical variable with categories labeled 1,2 and 3, with 1 the
#reference category
idvar="id"
timevar="time"
Yn="Y"
An="A"
Ybin=FALSE
Lny=c("L","U")
Lnp=c("L","U")
Cn<-NA
LnC<-NA
type=NA

gestcat(data,idvar,timevar,Yn,An,Ybin,Lny,Lnp,type=1)

#Example with censoring
datas<-dataexamples(n=1000,seed=123,Censoring=TRUE)
data=datas$datagestcat
Cn="C"
LnC=c("L","U")
gestcat(data,idvar,timevar,Yn,An,Ybin,Lny,Lnp,type=3,Cn,LnC)
```

gestmult *G-Estimation for a Time-Varying Outcome*

Description

Performs g-estimation of a structural nested mean model (SNMM), based on the outcome regression methods described in Sjolander and Vansteelandt (2016) and Dukes and Vansteelandt (2018). We assume a dataset with a time-varying outcome that is either binary or continuous, time-varying and/or baseline confounders, and a time-varying exposure that is either binary or continuous.

Usage

```
gestmult(
  data,
  idvar,
  timevar,
  Yn,
  An,
  Ybin,
  Abin,
  Lny,
  Lnp,
  type = 1,
  Cn = NA,
  LnC = NA,
  cutoff = NA,
  ...
)
```

Arguments

`data` A data frame in long format containing the data to be analysed. See description for details.

`idvar` Character string specifying the name of the ID variable in data.

`timevar` Character string specifying the name of the time variable in the data. Note that timevar must specify time periods as integer values starting from 1 (must not begin at 0).

`Yn` Character string specifying the name of the time-varying outcome variable.

`An` Character string specifying the name of the time-varying exposure variable.

`Ybin` TRUE or FALSE indicator of whether the outcome is binary.

`Abin` TRUE or FALSE indicator of whether the exposure is binary.

`Lny` Vector of character strings specifying the names of the confounders to be included in the outcome model in quotations.

`Lnp` Vector of character strings specifying the names of the confounders to be included in the model calculating the propensity scores.

`type` Value from 1-4 specifying SNMM type to fit. See details.

`Cn` Optional character string specifying the name of the censoring indicator variable. The variable specified in Cn should be a numeric vector taking values 0 or 1, with 1 indicating censored.

`LnC` Vector of character strings specifying the names of the covariates to be used in the censoring score model to calculate the censoring weights. Note that any variable in `LnC` should also be in `Lnp` for the validity of the censoring and propensity weights.

`cutoff` An integer taking value from 1 up to T, where T is the maximum value of timevar. Instructs the function to estimate causal effects based only on exposures up to `cutoff` time periods prior to the outcome. See details.

`...` Additional arguments, currently not in use.

Details

Suppose a series of time periods $1, \ldots, T + 1$ whereby a time-varying exposure and confounder (A_t and L_t) are measured over times $t = 1, \ldots, T$ and a time varying outcome Y_s is measured over times $s = 2, \ldots, T + 1$. Define $c = s - t$ as the step length, that is the number of time periods separating an exposure measurement, and subsequent outcome measurement. By using the transform $t = s - c$, `gestmult` estimates the causal parameters ψ of a SNMM of the form

$$E\{Y_s(\bar{a}_{s-c}, 0) - Y_s(\bar{a}_{s-c-1}, 0)|\bar{a}_{s-c-1}, \bar{l}_{s-c}\} = \psi z_{sc} a_{s-c} \; \forall c = 1, \ldots, T \; and \; \forall s > c$$

if Y is continuous or

$$\frac{E(Y_s(\bar{a}_{s-c}, 0)|\bar{a}_{s-c-1}, \bar{l}_{s-c})}{E(Y_s(\bar{a}_{s-c-1}, 0)|\bar{a}_{s-c-1}, \bar{l}_{s-c})} = exp(\psi z_{sc} a_{s-c}) \; \forall c = 1, \ldots, T \; and \; \forall s > c$$

if Y is binary. The SNMMs form is defined by the parameter z_{sc}, which can be controlled by the input `type` as follows

- `type=1` sets $z_{sc} = 1$. This implies that ψ is now the effect of exposure at any time t on all subsequent outcome periods.
- `type=2` sets $z_{sc} = c(1, l_{s-c})$ and adds affect modification by the first named variable in `Lny`, which we denote L_t. Now $\psi = c(\psi_0, \psi_1)$ where ψ_0 is the effect of exposure at any time t on all subsequent outcome periods, when $l_{s-c} = 0$ at all times t, modified by ψ_1 for each unit increase in l_{s-c} at all times t. Note that effect modification is currently only supported for binary or continuous confounders.
