

Package ‘gfoRmula’

April 9, 2022

Title Parametric G-Formula

Version 1.0.0

Description Implements the parametric g-formula algorithm of Robins (1986) <[doi:10.1016/0270-0255\(86\)90088-6](https://doi.org/10.1016/0270-0255(86)90088-6)>. The g-formula can be used to estimate the causal effects of hypothetical time-varying treatment interventions on the mean or risk of an outcome from longitudinal data with time-varying confounding. This package allows: 1) binary or continuous/multi-level time-varying treatments; 2) different types of outcomes (survival or continuous/binary end of follow-up); 3) data with competing events or truncation by death and loss to follow-up and other types of censoring events; 4) different options for handling competing events in the case of survival outcomes; 5) a random measurement/visit process; 6) joint interventions on multiple treatments; and 7) general incorporation of a priori knowledge of the data structure.

Depends R (>= 3.5.0)

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

Imports data.table, ggplot2, ggpubr, grDevices, nnet, parallel, progress, stats, stringr, survival, truncnorm, truncreg, utils

Suggests Hmisc

URL <https://github.com/CausalInference/gfoRmula>,
[https://www.cell.com/patterns/fulltext/S2666-3899\(20\)30008-8](https://www.cell.com/patterns/fulltext/S2666-3899(20)30008-8)

BugReports <https://github.com/CausalInference/gfoRmula/issues>

NeedsCompilation no

Author Victoria Lin [aut] (V. Lin and S. McGrath made equal contributions),
Sean McGrath [aut, cre] (<<https://orcid.org/0000-0002-7281-3516>>, V. Lin and S. McGrath made equal contributions),
Zilu Zhang [aut],
Roger W. Logan [aut],

Lucia C. Petito [aut],
 Jessica G. Young [aut] (<<https://orcid.org/0000-0002-2758-6932>>, M.A.
 Hernán and J.G. Young made equal contributions),
 Miguel A. Hernán [aut] (M.A. Hernán and J.G. Young made equal
 contributions),
 2019 The President and Fellows of Harvard College [cph]

Maintainer Sean McGrath <sean_mcgrath@g.harvard.edu>

Repository CRAN

Date/Publication 2022-04-09 20:02:28 UTC

R topics documented:

basicdata	2
basicdata_nocomp	3
binary_eofdata	4
carry_forward	4
sensor_data	6
coef.gformula	6
continuous_eofdata	8
continuous_eofdata_pb	9
gformula	9
gformula_binary_eof	19
gformula_continuous_eof	25
gformula_survival	31
lagged	39
plot.gformula_binary_eof	41
plot.gformula_continuous_eof	43
plot.gformula_survival	45
print.gformula_survival	48
simple_restriction	51
static	52
threshold	54
vcov.gformula	55
Index	57

basicdata

Example Dataset for a Survival Outcome with Censoring

Description

A dataset consisting of 11,332 observations on 2,500 individuals over 7 time points. Each row in the dataset corresponds to the record of one individual at one time point. Individuals who are censored at time $k + 1$ only have a total of $k + 1$ records, which correspond to time indices $0, \dots, k$.

Usage

basicdata

Format

A data table with 11,332 rows and 8 variables:

t0 Time index.

id Unique identifier for each individual.

L1 Binary time-varying covariate.

L2 Continuous time-varying covariate.

L3 Continuous baseline covariate. For each individual, the baseline values are repeated at each time point.

A Binary treatment variable.

D Competing event; time-varying indicator of failure.

Y Outcome of interest; time-varying indicator of failure.

basicdata_nocomp

Example Dataset for a Survival Outcome without Censoring

Description

A dataset consisting of 13,170 observations on 2,500 individuals over 7 time points. Each row in the dataset corresponds to the record of one individual at one time point.

Usage

basicdata_nocomp

Format

A data table with 13,170 rows and 7 variables:

t0 Time index.

id Unique identifier for each individual.

L1 Binary covariate.

L2 Continuous covariate.

L3 Continuous baseline covariate. For each individual, the baseline values are repeated at each time point.

A Binary treatment variable.

Y Outcome of interest; time-varying indicator of failure.

 binary_eofdata

Example Dataset for a Binary Outcome at End of Follow-Up

Description

A dataset consisting of 17,500 observations on 2,500 individuals over 7 time points. Each row in the dataset corresponds to the record of one individual at one time point.

Usage

binary_eofdata

Format

A data table with 17,500 rows and 7 variables:

time Time index.

id_num Unique identifier for each individual.

cov1 Binary time-varying covariate.

cov2 Continuous time-varying covariate.

cov3 Continuous baseline covariate. For each individual, the baseline values are repeated at each time point.

treat Binary treatment variable.

outcome Binary outcome of interest. Because this outcome is only defined at the end of follow-up, values of NA are given in all other time points.

 carry_forward

Carry Forward

Description

This function assists the implementation of a restriction on a covariate in the data table `newdf`. A particular covariate is simulated only when some condition (usually a covariate representing whether a doctor's visit occurred or not) is TRUE. If the condition is FALSE, the covariate value is not simulated for that time point and the value is instead carried over from the previous time point.

Usage

carry_forward(newdf, pool, restriction, time_name, t)

Arguments

newdf	Data table containing the simulated data at time t .
pool	Data table containing the simulated data at times before t .
restriction	List of vectors. Each vector contains as its first entry the covariate affected by the restriction; its second entry the condition that must be TRUE for the covariate to be modeled; its third entry a function that executes other specific actions based on the condition (in this case, this function); and its fourth entry some value used by the function (in this case, this entry is not used).
time_name	Character string specifying the name of the time variable in pool and newdf.
t	Integer specifying the current time index.

Value

No value is returned. The data table newdf is modified in place.

Examples

```
## Estimating the effect of static treatment strategies on risk of a
## failure event

id <- 'id'
time_points <- 7
time_name <- 't0'
covnames <- c('L1', 'L2', 'A')
outcome_name <- 'Y'
covtypes <- c('binary', 'bounded normal', 'binary')
histories <- c(lagged, lagavg)
histvars <- list(c('A', 'L1', 'L2'), c('L1', 'L2'))
covparams <- list(covmodels = c(L1 ~ lag1_A + lag_cumavg1_L1 + lag_cumavg1_L2 +
                                L3 + t0,
                                L2 ~ lag1_A + L1 + lag_cumavg1_L1 +
                                lag_cumavg1_L2 + L3 + t0,
                                A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 +
                                lag_cumavg1_L2 + L3 + t0))
ymodel <- Y ~ A + L1 + L2 + L3 + lag1_A + lag1_L1 + lag1_L2 + t0
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, time_points))),
                      list(c(static, rep(1, time_points))))
int_descript <- c('Never treat', 'Always treat')
nsimul <- 10000

# At t0 == 5, assign L1 its value at the previous time point
restrictions <- list(c('L2', 't0 != 5', carry_forward))

gform_basic <- gformula_survival(obs_data = basicdata_nocomp, id = id,
                                time_points = time_points,
                                time_name = time_name, covnames = covnames,
                                outcome_name = outcome_name,
                                covtypes = covtypes,
                                covparams = covparams, ymodel = ymodel,
```

```

                                intvars = intvars,
                                interventions = interventions,
                                int_descript = int_descript,
                                restrictions = restrictions,
                                histories = histories, histvars = histvars,
                                basecovs = c('L3'), nsimul = nsimul,
                                seed = 1234)
gform_basic

```

censor_data	<i>Example Dataset for a Survival Outcome with an Indicator of Censoring Variable</i>
-------------	---

Description

A dataset consisting of 118,725 observations on 20,000 individuals over 10 time points. Each row in the dataset corresponds to the record of one individual at one time point.

Usage

```
censor_data
```

Format

A data table with 22,500 rows and 7 variables:

- t0** Time index.
- id** Unique identifier for each individual.
- L** Binary time-varying covariate.
- A** Continuous treatment variable.
- C** Censoring indicator.
- Y** Outcome of interest; time-varying indicator of failure.

coef.gformula	<i>Coefficient method for objects of class "gformula"</i>
---------------	---

Description

This function extracts the coefficients of the fitted models for the time-varying covariates, outcome, and compevent event (if applicable).

Usage

```
## S3 method for class 'gformula'
coef(object, ...)
```

Arguments

```
object      Object of class "gformula".
...         Other arguments.
```

Value

If `bootdiag` was set to `FALSE` in `gformula`, this function returns a list of the coefficients of the fitted models to the observed data set. If bootstrapping was used and `bootdiag` was set to `TRUE` in `gformula`, this function returns a list described as follows. The first element (named 'Original sample') is a list of the coefficients of the fitted models to the observed data set. The *k*th element (named 'Bootstrap sample *k*-1') is a list of the coefficients of the fitted models corresponding to the *k*-1th bootstrap sample.

See Also

[gformula](#)

Examples

```
## Estimating the effect of static treatment strategies on risk of a
## failure event

id <- 'id'
time_points <- 7
time_name <- 't0'
covnames <- c('L1', 'L2', 'A')
outcome_name <- 'Y'
outcome_type <- 'survival'
covtypes <- c('binary', 'bounded normal', 'binary')
histories <- c('lagged', 'lagavg')
histvars <- list(c('A', 'L1', 'L2'), c('L1', 'L2'))
covparams <- list(covmodels = c(L1 ~ lag1_A + lag_cumavg1_L1 + lag_cumavg1_L2 +
                                L3 + t0,
                                L2 ~ lag1_A + L1 + lag_cumavg1_L1 +
                                lag_cumavg1_L2 + L3 + t0,
                                A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 +
                                lag_cumavg1_L2 + L3 + t0))
ymodel <- Y ~ A + L1 + L2 + L3 + lag1_A + lag1_L1 + lag1_L2 + t0
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, time_points))),
                      list(c(static, rep(1, time_points))))
int_descript <- c('Never treat', 'Always treat')
nsimul <- 10000

gform_basic <- gformula(obs_data = basicdata_nocomp, id = id,
                       time_points = time_points,
```

```

time_name = time_name, covnames = covnames,
outcome_name = outcome_name,
outcome_type = outcome_type, covtypes = covtypes,
covparams = covparams, ymodel = ymodel,
intvars = intvars,
interventions = interventions,
int_descript = int_descript,
histories = histories, histvars = histvars,
basecovs = c('L3'), nsimul = nsimul,
seed = 1234)
coef(gform_basic)

```

continuous_eofdata *Example Dataset for a Continuous Outcome at End of Follow-Up*

Description

A dataset consisting of 17,500 observations on 2,500 individuals over 7 time points. Each row in the dataset corresponds to the record of one individual at one time point.

Usage

```
continuous_eofdata
```

Format

A data table with 17,500 rows and 7 variables:

t0 Time index.

id Unique identifier for each individual.

L1 Categorical time-varying covariate.

L2 Continuous time-varying covariate.

L3 Continuous baseline covariate. For each individual, the baseline values are repeated at each time point.

A Binary treatment variable.