- `type=3` can posit a time-varying causal effect for each value of c, that is the causal effect for the exposure on outcome c time periods later. We set z_{sc} to a vector of zeros of length T with a 1 in the $c = s - t$'th position. Now $\psi = c(\psi_1, \ldots, \psi_T)$ where $\psi_(c)$ is the effect of exposure on outcome c time periods later for all outcome periods $s > c$ that is A_{s-c} on Y_s.
- `type=4` allows for a time-varying causal effect that can be modified by the first named variable in `Lny`, that is it allows for both time-varying effects and effect modification. It sets z_{sc} to a vector of zeros of length T with $c(1, l_{s-c})$ in the $c = s - t$'th position. Now $\psi = (\underline{\psi_1}, \ldots, \underline{\psi_T})$ where $\underline{\psi_c} = c(\psi_{0c}, \psi_{1c})$. Here ψ_{0c} is the effect of exposure on outcome c time periods later, given $l_{s-c} = 0$ for all $s > c$, modified by ψ_{1c} for each unit increase in l_{s-c} for all $s > c$. Note that effect modification is currently only supported for binary or continuous confounders.

The data must be in long format, where we assume the convention that each row with `time=t` contains A_t, L_t and C_{t+1}, Y_{t+1}. That is the censoring indicator for each row should indicate that a user is censored AFTER time t and the outcome indicates the first outcome that occurs AFTER A_t and L_t are measured. For example, data at time 1, should contain A_1, L_1, Y_2, and optionally C_2. If either A or Y are binary, they must be written as numeric vectors taking values either 0 or 1. The same is true for any covariate that is used for effect modification.

The data must be rectangular with a row entry for every individual for each exposure time 1 up to T. Data rows after censoring should be empty apart from the ID and time variables. This can be done using the function `FormatData`.

By default the censoring, propensity and outcome models include the exposure history at the previous time as a variable. One may consider also including all previous exposure and confounder history as variables in `Lny`,`Lnp`, and `LnC` if necessary.

Censoring weights are handled as described in Sjolander and Vansteelandt (2016). Note that it is necessary that any variable included in `LnC` must also be in `Lnp`. Missing data not due to censoring are automatically handled by removing rows with missing data prior to fitting the model. If outcome models fail to fit, consider removing covariates from `Lny` but keeping them in `Lnp` to reduce collinearity issues, or consider the sparseness of the data.

Value

List of the fitted causal parameters of the posited SNMM. These are labeled as follows for each SNMM type, where An is set to the name of the exposure variable, i is the current value of c, and Lny[1] is set to the name of the first confounder in `Lny`.

`type=1` An: The effect of exposure at any time t on outcome at all subsequent times.

`type=2` An: The effect of exposure on outcome at any time t, when Ln[1] is set to zero, on all subsequent outcome times.
An:Ln[1]: The effect modification by Lny[1], the additional effect of A on all subsequent Y for each unit increase in Lny[1] at all times t.

`type=3` c=i.An: The effect of exposure at any time t on outcome `c=i` time periods later.

`type=4` c=i.An: The effect of exposure at any time t on outcome `c=i` time periods later, when Ln[1] is set to zero.
c=i.An:Ln[1]: The effect modification by Lny[1], the additional effect of exposure on outcome c=i time periods later for each unit increase in Lny[1].

References

Vansteelandt, S., & Sjolander, A. (2016). Revisiting g-estimation of the Effect of a Time-varying Exposure Subject to Time-varying Confounding, Epidemiologic Methods, 5(1), 37-56. <doi:10.1515/em-2015-0005>.

Dukes, O., & Vansteelandt, S. (2018). A Note on g-Estimation of Causal Risk Ratios, American Journal of Epidemiology, 187(5), 1079–1084. <doi:10.1093/aje/kwx347>.

Examples

```
datas<-dataexamples(n=1000,seed=123,Censoring=TRUE)
data=datas$datagestmult
idvar="id"
timevar="time"
Yn="Y"
An="A"
Ybin=FALSE
Abin=TRUE
Lny=c("L","U")
Lnp=c("L","U")
Cn<-NA
LnC<-NA
type=NA

gestmult(data,idvar,timevar,Yn,An,Ybin,Abin,Lny,Lnp,type=1)

#Example with censoring
datas<-dataexamples(n=1000,seed=123,Censoring=TRUE)
data=datas$datagestmult
```