Y Continuous outcome of interest. Because this outcome is only defined at the end of follow-up, values of NA are given in all other time points.

continuous_eofdata_pb *Example Dataset for a Continuous Outcome at End of Follow-Up with Pre-Baseline Times*

Description

A dataset consisting of 22,500 observations on 2,500 individuals over 2 pre-baseline time points and follow-up 7 time points. Each row in the dataset corresponds to the record of one individual at one time point.

Usage

```
continuous_eofdata_pb
```

Format

A data table with 22,500 rows and 7 variables:

t0 Time index.

id Unique identifier for each individual.

L1 Categorical time-varying covariate.

L2 Continuous time-varying covariate.

L3 Continuous baseline covariate. For each individual, the baseline values are repeated at each time point.

A Binary treatment variable.

Y Continuous outcome of interest. Because this outcome is only defined at the end of follow-up, values of NA are given in all other time points.

gformula *Estimation of Survival Outcome, Continuous End-of-Follow-Up Outcome, or Binary End-of-Follow-Up Outcome Under the Parametric G-Formula*

Description

Based on an observed data set, this function estimates the risk over time (for survival outcomes), outcome mean at end-of-follow-up (for continuous end-of-follow-up outcomes), or outcome probability at end-of-follow-up (for binary end-of-follow-up outcomes) under multiple user-specified interventions using the parametric g-formula. See McGrath et al. (2020) for further details concerning the application and implementation of the parametric g-formula.

Usage

```
gformula(  
  obs_data,  
  id,  
  time_points = NULL,  
  time_name,  
  covnames,  
  covtypes,  
  covparams,  
  covfits_custom = NA,  
  covpredict_custom = NA,  
  histvars = NULL,  
  histories = NA,  
  basecovs = NA,  
  outcome_name,  
  outcome_type,  
  ymodel,  
  compevent_name = NULL,  
  compevent_model = NA,  
  compevent_cens = FALSE,  
  censor_name = NULL,  
  censor_model = NA,  
  intvars = NULL,  
  interventions = NULL,  
  int_times = NULL,  
  int_descript = NULL,  
  ref_int = 0,  
  intcomp = NA,  
  visitprocess = NA,  
  restrictions = NA,  
  yrestrictions = NA,  
  compevent_restrictions = NA,  
  baselags = FALSE,  
  nsimul = NA,  
  sim_data_b = FALSE,  
  seed,  
  nsamples = 0,  
  parallel = FALSE,  
  ncores = NA,  
  ci_method = "percentile",  
  threads,  
  model_fits = FALSE,  
  boot_diag = FALSE,  
  show_progress = TRUE,  
  ipw_cutoff_quantile = NULL,  
  ipw_cutoff_value = NULL,  
  ...  
)
```

Arguments

obs_data	Data table containing the observed data.
id	Character string specifying the name of the ID variable in obs_data.
time_points	Number of time points to simulate. By default, this argument is set equal to the maximum number of records that obs_data contains for any individual plus 1.
time_name	Character string specifying the name of the time variable in obs_data.
covnames	Vector of character strings specifying the names of the time-varying covariates in obs_data.
covtypes	Vector of character strings specifying the "type" of each time-varying covariate included in covnames. The possible "types" are: "binary", "normal", "categorical", "bounded normal", "zero-inflated normal", "truncated normal", "absorbing", "categorical time", and "custom".
covparams	List of vectors, where each vector contains information for one parameter used in the modeling of the time-varying covariates (e.g., model statement, family, link function, etc.). Each vector must be the same length as covnames and in the same order. If a parameter is not required for a certain covariate, it should be set to NA at that index.
covfits_custom	Vector containing custom fit functions for time-varying covariates that do not fall within the pre-defined covariate types. It should be in the same order covnames. If a custom fit function is not required for a particular covariate (e.g., if the first covariate is of type "binary" but the second is of type "custom"), then that index should be set to NA. The default is NA.
covpredict_custom	Vector containing custom prediction functions for time-varying covariates that do not fall within the pre-defined covariate types. It should be in the same order as covnames. If a custom prediction function is not required for a particular covariate, then that index should be set to NA. The default is NA.
histvars	List of vectors. The kth vector specifies the names of the variables for which the kth history function in histories is to be applied.
histories	Vector of history functions to apply to the variables specified in histvars. The default is NA.
basecovs	Vector of character strings specifying the names of baseline covariates in obs_data. These covariates are not simulated using a model but rather carry their value over all time points from the first time point of obs_data. These covariates should not be included in covnames. The default is NA.
outcome_name	Character string specifying the name of the outcome variable in obs_data.
outcome_type	Character string specifying the "type" of outcome. The possible "types" are: "survival", "continuous_eof", and "binary_eof".
ymodel	Model statement for the outcome variable.
compevent_name	Character string specifying the name of the competing event variable in obs_data. Only applicable for survival outcomes.
compevent_model	Model statement for the competing event variable. The default is NA. Only applicable for survival outcomes.

compevent_cens	Logical scalar indicating whether to treat competing events as censoring events. This argument is only applicable for survival outcomes and when a competing event model is supplied (i.e., <code>compevent_name</code> and <code>compevent_model</code> are specified). If this argument is set to TRUE, the competing event model will only be used to construct inverse probability weights to estimate the natural course means / risk from the observed data. If this argument is set to FALSE, the competing event model will be used in the parametric g-formula estimates of the risk and will not be used to construct inverse probability weights. See "Details". The default is FALSE.
censor_name	Character string specifying the name of the censoring variable in <code>obs_data</code> . Only applicable when using inverse probability weights to estimate the natural course means / risk from the observed data. See "Details".
censor_model	Model statement for the censoring variable. Only applicable when using inverse probability weights to estimate the natural course means / risk from the observed data. See "Details".
intvars	List, whose elements are vectors of character strings. The <i>k</i> th vector in <code>intvars</code> specifies the name(s) of the variable(s) to be intervened on in each round of the simulation under the <i>k</i> th intervention in <code>interventions</code> .
interventions	List, whose elements are lists of vectors. Each list in <code>interventions</code> specifies a unique intervention on the relevant variable(s) in <code>intvars</code> . Each vector contains a function implementing a particular intervention on a single variable, optionally followed by one or more "intervention values" (i.e., integers used to specify the treatment regime).
int_times	List, whose elements are lists of vectors. The <i>k</i> th list in <code>int_times</code> corresponds to the <i>k</i> th intervention in <code>interventions</code> . Each vector specifies the time points in which the relevant intervention is applied on the corresponding variable in <code>intvars</code> . When an intervention is not applied, the simulated natural course value is used. By default, this argument is set so that all interventions are applied in all time points.
int_descript	Vector of character strings, each describing an intervention. It must be in same order as the entries in <code>interventions</code> .
ref_int	Integer denoting the intervention to be used as the reference for calculating the risk ratio and risk difference. 0 denotes the natural course, while subsequent integers denote user-specified interventions in the order that they are named in <code>interventions</code> . The default is 0.
intcomp	List of two numbers indicating a pair of interventions to be compared by a hazard ratio. The default is NA, resulting in no hazard ratio calculation.
visitprocess	List of vectors. Each vector contains as its first entry the covariate name of a visit process; its second entry the name of a covariate whose modeling depends on the visit process; and its third entry the maximum number of consecutive visits that can be missed before an individual is censored. The default is NA.
restrictions	List of vectors. Each vector contains as its first entry a covariate for which <i>a priori</i> knowledge of its distribution is available; its second entry a condition under which no knowledge of its distribution is available and that must be TRUE for the distribution of that covariate given that condition to be estimated via a parametric model or other fitting procedure; its third entry a function for estimating the

distribution of that covariate given the condition in the second entry is false such that *a priori* knowledge of the covariate distribution is available; and its fourth entry a value used by the function in the third entry. The default is NA.

yrestrictions	List of vectors. Each vector contains as its first entry a condition and its second entry an integer. When the condition is TRUE, the outcome variable is simulated according to the fitted model; when the condition is FALSE, the outcome variable takes on the value in the second entry. The default is NA.
compevent_restrictions	List of vectors. Each vector contains as its first entry a condition and its second entry an integer. When the condition is TRUE, the competing event variable is simulated according to the fitted model; when the condition is FALSE, the competing event variable takes on the value in the second entry. The default is NA. Only applicable for survival outcomes.
baselags	Logical scalar for specifying the convention used for <code>lagi</code> and <code>lag_cumavgi</code> terms in the model statements when pre-baseline times are not included in <code>obs_data</code> and when the current time index, t , is such that $t < i$. If this argument is set to FALSE, the value of all <code>lagi</code> and <code>lag_cumavgi</code> terms in this context are set to 0 (for non-categorical covariates) or the reference level (for categorical covariates). If this argument is set to TRUE, the value of <code>lagi</code> and <code>lag_cumavgi</code> terms are set to their values at time 0. The default is FALSE.
nsimul	Number of subjects for whom to simulate data. By default, this argument is set equal to the number of subjects in <code>obs_data</code> .
sim_data_b	Logical scalar indicating whether to return the simulated data set. If bootstrap samples are used (i.e., <code>nsamples</code> is set to a value greater than 0), this argument must be set to FALSE. The default is FALSE.
seed	Starting seed for simulations and bootstrapping.
nsamples	Integer specifying the number of bootstrap samples to generate. The default is 0.
parallel	Logical scalar indicating whether to parallelize simulations of different interventions to multiple cores.
ncores	Integer specifying the number of CPU cores to use in parallel simulation. This argument is required when <code>parallel</code> is set to TRUE. In many applications, users may wish to set this argument equal to <code>parallel::detectCores() - 1</code> .
ci_method	Character string specifying the method for calculating the bootstrap 95% confidence intervals, if applicable. The options are "percentile" and "normal".
threads	Integer specifying the number of threads to be used in <code>data.table</code> . See setDTthreads for further details.
model_fits	Logical scalar indicating whether to return the fitted models. Note that if this argument is set to TRUE, the output of this function may use a lot of memory. The default is FALSE.
boot_diag	Logical scalar indicating whether to return the coefficients, standard errors, and variance-covariance matrices of the parameters of the fitted models in the bootstrap samples. The default is FALSE.

<code>show_progress</code>	Logical scalar indicating whether to print a progress bar for the number of bootstrap samples completed in the R console. This argument is only applicable when <code>parallel</code> is set to <code>FALSE</code> and bootstrap samples are used (i.e., <code>nsamples</code> is set to a value greater than 0). The default is <code>TRUE</code> .
<code>ipw_cutoff_quantile</code>	Percentile by which to truncate inverse probability weights. The default is <code>NULL</code> (i.e., no truncation). See "Details".
<code>ipw_cutoff_value</code>	Cutoff value by which to truncate inverse probability weights. The default is <code>NULL</code> (i.e., no truncation). See "Details".
<code>...</code>	Other arguments, which are passed to the functions in <code>covpredict_custom</code> .

Details

To assess model misspecification in the parametric g-formula, users can obtain inverse probability (IP) weighted estimates of the natural course risk and/or means of the time-varying covariates from the observed data. See Chiu et al. (2022) for details. In addition to the general requirements described in McGrath et al. (2020), the requirements for the input data set and the call to the `gformula` function for such analyses are described below.

Users need to include a column in `obs_data` with a time-varying censoring variable. Users need to indicate the name of the censoring variable and a model statement for the censoring variable with parameters `sensor_name` and `sensor_model`, respectively. When competing events are present, users need to include a column in `obs_data` with a time-varying indicator of the competing event variable and need to indicate the name of the competing event variable and the corresponding model statement with parameters `compevent_name` and `compevent_model`, respectively. Users need to indicate whether to treat competing events as censoring events with the `compevent_cens` parameter. Finally, users can specify how to truncate IP weights with the `ipw_cutoff_quantile` or `ipw_cutoff_value` parameters.

Value

An object of class `gformula_survival`. The object is a list with the following components:

<code>result</code>	Results table. For survival outcomes, this contains the estimated risk, risk difference, and risk ratio for all interventions (including the natural course) at each time point. For continuous end-of-follow-up outcomes, this contains estimated mean outcome, mean difference, and mean ratio for all interventions (including natural course) at the last time point. For binary end-of-follow-up outcomes, this contains the estimated outcome probability, probability difference, and probability ratio for all interventions (including natural course) at the last time point. If bootstrapping was used, the results table includes the bootstrap risk / mean / probability difference, ratio, standard error, and 95% confidence interval.
<code>coeffs</code>	A list of the coefficients of the fitted models.
<code>stderrs</code>	A list of the standard errors of the coefficients of the fitted models.
<code>vcovs</code>	A list of the variance-covariance matrices of the parameters of the fitted models.
<code>rmses</code>	A list of root mean square error (RMSE) values of the fitted models.

hazardratio_val	Hazard ratio between two interventions (if applicable).
fits	A list of the fitted models for the time-varying covariates, outcome, and competing event (if applicable). If <code>model_fits</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
sim_data	A list of data tables of the simulated data. Each element in the list corresponds to one of the interventions. If the argument <code>sim_data_b</code> is set to <code>FALSE</code> , a value of <code>NA</code> is given.
IP_weights	A numeric vector specifying the inverse probability weights. See "Details".
bootcoeffs	A list, where the <code>k</code> th element is a list containing the coefficients of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
bootstderrs	A list, where the <code>k</code> th element is a list containing the standard errors of the coefficients of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
bootvcovs	A list, where the <code>k</code> th element is a list containing the variance-covariance matrices of the parameters of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
...	Some additional elements.

The results for the g-formula simulation are printed with the `print.gformula_survival`, `print.gformula_continuous_eof` and `print.gformula_binary_eof` functions. To generate graphs comparing the mean estimated covariate values and risks over time and mean observed covariate values and risks over time, use the `plot.gformula_survival`, `plot.gformula_continuous_eof`, and `plot.gformula_binary_eof` functions.

References

Chiu YH, Wen L, McGrath S, Logan R, Dahabreh IJ, Hernán MA. Assessing model misspecification of the parametric g-formula in the presence of censoring. In preparation.

McGrath S, Lin V, Zhang Z, Petito LC, Logan RW, Hernán MA, and JG Young. `gfoRmula`: An R package for estimating the effects of sustained treatment strategies via the parametric g-formula. *Patterns*. 2020;1:100008.

Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period: application to the healthy worker survivor effect. *Mathematical Modelling*. 1986;7:1393–1512. [Errata (1987) in *Computers and Mathematics with Applications* 14, 917.-921. Addendum (1987) in *Computers and Mathematics with Applications* 14, 923-.945. Errata (1987) to addendum in *Computers and Mathematics with Applications* 18, 477.].

Examples

```
## Estimating the effect of static treatment strategies on risk of a
## failure event

id <- 'id'
time_points <- 7
time_name <- 't0'
```



```

      A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 +
      lag_cumavg1_L2 + L3 + as.factor(t0)))
ymodel <- Y ~ A + L1 + L2 + lag1_A + lag1_L1 + lag1_L2 + L3 + as.factor(t0)
compevent_model <- D ~ A + L1 + L2 + lag1_A + lag1_L1 + lag1_L2 + L3 + as.factor(t0)
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, time_points))),
  list(c(static, rep(1, time_points))))
int_descript <- c('Never treat', 'Always treat')
nsimul <- 10000

gform_basic <- gformula(obs_data = basicdata, id = id,
  time_points = time_points,
  time_name = time_name, covnames = covnames,
  outcome_name = outcome_name,
  outcome_type = outcome_type,
  compevent_name = compevent_name,
  covtypes = covtypes,
  covparams = covparams, ymodel = ymodel,
  compevent_model = compevent_model,
  intvars = intvars, interventions = interventions,
  int_descript = int_descript,
  histories = histories, histvars = histvars,
  basecovs = c('L3'), nsimul = nsimul,
  seed = 1234)

gform_basic

## Estimating the effect of treatment strategies on the mean of a continuous
## end of follow-up outcome

library('Hmisc')
id <- 'id'
time_name <- 't0'
covnames <- c('L1', 'L2', 'A')
outcome_name <- 'Y'
outcome_type <- 'continuous_eof'
covtypes <- c('categorical', 'normal', 'binary')
histories <- c(lagged)
histvars <- list(c('A', 'L1', 'L2'))
covparams <- list(covmodels = c(L1 ~ lag1_A + lag1_L1 + L3 + t0 +
  rcspline.eval(lag1_L2, knots = c(-1, 0, 1)),
  L2 ~ lag1_A + L1 + lag1_L1 + lag1_L2 + L3 + t0,
  A ~ lag1_A + L1 + L2 + lag1_L1 + lag1_L2 + L3 + t0))
ymodel <- Y ~ A + L1 + L2 + lag1_A + lag1_L1 + lag1_L2 + L3
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, 7))),
  list(c(static, rep(1, 7))))
int_descript <- c('Never treat', 'Always treat')
nsimul <- 10000

gform_cont_eof <- gformula(obs_data = continuous_eofdata,
  id = id, time_name = time_name,

```

```

covnames = covnames, outcome_name = outcome_name,
outcome_type = outcome_type, covtypes = covtypes,
covparams = covparams, ymodel = ymodel,
intvars = intvars, interventions = interventions,
int_descript = int_descript,
histories = histories, histvars = histvars,
basecovs = c("L3"), nsimul = nsimul, seed = 1234)
gform_cont_eof

## Estimating the effect of threshold interventions on the mean of a binary
## end of follow-up outcome

outcome_type <- 'binary_eof'
id <- 'id_num'
time_name <- 'time'
covnames <- c('cov1', 'cov2', 'treat')
outcome_name <- 'outcome'
histories <- c(lagged, cumavg)
histvars <- list(c('treat', 'cov1', 'cov2'), c('cov1', 'cov2'))
covtypes <- c('binary', 'zero-inflated normal', 'normal')
covparams <- list(covmodels = c(cov1 ~ lag1_treat + lag1_cov1 + lag1_cov2 +
  cov3 + time,
  cov2 ~ lag1_treat + cov1 + lag1_cov1 +
  lag1_cov2 + cov3 + time,
  treat ~ lag1_treat + cumavg_cov1 +
  cumavg_cov2 + cov3 + time))
ymodel <- outcome ~ treat + cov1 + cov2 + lag1_cov1 + lag1_cov2 + cov3
intvars <- list('treat', 'treat')
interventions <- list(list(c(static, rep(0, 7))),
  list(c(threshold, 1, Inf)))
int_descript <- c('Never treat', 'Threshold - lower bound 1')
nsimul <- 10000
ncores <- 2

gform_bin_eof <- gformula(obs_data = binary_eofdata,
  outcome_type = outcome_type, id = id,
  time_name = time_name, covnames = covnames,
  outcome_name = outcome_name, covtypes = covtypes,
  covparams = covparams, ymodel = ymodel,
  intvars = intvars, interventions = interventions,
  int_descript = int_descript, histories = histories,
  histvars = histvars, basecovs = c("cov3"),
  seed = 1234, parallel = TRUE, nsamples = 5,
  nsimul = nsimul, ncores = ncores)
gform_bin_eof

## Using IP weighting to estimate natural course risk
## Only the natural course intervention is included for simplicity

covnames <- c('L', 'A')
```

```

histories <- c(lagged)
histvars <- list(c('A', 'L'))
ymodel <- Y ~ L + A
covtypes <- c('binary', 'normal')
covparams <- list(covmodels = c(L ~ lag1_L + lag1_A,
                                A ~ lag1_L + L + lag1_A))

censor_name <- 'C'
censor_model <- C ~ L
res_censor <- gformula(obs_data = censor_data, id = 'id',
                      time_name = 't0', covnames = covnames,
                      outcome_name = 'Y', outcome_type = 'survival',
                      censor_name = censor_name, censor_model = censor_model,
                      covtypes = covtypes,
                      covparams = covparams, ymodel = ymodel,
                      intvars = NULL, interventions = NULL, int_descript = NULL,
                      histories = histories, histvars = histvars,
                      seed = 1234)

plot(res_censor)

```

gformula_binary_eof *Estimation of Binary End-of-Follow-Up Outcome Under the Parametric G-Formula*

Description

Based on an observed data set, this internal function estimates the outcome probability at end-of-follow-up under multiple user-specified interventions using the parametric g-formula. See McGrath et al. (2020) for further details concerning the application and implementation of the parametric g-formula.

Usage

```

gformula_binary_eof(
  obs_data,
  id,
  time_name,
  covnames,
  covtypes,
  covparams,
  covfits_custom = NA,
  covpredict_custom = NA,
  histvars = NULL,
  histories = NA,
  basecovs = NA,
  censor_name = NULL,
  censor_model = NA,
  outcome_name,

```

```

ymodel,
intvars = NULL,
interventions = NULL,
int_times = NULL,
int_descript = NULL,
ref_int = 0,
visitprocess = NA,
restrictions = NA,
yrestrictions = NA,
baselags = FALSE,
nsimul = NA,
sim_data_b = FALSE,
seed,
nsamples = 0,
parallel = FALSE,
ncores = NA,
ci_method = "percentile",
threads,
model_fits = FALSE,
boot_diag = FALSE,
show_progress = TRUE,
ipw_cutoff_quantile = NULL,
ipw_cutoff_value = NULL,
...
)

```

Arguments

<code>obs_data</code>	Data table containing the observed data.
<code>id</code>	Character string specifying the name of the ID variable in <code>obs_data</code> .
<code>time_name</code>	Character string specifying the name of the time variable in <code>obs_data</code> .
<code>covnames</code>	Vector of character strings specifying the names of the time-varying covariates in <code>obs_data</code> .
<code>covtypes</code>	Vector of character strings specifying the "type" of each time-varying covariate included in <code>covnames</code> . The possible "types" are: "binary", "normal", "categorical", "bounded normal", "zero-inflated normal", "truncated normal", "absorbing", "categorical time", and "custom".
<code>covparams</code>	List of vectors, where each vector contains information for one parameter used in the modeling of the time-varying covariates (e.g., model statement, family, link function, etc.). Each vector must be the same length as <code>covnames</code> and in the same order. If a parameter is not required for a certain covariate, it should be set to NA at that index.
<code>covfits_custom</code>	Vector containing custom fit functions for time-varying covariates that do not fall within the pre-defined covariate types. It should be in the same order <code>covnames</code> . If a custom fit function is not required for a particular covariate (e.g., if the first covariate is of type "binary" but the second is of type "custom"), then that index should be set to NA. The default is NA.