```
Cn=”C”
LnC=c(”L”,”U”)
gestmult(data,idvar,timevar,Yn,An,Ybin,Abin,Lny,Lnp,type=3,Cn,LnC,
cutoff=2)
```

gestmultcat *G-Estimation for a Time-Varying Outcome and Categorical Time-Varying Exposure*

Description

Performs g-estimation of a structural nested mean model (SNMM), based on the outcome regression methods described in Sjolander and Vansteelandt (2016) and Dukes and Vansteelandt (2018). We assume a dataset with a time-varying outcome that is either binary or continuous, time-varying and/or baseline confounders, and a categorical time-varying exposure of three or more categories.

Usage

```
gestmultcat(
  data,
  idvar,
  timevar,
  Yn,
  An,
  Ybin,
  Lny,
  Lnp,
  type = 1,
  Cn = NA,
  LnC = NA,
  cutoff = NA,
  ...
)
```

Arguments

`data` A data frame in long format containing the data to be analysed. See description for details.

`idvar` Character string specifying the name of the ID variable in data.

`timevar` Character string specifying the name of the time variable in the data. Note that timevar must specify time periods as integer values starting from 1 (must not begin at 0).

`Yn` Character string specifying the name of the time-varying outcome variable.

`An` Character string specifying the name of the time-varying exposure variable.

`Ybin` TRUE or FALSE indicator of whether the outcome is binary.

`Lny` Vector of character strings specifying the names of the confounders to be included in the outcome model in quotations.

`Lnp` Vector of character strings specifying the names of the confounders to be included in the model calculating the propensity scores.

`type` Value from 1-4 specifying SNMM type to fit. See details.

`Cn` Optional character string specifying the name of the censoring indicator variable. The variable specified in Cn should be a numeric vector taking values 0 or 1, with 1 indicating censored.

`LnC` Vector of character strings specifying the names of the covariates to be used in the censoring score model to calculate the censoring weights. Note that any variable in `LnC` should also be in `Lnp` for the validity of the censoring and propensity weights.

`cutoff` An integer taking value from 1 up to T, where T is the maximum value of timevar. Instructs the function to estimate causal effects based only on exposures up to `cutoff` time periods prior to outcomes. See details.

`...` Additional arguments, currently not in use.

Details

Suppose a series of time periods $1,\dots,T+1$ whereby a time-varying exposure and confounder (A_t and L_t) are measured over times $t=1,\dots,T$ and a time varying outcome Y_s is measured over times $s=2,\dots,T+1$. Define $c=s-t$ as the step length, that is the number of time periods separating an exposure measurement, and subsequent outcome measurement. Also suppose that $A_t=a_t$ is a categorical variable consisting of $k>2$ categories. These categories may take any arbitrary list of names, but we assume for theory purposes they are labeled as $j=0,1\dots,k$ where $j=0$ denotes no exposure, or some reference category. Define binary variables A^j_t $j=0,1,\dots,k$ where $A^j_t=1$ if $A_t=j$ and 0 otherwise. By using the transform $t=s-c$, `gestmultcat` estimates the causal parameters ψ of a SNMM of the form

$$E\{Y_s(\bar{a}_{s-c},a^0)-Y_s(\bar{a}_{s-c-1},a^0)|\bar{a}_{s-c-1},\bar{l}_{s-c}\}=\sum_{j=1}^{k}\psi^j z_{sc}a^j_{s-c}\ \forall c=1,\dots,T\ and\ \forall s>c$$

if Y is continuous or

$$\frac{E(Y_s(\bar{a}_{s-c},a^0)|\bar{a}_{s-c-1},\bar{l}_{s-c})}{E(Y_s(\bar{a}_{s-c-1},a^0)|\bar{a}_{s-c-1},\bar{l}_{s-c})}=exp(\sum_{j=1}^{k}\psi^j z_{sc}a^j_{s-c})\ \forall c=1,\dots,T\ and\ \forall s>c$$

if Y is binary. The SNNM fits a separate set of causal parameters ψ^j, for the effect of exposure at category j on outcome, compared to exposure at the reference category 0, for each non-reference category. The models form is defined by the parameter z_{sc}, which can be controlled by the input `type` as follows

- `type=1` sets $z_{sc}=1$. This implies that ψ^j is now the effect of exposure when set to category j, compared to when set to the reference category, at any time t on all subsequent outcome periods.

- type=2 sets $z_{sc} = c(1, l_{s-c})$ and adds affect modification by the first named variable in Lny, which we denote L_t. Now $\psi^j = c(\psi_0^j, \psi_1^j)$ where ψ_0^j is the effect of exposure when set to category j, compared to when set to the reference category, at any time t on all subsequent outcome periods when $l_{s-c} = 0$ for all t, modified by ψ_1^j for each unit increase in l_{s-c} at all times t. Note that effect modification is currently only supported for binary or continuous confounders.