covpredict_custom	Vector containing custom prediction functions for time-varying covariates that do not fall within the pre-defined covariate types. It should be in the same order as covnames. If a custom prediction function is not required for a particular covariate, then that index should be set to NA. The default is NA.
histvars	List of vectors. The kth vector specifies the names of the variables for which the kth history function in histories is to be applied.
histories	Vector of history functions to apply to the variables specified in histvars. The default is NA.
basecovs	Vector of character strings specifying the names of baseline covariates in obs_data. These covariates are not simulated using a model but rather carry their value over all time points from the first time point of obs_data. These covariates should not be included in covnames. The default is NA.
censor_name	Character string specifying the name of the censoring variable in obs_data. Only applicable when using inverse probability weights to estimate the natural course means / risk from the observed data. See "Details".
censor_model	Model statement for the censoring variable. Only applicable when using inverse probability weights to estimate the natural course means / risk from the observed data. See "Details".
outcome_name	Character string specifying the name of the outcome variable in obs_data.
ymodel	Model statement for the outcome variable.
intvars	List, whose elements are vectors of character strings. The kth vector in intvars specifies the name(s) of the variable(s) to be intervened on in each round of the simulation under the kth intervention in interventions.
interventions	List, whose elements are lists of vectors. Each list in interventions specifies a unique intervention on the relevant variable(s) in intvars. Each vector contains a function implementing a particular intervention on a single variable, optionally followed by one or more "intervention values" (i.e., integers used to specify the treatment regime).
int_times	List, whose elements are lists of vectors. The kth list in int_times corresponds to the kth intervention in interventions. Each vector specifies the time points in which the relevant intervention is applied on the corresponding variable in intvars. When an intervention is not applied, the simulated natural course value is used. By default, this argument is set so that all interventions are applied in all time points.
int_descript	Vector of character strings, each describing an intervention. It must be in same order as the entries in interventions.
ref_int	Integer denoting the intervention to be used as the reference for calculating the end-of-follow-up mean ratio and mean difference. 0 denotes the natural course, while subsequent integers denote user-specified interventions in the order that they are named in interventions. The default is 0.
visitprocess	List of vectors. Each vector contains as its first entry the covariate name of a visit process; its second entry the name of a covariate whose modeling depends on the visit process; and its third entry the maximum number of consecutive visits that can be missed before an individual is censored. The default is NA.

restrictions	List of vectors. Each vector contains as its first entry a covariate for which <i>a priori</i> knowledge of its distribution is available; its second entry a condition under which no knowledge of its distribution is available and that must be TRUE for the distribution of that covariate given that condition to be estimated via a parametric model or other fitting procedure; its third entry a function for estimating the distribution of that covariate given the condition in the second entry is false such that <i>a priori</i> knowledge of the covariate distribution is available; and its fourth entry a value used by the function in the third entry. The default is NA.
yrestrictions	List of vectors. Each vector contains as its first entry a condition and its second entry an integer. When the condition is TRUE, the outcome variable is simulated according to the fitted model; when the condition is FALSE, the outcome variable takes on the value in the second entry. The default is NA.
baselags	Logical scalar for specifying the convention used for <code>lagi</code> and <code>lag_cumavgi</code> terms in the model statements when pre-baseline times are not included in <code>obs_data</code> and when the current time index, t , is such that $t < i$. If this argument is set to FALSE, the value of all <code>lagi</code> and <code>lag_cumavgi</code> terms in this context are set to 0 (for non-categorical covariates) or the reference level (for categorical covariates). If this argument is set to TRUE, the value of <code>lagi</code> and <code>lag_cumavgi</code> terms are set to their values at time 0. The default is FALSE.
nsimul	Number of subjects for whom to simulate data. By default, this argument is set equal to the number of subjects in <code>obs_data</code> .
sim_data_b	Logical scalar indicating whether to return the simulated data set. If bootstrap samples are used (i.e., <code>nsamples</code> is set to a value greater than 0), this argument must be set to FALSE. The default is FALSE.
seed	Starting seed for simulations and bootstrapping.
nsamples	Integer specifying the number of bootstrap samples to generate. The default is 0.
parallel	Logical scalar indicating whether to parallelize simulations of different interventions to multiple cores.
ncores	Integer specifying the number of CPU cores to use in parallel simulation. This argument is required when <code>parallel</code> is set to TRUE. In many applications, users may wish to set this argument equal to <code>parallel::detectCores() - 1</code> .
ci_method	Character string specifying the method for calculating the bootstrap 95% confidence intervals, if applicable. The options are "percentile" and "normal".
threads	Integer specifying the number of threads to be used in <code>data.table</code> . See setDTthreads for further details.
model_fits	Logical scalar indicating whether to return the fitted models. Note that if this argument is set to TRUE, the output of this function may use a lot of memory. The default is FALSE.
boot_diag	Logical scalar indicating whether to return the coefficients, standard errors, and variance-covariance matrices of the parameters of the fitted models in the bootstrap samples. The default is FALSE.
show_progress	Logical scalar indicating whether to print a progress bar for the number of bootstrap samples completed in the R console. This argument is only applicable when <code>parallel</code> is set to FALSE and bootstrap samples are used (i.e., <code>nsamples</code> is set to a value greater than 0). The default is TRUE.

ipw_cutoff_quantile	Percentile by which to truncate inverse probability weights. The default is NULL (i.e., no truncation). See "Details".
ipw_cutoff_value	Cutoff value by which to truncate inverse probability weights. The default is NULL (i.e., no truncation). See "Details".
...	Other arguments, which are passed to the functions in covpredict_custom.

Details

To assess model misspecification in the parametric g-formula, users can obtain inverse probability (IP) weighted estimates of the natural course means of the time-varying covariates from the observed data. See Chiu et al. (2022) for details. In addition to the general requirements described in McGrath et al. (2020), the requirements for the input data set and the call to the gformula function for such analyses are described below.

Users need to include a column in `obs_data` with a time-varying censoring variable. Users need to indicate the name of the censoring variable and a model statement for the censoring variable with parameters `sensor_name` and `sensor_model`, respectively. Finally, users can specify how to truncate IP weights with the `ipw_cutoff_quantile` or `ipw_cutoff_value` parameters.

Value

An object of class `gformula_binary_eof`. The object is a list with the following components:

<code>result</code>	Results table containing the estimated outcome probability for all interventions (including natural course) at the last time point. If bootstrapping was used, the results table includes the bootstrap end-of-follow-up mean ratio, standard error, and 95% confidence interval.
<code>coeffs</code>	A list of the coefficients of the fitted models.
<code>stderrs</code>	A list of the standard errors of the coefficients of the fitted models.
<code>vcovs</code>	A list of the variance-covariance matrices of the parameters of the fitted models.
<code>rmses</code>	A list of root mean square error (RMSE) values of the fitted models.
<code>fits</code>	A list of the fitted models for the time-varying covariates and outcome. If <code>model_fits</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
<code>sim_data</code>	A list of data tables of the simulated data. Each element in the list corresponds to one of the interventions. If the argument <code>sim_data_b</code> is set to <code>FALSE</code> , a value of <code>NA</code> is given.
<code>IP_weights</code>	A numeric vector specifying the inverse probability weights. See "Details".
<code>bootcoeffs</code>	A list, where the <code>k</code> th element is a list containing the coefficients of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
<code>bootstderrs</code>	A list, where the <code>k</code> th element is a list containing the standard errors of the coefficients of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.

bootvcovs A list, where the kth element is a list containing the variance-covariance matrices of the parameters of the fitted models corresponding to the kth bootstrap sample. If boot_diag is set to FALSE, a value of NULL is given.

... Some additional elements.

The results for the g-formula simulation under various interventions for the last time point are printed with the `print.gformula_binary_eof` function. To generate graphs comparing the mean estimated and observed covariate values over time, use the `plot.gformula_binary_eof` function.

References

Chiu YH, Wen L, McGrath S, Logan R, Dahabreh IJ, Hernán MA. Assessing model misspecification of the parametric g-formula in the presence of censoring. In preparation.

McGrath S, Lin V, Zhang Z, Petito LC, Logan RW, Hernán MA, and JG Young. gfoRmula: An R package for estimating the effects of sustained treatment strategies via the parametric g-formula. *Patterns*. 2020;1:100008.

Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period: application to the healthy worker survivor effect. *Mathematical Modelling*. 1986;7:1393–1512. [Errata (1987) in *Computers and Mathematics with Applications* 14, 917.-921. Addendum (1987) in *Computers and Mathematics with Applications* 14, 923-.945. Errata (1987) to addendum in *Computers and Mathematics with Applications* 18, 477.].

See Also

[gformula](#)

Examples

```
## Estimating the effect of threshold interventions on the mean of a binary
## end of follow-up outcome

id <- 'id_num'
time_name <- 'time'
covnames <- c('cov1', 'cov2', 'treat')
outcome_name <- 'outcome'
histories <- c(lagged, cumavg)
histvars <- list(c('treat', 'cov1', 'cov2'), c('cov1', 'cov2'))
covtypes <- c('binary', 'zero-inflated normal', 'normal')
covparams <- list(covmodels = c(cov1 ~ lag1_treat + lag1_cov1 + lag1_cov2 + cov3 +
                                time,
                                cov2 ~ lag1_treat + cov1 + lag1_cov1 + lag1_cov2 +
                                cov3 + time,
                                treat ~ lag1_treat + cumavg_cov1 +
                                cumavg_cov2 + cov3 + time))
ymodel <- outcome ~ treat + cov1 + cov2 + lag1_cov1 + lag1_cov2 + cov3
intvars <- list('treat', 'treat')
interventions <- list(list(c(static, rep(0, 7))),
                      list(c(threshold, 1, Inf)))
int_descript <- c('Never treat', 'Threshold - lower bound 1')
```



```

nsimul <- 10000
ncores <- 2

gform_bin_eof <- gformula_binary_eof(obs_data = binary_eofdata, id = id,
                                     time_name = time_name,
                                     covnames = covnames,
                                     outcome_name = outcome_name,
                                     covtypes = covtypes,
                                     covparams = covparams,
                                     ymodel = ymodel,
                                     intvars = intvars,
                                     interventions = interventions,
                                     int_descript = int_descript,
                                     histories = histories, histvars = histvars,
                                     basecovs = c("cov3"), seed = 1234,
                                     parallel = TRUE, nsamples = 5,
                                     nsimul = nsimul, ncores = ncores)

gform_bin_eof

```

gformula_continuous_eof

Estimation of Continuous End-of-Follow-Up Outcome Under the Parametric G-Formula

Description

Based on an observed data set, this internal function estimates the outcome mean at end-of-follow-up under multiple user-specified interventions using the parametric g-formula. See McGrath et al. (2020) for further details concerning the application and implementation of the parametric g-formula.

Usage

```

gformula_continuous_eof(
  obs_data,
  id,
  time_name,
  covnames,
  covtypes,
  covparams,
  covfits_custom = NA,
  covpredict_custom = NA,
  histvars = NULL,
  histories = NA,
  basecovs = NA,
  outcome_name,
  ymodel,

```

```

  censor_name = NULL,
  censor_model = NA,
  intvars = NULL,
  interventions = NULL,
  int_times = NULL,
  int_descript = NULL,
  ref_int = 0,
  visitprocess = NA,
  restrictions = NA,
  yrestrictions = NA,
  baselags = FALSE,
  nsimul = NA,
  sim_data_b = FALSE,
  seed,
  nsamples = 0,
  parallel = FALSE,
  ncores = NA,
  ci_method = "percentile",
  threads,
  model_fits = FALSE,
  boot_diag = FALSE,
  show_progress = TRUE,
  ipw_cutoff_quantile = NULL,
  ipw_cutoff_value = NULL,
  ...
)

```

Arguments

<code>obs_data</code>	Data table containing the observed data.
<code>id</code>	Character string specifying the name of the ID variable in <code>obs_data</code> .
<code>time_name</code>	Character string specifying the name of the time variable in <code>obs_data</code> .
<code>covnames</code>	Vector of character strings specifying the names of the time-varying covariates in <code>obs_data</code> .
<code>covtypes</code>	Vector of character strings specifying the "type" of each time-varying covariate included in <code>covnames</code> . The possible "types" are: "binary", "normal", "categorical", "bounded normal", "zero-inflated normal", "truncated normal", "absorbing", "categorical time", and "custom".
<code>covparams</code>	List of vectors, where each vector contains information for one parameter used in the modeling of the time-varying covariates (e.g., model statement, family, link function, etc.). Each vector must be the same length as <code>covnames</code> and in the same order. If a parameter is not required for a certain covariate, it should be set to NA at that index.
<code>covfits_custom</code>	Vector containing custom fit functions for time-varying covariates that do not fall within the pre-defined covariate types. It should be in the same order <code>covnames</code> . If a custom fit function is not required for a particular covariate (e.g., if the first covariate is of type "binary" but the second is of type "custom"), then that index should be set to NA. The default is NA.

covpredict_custom	Vector containing custom prediction functions for time-varying covariates that do not fall within the pre-defined covariate types. It should be in the same order as covnames. If a custom prediction function is not required for a particular covariate, then that index should be set to NA. The default is NA.
histvars	List of vectors. The kth vector specifies the names of the variables for which the kth history function in histories is to be applied.
histories	Vector of history functions to apply to the variables specified in histvars. The default is NA.
basecovs	Vector of character strings specifying the names of baseline covariates in obs_data. These covariates are not simulated using a model but rather carry their value over all time points from the first time point of obs_data. These covariates should not be included in covnames. The default is NA.
outcome_name	Character string specifying the name of the outcome variable in obs_data.
ymodel	Model statement for the outcome variable.
censor_name	Character string specifying the name of the censoring variable in obs_data. Only applicable when using inverse probability weights to estimate the natural course means / risk from the observed data. See "Details".
censor_model	Model statement for the censoring variable. Only applicable when using inverse probability weights to estimate the natural course means / risk from the observed data. See "Details".
intvars	List, whose elements are vectors of character strings. The kth vector in intvars specifies the name(s) of the variable(s) to be intervened on in each round of the simulation under the kth intervention in interventions.
interventions	List, whose elements are lists of vectors. Each list in interventions specifies a unique intervention on the relevant variable(s) in intvars. Each vector contains a function implementing a particular intervention on a single variable, optionally followed by one or more "intervention values" (i.e., integers used to specify the treatment regime).
int_times	List, whose elements are lists of vectors. The kth list in int_times corresponds to the kth intervention in interventions. Each vector specifies the time points in which the relevant intervention is applied on the corresponding variable in intvars. When an intervention is not applied, the simulated natural course value is used. By default, this argument is set so that all interventions are applied in all time points.
int_descript	Vector of character strings, each describing an intervention. It must be in same order as the entries in interventions.
ref_int	Integer denoting the intervention to be used as the reference for calculating the end-of-follow-up mean ratio and mean difference. 0 denotes the natural course, while subsequent integers denote user-specified interventions in the order that they are named in interventions. The default is 0.
visitprocess	List of vectors. Each vector contains as its first entry the covariate name of a visit process; its second entry the name of a covariate whose modeling depends on the visit process; and its third entry the maximum number of consecutive visits that can be missed before an individual is censored. The default is NA.

restrictions	List of vectors. Each vector contains as its first entry a covariate for which <i>a priori</i> knowledge of its distribution is available; its second entry a condition under which no knowledge of its distribution is available and that must be TRUE for the distribution of that covariate given that condition to be estimated via a parametric model or other fitting procedure; its third entry a function for estimating the distribution of that covariate given the condition in the second entry is false such that <i>a priori</i> knowledge of the covariate distribution is available; and its fourth entry a value used by the function in the third entry. The default is NA.
yrestrictions	List of vectors. Each vector contains as its first entry a condition and its second entry an integer. When the condition is TRUE, the outcome variable is simulated according to the fitted model; when the condition is FALSE, the outcome variable takes on the value in the second entry. The default is NA.
baselags	Logical scalar for specifying the convention used for <code>lagi</code> and <code>lag_cumavgi</code> terms in the model statements when pre-baseline times are not included in <code>obs_data</code> and when the current time index, t , is such that $t < i$. If this argument is set to FALSE, the value of all <code>lagi</code> and <code>lag_cumavgi</code> terms in this context are set to 0 (for non-categorical covariates) or the reference level (for categorical covariates). If this argument is set to TRUE, the value of <code>lagi</code> and <code>lag_cumavgi</code> terms are set to their values at time 0. The default is FALSE.
nsimul	Number of subjects for whom to simulate data. By default, this argument is set equal to the number of subjects in <code>obs_data</code> .
sim_data_b	Logical scalar indicating whether to return the simulated data set. If bootstrap samples are used (i.e., <code>nsamples</code> is set to a value greater than 0), this argument must be set to FALSE. The default is FALSE.
seed	Starting seed for simulations and bootstrapping.
nsamples	Integer specifying the number of bootstrap samples to generate. The default is 0.
parallel	Logical scalar indicating whether to parallelize simulations of different interventions to multiple cores.
ncores	Integer specifying the number of CPU cores to use in parallel simulation. This argument is required when <code>parallel</code> is set to TRUE. In many applications, users may wish to set this argument equal to <code>parallel::detectCores() - 1</code> .
ci_method	Character string specifying the method for calculating the bootstrap 95% confidence intervals, if applicable. The options are "percentile" and "normal".
threads	Integer specifying the number of threads to be used in <code>data.table</code> . See setDTthreads for further details.
model_fits	Logical scalar indicating whether to return the fitted models. Note that if this argument is set to TRUE, the output of this function may use a lot of memory. The default is FALSE.
boot_diag	Logical scalar indicating whether to return the coefficients, standard errors, and variance-covariance matrices of the parameters of the fitted models in the bootstrap samples. The default is FALSE.
show_progress	Logical scalar indicating whether to print a progress bar for the number of bootstrap samples completed in the R console. This argument is only applicable when <code>parallel</code> is set to FALSE and bootstrap samples are used (i.e., <code>nsamples</code> is set to a value greater than 0). The default is TRUE.

ipw_cutoff_quantile	Percentile by which to truncate inverse probability weights. The default is NULL (i.e., no truncation). See "Details".
ipw_cutoff_value	Cutoff value by which to truncate inverse probability weights. The default is NULL (i.e., no truncation). See "Details".
...	Other arguments, which are passed to the functions in covpredict_custom.

Details

To assess model misspecification in the parametric g-formula, users can obtain inverse probability (IP) weighted estimates of the natural course means of the time-varying covariates from the observed data. See Chiu et al. (2022) for details. In addition to the general requirements described in McGrath et al. (2020), the requirements for the input data set and the call to the gformula function for such analyses are described below.

Users need to include a column in `obs_data` with a time-varying censoring variable. Users need to indicate the name of the censoring variable and a model statement for the censoring variable with parameters `sensor_name` and `sensor_model`, respectively. Finally, users can specify how to truncate IP weights with the `ipw_cutoff_quantile` or `ipw_cutoff_value` parameters.

Value

An object of class `gformula_continuous_eof`. The object is a list with the following components:

<code>result</code>	Results table containing the estimated mean outcome for all interventions (including natural course) at the last time point. If bootstrapping was used, the results table includes the bootstrap end-of-follow-up mean ratio, standard error, and 95% confidence interval.
<code>coeffs</code>	A list of the coefficients of the fitted models.
<code>stderrs</code>	A list of the standard errors of the coefficients of the fitted models.
<code>vcovs</code>	A list of the variance-covariance matrices of the parameters of the fitted models.
<code>rmses</code>	A list of root mean square error (RMSE) values of the fitted models.
<code>fits</code>	A list of the fitted models for the time-varying covariates and outcome. If <code>model_fits</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
<code>sim_data</code>	A list of data tables of the simulated data. Each element in the list corresponds to one of the interventions. If the argument <code>sim_data_b</code> is set to <code>FALSE</code> , a value of <code>NA</code> is given.
<code>IP_weights</code>	A numeric vector specifying the inverse probability weights. See "Details".
<code>bootcoeffs</code>	A list, where the <code>k</code> th element is a list containing the coefficients of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
<code>bootstderrs</code>	A list, where the <code>k</code> th element is a list containing the standard errors of the coefficients of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.

bootvcovs A list, where the kth element is a list containing the variance-covariance matrices of the parameters of the fitted models corresponding to the kth bootstrap sample. If boot_diag is set to FALSE, a value of NULL is given.

... Some additional elements.

The results for the g-formula simulation under various interventions for the last time point are printed with the `print.gformula_continuous_eof` function. To generate graphs comparing the mean estimated and observed covariate values over time, use the `print.gformula_continuous_eof` function.

References

Chiu YH, Wen L, McGrath S, Logan R, Dahabreh IJ, Hernán MA. Assessing model misspecification of the parametric g-formula in the presence of censoring. In preparation.

McGrath S, Lin V, Zhang Z, Petito LC, Logan RW, Hernán MA, and JG Young. gfoRmula: An R package for estimating the effects of sustained treatment strategies via the parametric g-formula. *Patterns*. 2020;1:100008.

Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period: application to the healthy worker survivor effect. *Mathematical Modelling*. 1986;7:1393–1512. [Errata (1987) in *Computers and Mathematics with Applications* 14, 917.-921. Addendum (1987) in *Computers and Mathematics with Applications* 14, 923-.945. Errata (1987) to addendum in *Computers and Mathematics with Applications* 18, 477.].