- type=3 can posit a time-varying causal effect for each value of c, that is the causal effect for the exposure on outcome c time periods later. We set z_{sc} to a vector of zeros of length T with a 1 in the $c = s - t$'th position. Now $\psi^j = c(\psi_1^j, \ldots, \psi_T^j)$ where ψ_c^j is the effect of exposure, when set to category j, on outcome c time periods later for all $s > c$ that is A_{s-c}^j on Y_s for all $s > c$.
- type=4 allows for a time-varying causal effect that can be modified by the first named variable in Lny, that is it allows for both time-varying effects and effect modification. It sets z_{sc} to a vector of zeros of length T with $c(1, l_{s-c})$ in the $c = s-t$'th position. Now $\psi^j = (\underline{\psi_1^j}, \ldots, \underline{\psi_T^j})$ where $\underline{\psi_c^j} = c(\psi_{0c}^j, \psi_{1c}^j)$. Here ψ_{0c}^j is the effect of exposure when set to category j on outcome c time periods later, given $l_{s-c} = 0$, for all $s > c$, modified by ψ_{1c}^j for each unit increase in l_{s-c} for all $s > c$. Note that effect modification is currently only supported for binary or continuous confounders.

The data must be in long format, where we assume the convention that each row with time=t contains A_t, L_t and C_{t+1}, Y_{t+1}. That is the censoring indicator for each row should indicate that a user is censored AFTER time t, and the outcome the first outcome that occurs AFTER A_t and L_t are measured. For example, data at time 1, should contain A_1, L_1, Y_2, and optionally C_2. If Y is binary, it must be written as a numeric vector taking values either 0 or 1. The same is true for any covariate that is used for effect modification.
The data must be rectangular with a row entry for every individual for each exposure time 1 up to T. Data rows after censoring should be empty apart from the ID and time variables. This can be done using the function FormatData.
By default the censoring, propensity and outcome models include the exposure history at the previous time as a variable. One may consider also including all previous exposure and confounder history as variables in Lny,Lnp, and LnC if necessary.
Censoring weights are handled as described in Sjolander and Vansteelandt (2016). Note that it is necessary that any variable included in LnC must also be in Lnp. Missing data not due to censoring are automatically handled by removing rows with missing data prior to fitting the model. If outcome models fail to fit, consider removing covariates from Lny but keeping them in Lnp to reduce collinearity issues, or consider the sparseness of the data.

Value

List of the fitted causal parameters of the posited SNMM. These are labeled as follows for each SNMM type, where An is set to the name of the exposure variable, i is the current value of c, j is the category level, and Lny[1] is set to the name of the first confounder in Lny.

type=1 Anj: The effect of exposure at category j, at any time t on all subsequent outcome times.

type=2 Anj: The effect of exposure at category j on outcome at any time t, when Ln[1] is set to zero, on all subsequent outcome times.

Anj:Ln[1]: The effect modification by Lny[1], the additional effect of A at category j on all subsequent Y for each unit increase in Lny[1].

`type=3` c=i.Anj: The effect of exposure at any time t on outcome `c=i` time periods later.

`type=4` c=i.Anj: The effect of exposure at any time t on outcome `c=i` time periods later, when Ln[1] is set to zero.
c=i.Anj:Ln[1]: The effect modification by Lny[1], the additional effect of exposure at category j, on outcome c=i time periods later for each unit increase in Lny[1].

References

Vansteelandt, S., & Sjolander, A. (2016). Revisiting g-estimation of the Effect of a Time-varying Exposure Subject to Time-varying Confounding, Epidemiologic Methods, 5(1), 37-56. <doi:10.1515/em-2015-0005>.

Dukes, O., & Vansteelandt, S. (2018). A Note on g-Estimation of Causal Risk Ratios, American Journal of Epidemiology, 187(5), 1079–1084. <doi:10.1093/aje/kwx347>.

Examples

```
datas<-dataexamples(n=1000,seed=123,Censoring=FALSE)
data=datas$datagestmultcat
#A is a categorical variable with categories labeled 1,2 and 3, with 1 the
#reference category
idvar=”id”
timevar=”time”
Yn=”Y”
An=”A”
Ybin=FALSE
#Remove U from Y to avoid collinearity
Lny=c(”L”,”U”)
Lnp=c(”L”,”U”)
Cn<-NA
LnC<-NA
type=NA


gestmultcat(data,idvar,timevar,Yn,An,Ybin,Lny,Lnp,type=1)

#Example with censoring
datas<-dataexamples(n=1000,seed=123,Censoring=TRUE)
data=datas$datagestmultcat
Cn=”C”
LnC=c(”L”,”U”)
gestmultcat(data,idvar,timevar,Yn,An,Ybin,Lny,Lnp,type=3,Cn,LnC,
cutoff=2)
```

Index