See Also

[gformula](#)

Examples

```
## Estimating the effect of treatment strategies on the mean of a continuous
## end of follow-up outcome

library('Hmisc')
id <- 'id'
time_name <- 't0'
covnames <- c('L1', 'L2', 'A')
outcome_name <- 'Y'
covtypes <- c('categorical', 'normal', 'binary')
histories <- c(lagged)
histvars <- list(c('A', 'L1', 'L2'))
covparams <- list(covmodels = c(L1 ~ lag1_A + lag1_L1 + L3 + t0 +
                               rcspline.eval(lag1_L2, knots = c(-1, 0, 1)),
                               L2 ~ lag1_A + L1 + lag1_L1 + lag1_L2 + L3 + t0,
                               A ~ lag1_A + L1 + L2 + lag1_L1 + lag1_L2 + L3 + t0))
ymodel <- Y ~ A + L1 + L2 + lag1_A + lag1_L1 + lag1_L2 + L3
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, 7))),
                      list(c(static, rep(1, 7))))
int_descript <- c('Never treat', 'Always treat')
```

```
nsimul <- 10000

gform_cont_eof <- gformula_continuous_eof(obs_data = continuous_eofdata,
                                         id = id,
                                         time_name = time_name,
                                         covnames = covnames,
                                         outcome_name = outcome_name,
                                         covtypes = covtypes,
                                         covparams = covparams, ymodel = ymodel,
                                         intvars = intvars,
                                         interventions = interventions,
                                         int_descript = int_descript,
                                         histories = histories, histvars = histvars,
                                         basecovs = c("L3"),
                                         nsimul = nsimul, seed = 1234)

gform_cont_eof
```

gformula_survival

Estimation of Survival Outcome Under the Parametric G-Formula

Description

Based on an observed data set, this internal function estimates the risk over time under multiple user-specified interventions using the parametric g-formula. See McGrath et al. (2020) for further details concerning the application and implementation of the parametric g-formula.

Usage

```
gformula_survival(
  obs_data,
  id,
  time_points = NULL,
  time_name,
  covnames,
  covtypes,
  covparams,
  covfits_custom = NA,
  covpredict_custom = NA,
  histvars = NULL,
  histories = NA,
  basecovs = NA,
  outcome_name,
  ymodel,
  compevent_name = NULL,
  compevent_model = NA,
  compevent_cens = FALSE,
```

```

  censor_name = NULL,
  censor_model = NA,
  intvars = NULL,
  interventions = NULL,
  int_times = NULL,
  int_descript = NULL,
  ref_int = 0,
  intcomp = NA,
  visitprocess = NA,
  restrictions = NA,
  yrestrictions = NA,
  compevent_restrictions = NA,
  baselags = FALSE,
  nsimul = NA,
  sim_data_b = FALSE,
  seed,
  nsamples = 0,
  parallel = FALSE,
  ncores = NA,
  ci_method = "percentile",
  threads,
  model_fits = FALSE,
  boot_diag = FALSE,
  show_progress = TRUE,
  ipw_cutoff_quantile = NULL,
  ipw_cutoff_value = NULL,
  ...
)

```

Arguments

<code>obs_data</code>	Data table containing the observed data.
<code>id</code>	Character string specifying the name of the ID variable in <code>obs_data</code> .
<code>time_points</code>	Number of time points to simulate. By default, this argument is set equal to the maximum number of records that <code>obs_data</code> contains for any individual.
<code>time_name</code>	Character string specifying the name of the time variable in <code>obs_data</code> .
<code>covnames</code>	Vector of character strings specifying the names of the time-varying covariates in <code>obs_data</code> .
<code>covtypes</code>	Vector of character strings specifying the "type" of each time-varying covariate included in <code>covnames</code> . The possible "types" are: "binary", "normal", "categorical", "bounded normal", "zero-inflated normal", "truncated normal", "absorbing", "categorical time", and "custom".
<code>covparams</code>	List of vectors, where each vector contains information for one parameter used in the modeling of the time-varying covariates (e.g., model statement, family, link function, etc.). Each vector must be the same length as <code>covnames</code> and in the same order. If a parameter is not required for a certain covariate, it should be set to NA at that index.

covfits_custom	Vector containing custom fit functions for time-varying covariates that do not fall within the pre-defined covariate types. It should be in the same order covnames. If a custom fit function is not required for a particular covariate (e.g., if the first covariate is of type "binary" but the second is of type "custom"), then that index should be set to NA. The default is NA.
covpredict_custom	Vector containing custom prediction functions for time-varying covariates that do not fall within the pre-defined covariate types. It should be in the same order as covnames. If a custom prediction function is not required for a particular covariate, then that index should be set to NA. The default is NA.
histvars	List of vectors. The kth vector specifies the names of the variables for which the kth history function in histories is to be applied.
histories	Vector of history functions to apply to the variables specified in histvars. The default is NA.
basecovs	Vector of character strings specifying the names of baseline covariates in obs_data. These covariates are not simulated using a model but rather carry their value over all time points from the first time point of obs_data. These covariates should not be included in covnames. The default is NA.
outcome_name	Character string specifying the name of the outcome variable in obs_data.
ymodel	Model statement for the outcome variable.
compevent_name	Character string specifying the name of the competing event variable in obs_data.
compevent_model	Model statement for the competing event variable. The default is NA.
compevent_cens	Logical scalar indicating whether to treat competing events as censoring events. This argument is only applicable for survival outcomes and when a competing even model is supplied (i.e., compevent_name and compevent_model are specified). If this argument is set to TRUE, the competing event model will only be used to construct inverse probability weights to estimate the natural course means / risk from the observed data. If this argument is set to FALSE, the competing event model will be used in the parametric g-formula estimates of the risk and will not be used to construct inverse probability weights. See "Details". The default is FALSE.
censor_name	Character string specifying the name of the censoring variable in obs_data. Only applicable when using inverse probability weights to estimate the natural course means / risk from the observed data. See "Details".
censor_model	Model statement for the censoring variable. Only applicable when using inverse probability weights to estimate the natural course means / risk from the observed data. See "Details".
intvars	List, whose elements are vectors of character strings. The kth vector in intvars specifies the name(s) of the variable(s) to be intervened on in each round of the simulation under the kth intervention in interventions.
interventions	List, whose elements are lists of vectors. Each list in interventions specifies a unique intervention on the relevant variable(s) in intvars. Each vector contains a function implementing a particular intervention on a single variable, optionally followed by one or more "intervention values" (i.e., integers used to specify the treatment regime).

int_times	List, whose elements are lists of vectors. The kth list in int_times corresponds to the kth intervention in interventions. Each vector specifies the time points in which the relevant intervention is applied on the corresponding variable in intvars. When an intervention is not applied, the simulated natural course value is used. By default, this argument is set so that all interventions are applied in all time points.
int_descript	Vector of character strings, each describing an intervention. It must be in same order as the entries in interventions.
ref_int	Integer denoting the intervention to be used as the reference for calculating the risk ratio and risk difference. 0 denotes the natural course, while subsequent integers denote user-specified interventions in the order that they are named in interventions. The default is 0.
intcomp	List of two numbers indicating a pair of interventions to be compared by a hazard ratio. The default is NA, resulting in no hazard ratio calculation.
visitprocess	List of vectors. Each vector contains as its first entry the covariate name of a visit process; its second entry the name of a covariate whose modeling depends on the visit process; and its third entry the maximum number of consecutive visits that can be missed before an individual is censored. The default is NA.
restrictions	List of vectors. Each vector contains as its first entry a covariate for which <i>a priori</i> knowledge of its distribution is available; its second entry a condition under which no knowledge of its distribution is available and that must be TRUE for the distribution of that covariate given that condition to be estimated via a parametric model or other fitting procedure; its third entry a function for estimating the distribution of that covariate given the condition in the second entry is false such that <i>a priori</i> knowledge of the covariate distribution is available; and its fourth entry a value used by the function in the third entry. The default is NA.
yrestrictions	List of vectors. Each vector contains as its first entry a condition and its second entry an integer. When the condition is TRUE, the outcome variable is simulated according to the fitted model; when the condition is FALSE, the outcome variable takes on the value in the second entry. The default is NA.
compevent_restrictions	List of vectors. Each vector contains as its first entry a condition and its second entry an integer. When the condition is TRUE, the competing event variable is simulated according to the fitted model; when the condition is FALSE, the competing event variable takes on the value in the second entry. The default is NA.
baselags	Logical scalar for specifying the convention used for lagi and lag_cumavgi terms in the model statements when pre-baseline times are not included in obs_data and when the current time index, t , is such that $t < i$. If this argument is set to FALSE, the value of all lagi and lag_cumavgi terms in this context are set to 0 (for non-categorical covariates) or the reference level (for categorical covariates). If this argument is set to TRUE, the value of lagi and lag_cumavgi terms are set to their values at time 0. The default is FALSE.
nsimul	Number of subjects for whom to simulate data. By default, this argument is set equal to the number of subjects in obs_data.

sim_data_b	Logical scalar indicating whether to return the simulated data set. If bootstrap samples are used (i.e., nsamples is set to a value greater than 0), this argument must be set to FALSE. The default is FALSE.
seed	Starting seed for simulations and bootstrapping.
nsamples	Integer specifying the number of bootstrap samples to generate. The default is 0.
parallel	Logical scalar indicating whether to parallelize simulations of different interventions to multiple cores.
ncores	Integer specifying the number of CPU cores to use in parallel simulation. This argument is required when parallel is set to TRUE. In many applications, users may wish to set this argument equal to <code>parallel::detectCores() - 1</code> .
ci_method	Character string specifying the method for calculating the bootstrap 95% confidence intervals, if applicable. The options are "percentile" and "normal".
threads	Integer specifying the number of threads to be used in data.table. See setDTthreads for further details.
model_fits	Logical scalar indicating whether to return the fitted models. Note that if this argument is set to TRUE, the output of this function may use a lot of memory. The default is FALSE.
boot_diag	Logical scalar indicating whether to return the coefficients, standard errors, and variance-covariance matrices of the parameters of the fitted models in the bootstrap samples. The default is FALSE.
show_progress	Logical scalar indicating whether to print a progress bar for the number of bootstrap samples completed in the R console. This argument is only applicable when parallel is set to FALSE and bootstrap samples are used (i.e., nsamples is set to a value greater than 0). The default is TRUE.
ipw_cutoff_quantile	Percentile by which to truncate inverse probability weights. The default is NULL (i.e., no truncation). See "Details".
ipw_cutoff_value	Cutoff value by which to truncate inverse probability weights. The default is NULL (i.e., no truncation). See "Details".
...	Other arguments, which are passed to the functions in covpredict_custom.

Details

To assess model misspecification in the parametric g-formula, users can obtain inverse probability (IP) weighted estimates of the natural course risk and/or means of the time-varying covariates from the observed data. See Chiu et al. (2022) for details. In addition to the general requirements described in McGrath et al. (2020), the requirements for the input data set and the call to the gformula function for such analyses are described below.

Users need to include a column in obs_data with a time-varying censoring variable. Users need to indicate the name of the censoring variable and a model statement for the censoring variable with parameters censor_name and censor_model, respectively. When competing events are present, users need to include a column in obs_data with a time-varying indicator of the competing event variable and need to indicate the name of the competing event variable and the corresponding model

statement with parameters `compevent_name` and `compevent_model`, respectively. Users need to indicate whether to treat competing events as censoring events with the `compevent_cens` parameter. Finally, users can specify how to truncate IP weights with the `ipw_cutoff_quantile` or `ipw_cutoff_value` parameters.

Value

An object of class `gformula_survival`. The object is a list with the following components:

<code>result</code>	Results table containing the estimated risk and risk ratio for all interventions (including the natural course) at each time point. If bootstrapping was used, the results table includes the bootstrap mean risk ratio, standard error, and 95% confidence interval.
<code>coeffs</code>	A list of the coefficients of the fitted models.
<code>stderrs</code>	A list of the standard errors of the coefficients of the fitted models.
<code>vcovs</code>	A list of the variance-covariance matrices of the parameters of the fitted models.
<code>rmses</code>	A list of root mean square error (RMSE) values of the fitted models.
<code>hazardratio_val</code>	Hazard ratio between two interventions (if applicable).
<code>fits</code>	A list of the fitted models for the time-varying covariates, outcome, and competing event (if applicable). If <code>model_fits</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
<code>sim_data</code>	A list of data tables of the simulated data. Each element in the list corresponds to one of the interventions. If the argument <code>sim_data_b</code> is set to <code>FALSE</code> , a value of <code>NA</code> is given.
<code>IP_weights</code>	A numeric vector specifying the inverse probability weights. See "Details".
<code>bootcoeffs</code>	A list, where the <code>k</code> th element is a list containing the coefficients of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
<code>bootstderrs</code>	A list, where the <code>k</code> th element is a list containing the standard errors of the coefficients of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
<code>bootvcovs</code>	A list, where the <code>k</code> th element is a list containing the variance-covariance matrices of the parameters of the fitted models corresponding to the <code>k</code> th bootstrap sample. If <code>boot_diag</code> is set to <code>FALSE</code> , a value of <code>NULL</code> is given.
<code>...</code>	Some additional elements.

The results for the g-formula simulation under various interventions only for the first and last time points are printed with the `print.gformula_survival` function. To generate graphs comparing the mean estimated covariate values and risks over time and mean observed covariate values and risks over time, use the `plot.gformula_survival` function.

References

Chiu YH, Wen L, McGrath S, Logan R, Dahabreh IJ, Hernán MA. Assessing model misspecification of the parametric g-formula in the presence of censoring. In preparation.


```

## Estimating the effect of treatment strategies on risk of a failure event
## when competing events exist

id <- 'id'
time_points <- 7
time_name <- 't0'
covnames <- c('L1', 'L2', 'A')
outcome_name <- 'Y'
compevent_name <- 'D'
covtypes <- c('binary', 'bounded normal', 'binary')
histories <- c(lagged, lagavg)
histvars <- list(c('A', 'L1', 'L2'), c('L1', 'L2'))
covparams <- list(covlink = c('logit', 'identity', 'logit'),
                 covmodels = c(L1 ~ lag1_A + lag_cumavg1_L1 + lag_cumavg1_L2 +
                               L3 + as.factor(t0),
                               L2 ~ lag1_A + L1 + lag_cumavg1_L1 +
                               lag_cumavg1_L2 + L3 + as.factor(t0),
                               A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 +
                               lag_cumavg1_L2 + L3 + as.factor(t0)))
ymodel <- Y ~ A + L1 + L2 + lag1_A + lag1_L1 + lag1_L2 + L3 + as.factor(t0)
compevent_model <- D ~ A + L1 + L2 + lag1_A + lag1_L1 + lag1_L2 + L3 + as.factor(t0)
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, time_points))),
                      list(c(static, rep(1, time_points))))
int_descript <- c('Never treat', 'Always treat')
nsimul <- 10000

gform_basic <- gformula_survival(obs_data = basicdata, id = id,
                                time_points = time_points,
                                time_name = time_name, covnames = covnames,
                                outcome_name = outcome_name,
                                compevent_name = compevent_name,
                                covtypes = covtypes,
                                covparams = covparams, ymodel = ymodel,
                                compevent_model = compevent_model,
                                intvars = intvars, interventions = interventions,
                                int_descript = int_descript,
                                histories = histories, histvars = histvars,
                                basecovs = c('L3'), nsimul = nsimul,
                                seed = 1234)

gform_basic

## Using IP weighting to estimate natural course risk
## Only the natural course intervention is included for simplicity

covnames <- c('L', 'A')
histories <- c(lagged)
histvars <- list(c('A', 'L'))
ymodel <- Y ~ L + A

```

```

covtypes <- c('binary', 'normal')
covparams <- list(covmodels = c(L ~ lag1_L + lag1_A,
                                A ~ lag1_L + L + lag1_A))

censor_name <- 'C'
censor_model <- C ~ L
res_censor <- gformula(obs_data = censor_data, id = 'id',
                       time_name = 't0', covnames = covnames,
                       outcome_name = 'Y', outcome_type = 'survival',
                       censor_name = censor_name, censor_model = censor_model,
                       covtypes = covtypes,
                       covparams = covparams, ymodel = ymodel,
                       intvars = NULL, interventions = NULL, int_descript = NULL,
                       histories = histories, histvars = histvars,
                       seed = 1234)

plot(res_censor)

```

lagged

History functions

Description

These functions create new columns in an input data table for covariate histories. Users must specify which covariates are to be used in the history functions.

Usage

```

lagged(
  pool,
  histvars,
  histvals,
  time_name,
  t,
  id_name,
  baselags,
  below_zero_indicator
)

```

```
cumavg(pool, histvars, time_name, t, id_name, below_zero_indicator)
```

```

lagavg(
  pool,
  histvars,
  histvals,
  time_name,
  t,
  id_name,

```

```

    baselags,
    below_zero_indicator
  )

```

Arguments

<code>pool</code>	Data table containing all information prior to time t (t noninclusive).
<code>histvars</code>	Vector of character strings specifying the names of the variables for which history functions are to be applied.
<code>histvals</code>	For <code>lagged</code> , this argument is a vector specifying the lags used in the model statements (e.g., if <code>lag1_varname</code> and <code>lag2_varname</code> were included in the model statements, this vector would be <code>c(1, 2)</code>). For <code>lagavg</code> , this argument is a numeric vector specifying the lag averages used in the model statements.
<code>time_name</code>	Character string specifying the name of the time variable in <code>pool</code> .
<code>t</code>	Integer specifying the current time index.
<code>id_name</code>	Character string specifying the name of the ID variable in <code>pool</code> .
<code>baselags</code>	Logical scalar for specifying the convention used for <code>lagi</code> and <code>lag_cumavgi</code> terms in the model statements when pre-baseline times are not included in <code>obs_data</code> and when the current time index, t , is such that $t < i$. If this argument is set to <code>FALSE</code> , the value of all <code>lagi</code> and <code>lag_cumavgi</code> terms in this context are set to 0 (for non-categorical covariates) or the reference level (for categorical covariates). If this argument is set to <code>TRUE</code> , the value of <code>lagi</code> and <code>lag_cumavgi</code> terms are set to their values at time 0. The default is <code>FALSE</code> .
<code>below_zero_indicator</code>	Logical scalar indicating whether the observed data set contains rows for time $t < 0$.

Details

`lagged` creates new columns for lagged versions of existing variables in the dataset. The user must specify which variables are to be lagged.

`cumavg` creates new columns for the cumulative average up until time t of existing variables in the dataset.

`lagavg` creates new columns for the "lagged cumulative average" (cumulative average up until time t , then lagged by one time unit) up until time t of existing variables in the dataset.

Value

No value is returned. The data table `pool` is modified in place.

Examples

```

## Estimating the effect of static treatment strategies on risk of a
## failure event

id <- 'id'
time_points <- 7

```



```

time_name <- 't0'
covnames <- c('L1', 'L2', 'A')
outcome_name <- 'Y'
covtypes <- c('binary', 'bounded normal', 'binary')
histories <- c(lagged, lagavg)
histvars <- list(c('A', 'L1', 'L2'), c('L1', 'L2'))
covparams <- list(covmodels = c(L1 ~ lag1_A + lag_cumavg1_L1 + lag_cumavg1_L2 +
                                L3 + t0,
                                L2 ~ lag1_A + L1 + lag_cumavg1_L1 +
                                lag_cumavg1_L2 + L3 + t0,
                                A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 +
                                lag_cumavg1_L2 + L3 + t0))
ymodel <- Y ~ A + L1 + L2 + L3 + lag1_A + lag1_L1 + lag1_L2 + t0
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, time_points))),
                      list(c(static, rep(1, time_points))))
int_descript <- c('Never treat', 'Always treat')
nsimul <- 10000

gform_basic <- gformula_survival(obs_data = basicdata_nocomp, id = id,
                                time_points = time_points,
                                time_name = time_name, covnames = covnames,
                                outcome_name = outcome_name,
                                covtypes = covtypes,
                                covparams = covparams, ymodel = ymodel,
                                intvars = intvars,
                                interventions = interventions,
                                int_descript = int_descript,
                                histories = histories, histvars = histvars,
                                basecovs = c('L3'), nsimul = nsimul,
                                seed = 1234)

gform_basic

```

```
plot.gformula_binary_eof
```

Plot method for objects of class "gformula_binary_eof"

Description

This function generates graphs of the mean simulated vs. observed values at each time point of the time-varying covariates under the natural course. For categorical covariates, the observed and simulated probability of each level are plotted at each time point.

Usage

```

## S3 method for class 'gformula_binary_eof'
plot(
  x,

```

```

covnames = NULL,
ncol = NULL,
nrow = NULL,
common.legend = TRUE,
legend = "bottom",
xlab = NULL,
ylab_cov = NULL,
...
)

```

Arguments

x	Object of class "gformula_binary_eof".
covnames	Vector of character strings specifying the names of the time-varying covariates to be plotted. The ordering of covariates given here is used in the plot grid. Time-varying covariates of type "categorical time" cannot be included. By default, this argument is set equal to the covnames argument used in gformula_binary_eof , where covariates of type "categorical time" are removed.
ncol	Number of columns in the plot grid. By default, two columns are used when there is at least two plots.
nrow	Number of rows in the plot grid. By default, a maximum of six rows is used and additional plots are included in subsequent pages.
common.legend	Logical scalar indicating whether to include a legend. The default is TRUE.
legend	Character string specifying the legend position. Valid values are "top", "bottom", "left", "right", and "none". The default is "bottom".
xlab	Character string for the x axes of all plots. By default, this argument is set to the time_name argument specified in gformula_binary_eof .
ylab_cov	Vector of character strings for the y axes of the plots for the covariates. This argument must be the same length as covnames. The i-th element of this argument corresponds to the plot for the i-th element of covnames.
...	Other arguments, which are passed to ggarrange .

Value

An object of class "ggarrange". See documentation of [ggarrange](#).

See Also

[gformula_binary_eof](#)

Examples

```

## Estimating the effect of threshold interventions on the mean of a binary
## end of follow-up outcome

id <- 'id_num'
time_name <- 'time'

```

```

covnames <- c('cov1', 'cov2', 'treat')
outcome_name <- 'outcome'
histories <- c(lagged, cumavg)
histvars <- list(c('treat', 'cov1', 'cov2'), c('cov1', 'cov2'))
covtypes <- c('binary', 'zero-inflated normal', 'normal')
covparams <- list(covmodels = c(cov1 ~ lag1_treat + lag1_cov1 + lag1_cov2 + cov3 +
                                time,
                                cov2 ~ lag1_treat + cov1 + lag1_cov1 + lag1_cov2 +
                                cov3 + time,
                                treat ~ lag1_treat + cumavg_cov1 +
                                cumavg_cov2 + cov3 + time))
ymodel <- outcome ~ treat + cov1 + cov2 + lag1_cov1 + lag1_cov2 + cov3
intvars <- list('treat', 'treat')
interventions <- list(list(c(static, rep(0, 7))),
                      list(c(threshold, 1, Inf)))
int_descript <- c('Never treat', 'Threshold - lower bound 1')
nsimul <- 10000
ncores <- 2

gform_bin_eof <- gformula_binary_eof(obs_data = binary_eofdata, id = id,
                                     time_name = time_name,
                                     covnames = covnames,
                                     outcome_name = outcome_name,
                                     covtypes = covtypes,
                                     covparams = covparams,
                                     ymodel = ymodel,
                                     intvars = intvars,
                                     interventions = interventions,
                                     int_descript = int_descript,
                                     histories = histories, histvars = histvars,
                                     basecovs = c("cov3"), seed = 1234,
                                     parallel = TRUE, nsamples = 5,
                                     nsimul = nsimul, ncores = ncores)

plot(gform_bin_eof)

```

```
plot.gformula_continuous_eof
```

Plot method for objects of class "gformula_continuous_eof"

Description

This function generates graphs of the mean simulated vs. observed values at each time point of the time-varying covariates under the natural course. For categorical covariates, the observed and simulated probability of each level are plotted at each time point.

Usage

```
## S3 method for class 'gformula_continuous_eof'
```

```

plot(
  x,
  covnames = NULL,
  ncol = NULL,
  nrow = NULL,
  common.legend = TRUE,
  legend = "bottom",
  xlab = NULL,
  ylab_cov = NULL,
  ...
)

```

Arguments

<code>x</code>	Object of class "gformula_continuous_eof".
<code>covnames</code>	Vector of character strings specifying the names of the time-varying covariates to be plotted. The ordering of covariates given here is used in the plot grid. Time-varying covariates of type "categorical time" cannot be included. By default, this argument is set equal to the <code>covnames</code> argument used in gformula_continuous_eof , where covariates of type "categorical time" are removed.
<code>ncol</code>	Number of columns in the plot grid. By default, two columns are used when there is at least two plots.
<code>nrow</code>	Number of rows in the plot grid. By default, a maximum of six rows is used and additional plots are included in subsequent pages.
<code>common.legend</code>	Logical scalar indicating whether to include a legend. The default is TRUE.
<code>legend</code>	Character string specifying the legend position. Valid values are "top", "bottom", "left", "right", and "none". The default is "bottom".
<code>xlab</code>	Character string for the x axes of all plots. By default, this argument is set to the <code>time_name</code> argument specified in gformula_continuous_eof .
<code>ylab_cov</code>	Vector of character strings for the y axes of the plots for the covariates. This argument must be the same length as <code>covnames</code> . The <i>i</i> -th element of this argument corresponds to the plot for the <i>i</i> -th element of <code>covnames</code> .
<code>...</code>	Other arguments, which are passed to ggarrange .

Value

An object of class "ggarrange". See documentation of [ggarrange](#).

See Also

[gformula_continuous_eof](#)

Examples

```

## Estimating the effect of treatment strategies on the mean of a continuous
## end of follow-up outcome

```

```

library('Hmisc')
id <- 'id'
time_name <- 't0'
covnames <- c('L1', 'L2', 'A')
outcome_name <- 'Y'
covtypes <- c('categorical', 'normal', 'binary')
histories <- c('lagged')
histvars <- list(c('A', 'L1', 'L2'))
covparams <- list(covmodels = c(L1 ~ lag1_A + lag1_L1 + L3 + t0 +
                                rcspline.eval(lag1_L2, knots = c(-1, 0, 1)),
                                L2 ~ lag1_A + L1 + lag1_L1 + lag1_L2 + L3 + t0,
                                A ~ lag1_A + L1 + L2 + lag1_L1 + lag1_L2 + L3 + t0))
ymodel <- Y ~ A + L1 + L2 + lag1_A + lag1_L1 + lag1_L2 + L3
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, 7))),
                      list(c(static, rep(1, 7))))
int_descript <- c('Never treat', 'Always treat')
nsimul <- 10000

gform_cont_eof <- gformula_continuous_eof(obs_data = continuous_eofdata,
                                         id = id,
                                         time_name = time_name,
                                         covnames = covnames,
                                         outcome_name = outcome_name,
                                         covtypes = covtypes,
                                         covparams = covparams, ymodel = ymodel,
                                         intvars = intvars,
                                         interventions = interventions,
                                         int_descript = int_descript,
                                         histories = histories, histvars = histvars,
                                         basecovs = c("L3"),
                                         nsimul = nsimul, seed = 1234)

plot(gform_cont_eof)

```

```
plot.gformula_survival
```

Plot method for objects of class "gformula_survival"

Description

This function generates graphs of the mean simulated vs. observed values at each time point of the time-varying covariates, risk, and survival under the natural course. For categorical covariates, the observed and simulated probability of each level are plotted at each time point.

Usage

```
## S3 method for class 'gformula_survival'
plot(
```

```

x,
covnames = NULL,
risk = TRUE,
survival = FALSE,
ncol = NULL,
nrow = NULL,
common.legend = TRUE,
legend = "bottom",
xlab = NULL,
ylab_cov = NULL,
ylab_risk = "risk",
ylab_surv = "survival",
pos_risk = NULL,
pos_surv = NULL,
ci_risk = FALSE,
...
)

```

Arguments

x	Object of class "gformula_survival".
covnames	Vector of character strings specifying the names of the time-varying covariates to be plotted. The ordering of covariates given here is used in the plot grid. Time-varying covariates of type "categorical time" cannot be included. To plot none of the time-varying covariates, set this argument to NA. By default, this argument is set equal to the covnames argument used in gformula_survival , where covariates of type 'categorical time' are removed.
risk	Logical scalar indicating whether to include a plot for the risk. The default is TRUE.
survival	Logical scalar indicating whether to include a plot for the survival. The default is FALSE.
ncol	Number of columns in the plot grid. By default, two columns are used when there is at least two plots.
nrow	Number of rows in the plot grid. By default, a maximum of six rows is used and additional plots are included in subsequent pages.
common.legend	Logical scalar indicating whether to include a legend. The default is TRUE.
legend	Character string specifying the legend position. Valid values are "top", "bottom", "left", "right", and "none". The default is "bottom".
xlab	Character string for the x axes of all plots. By default, this argument is set to the time_name argument specified in gformula_survival .
ylab_cov	Vector of character strings for the y axes of the plots for the covariates. This argument must be the same length as covnames. The i-th element of this argument corresponds to the plot for the i-th element of covnames.
ylab_risk	Character string for the y axis of the plot for the risk (if applicable). The default is "risk".


```

time_name = time_name, covnames = covnames,
outcome_name = outcome_name,
covtypes = covtypes,
covparams = covparams, ymodel = ymodel,
intvars = intvars,
interventions = interventions,
int_descript = int_descript,
histories = histories, histvars = histvars,
basecovs = c('L3'), nsimul = nsimul,
seed = 1234)

plot(gform_basic)

```

```
print.gformula_survival
```

Print and summary methods for "gformula" objects

Description

Print and summary method for objects of class "gformula_survival", "gformula_continuous_eof", or "gformula_binary_eof".

Usage

```

## S3 method for class 'gformula_survival'
print(
  x,
  all_times = FALSE,
  coefficients = FALSE,
  stderrs = FALSE,
  rmsees = FALSE,
  hazardratio = FALSE,
  fits = FALSE,
  ...
)

## S3 method for class 'gformula_continuous_eof'
print(
  x,
  coefficients = FALSE,
  stderrs = FALSE,
  rmsees = FALSE,
  fits = FALSE,
  ...
)

```



```

## S3 method for class 'gformula_binary_eof'
print(
  x,
  coefficients = FALSE,
  stderrs = FALSE,
  rmses = FALSE,
  fits = FALSE,
  ...
)

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

## S3 method for class 'summary.gformula'
print(
  x,
  all_times = TRUE,
  coefficients = FALSE,
  stderrs = FALSE,
  rmses = FALSE,
  hazardratio = FALSE,
  fits = TRUE,
  ...
)

```

Arguments

<code>x</code>	Object of class "gformula_survival", "gformula_continuous_eof", "gformula_binary_eof", or "summary.gformula" (for print).
<code>all_times</code>	Logical scalar indicating whether to print the results for all time points. This argument is only applicable to objects of class "gformula_survival". If this argument is set to FALSE, the results are only shown for the final time point. The default is FALSE for print and TRUE for summary.
<code>coefficients</code>	Logical scalar indicating whether to print the model coefficients. The default is FALSE.
<code>stderrs</code>	Logical scalar indicating whether to print the standard error of the model coefficients. The default is FALSE.
<code>rmses</code>	Logical scalar indicating whether to print the model root mean square errors (RMSEs). The default is FALSE.
<code>hazardratio</code>	Logical scalar indicating whether to print the hazard ratio between two interventions (if computed). If bootstrapping was used, 95% confidence intervals will be given. This argument is only applicable to objects of class "gformula_survival". The default is FALSE.
<code>fits</code>	Logical scalar indicating whether to print summaries of the fitted models for the time-varying covariates, outcome, and competing event (if applicable). This argument is only effective if the argument <code>model_fits</code> was set to TRUE in <code>gformula</code> . The default is FALSE for print and TRUE for summary.


```

                                lag_cumavg1_L2 + L3 + t0,
                                A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 +
                                lag_cumavg1_L2 + L3 + t0))
ymodel <- Y ~ A + L1 + L2 + L3 + lag1_A + lag1_L1 + lag1_L2 + t0
intvars <- list('A', 'A')
interventions <- list(list(c(static, rep(0, time_points))),
                      list(c(static, rep(1, time_points))))
int_descript <- c('Never treat', 'Always treat')
nsimul <- 10000

# At t0 == 5, assume we have deterministic knowledge that L1 equals 0
restrictions <- list(c('L1', 't0 != 5', simple_restriction, 0))

gform_basic <- gformula_survival(obs_data = basicdata_nocomp, id = id,
                                time_points = time_points,
                                time_name = time_name, covnames = covnames,
                                outcome_name = outcome_name,
                                covtypes = covtypes,
                                covparams = covparams, ymodel = ymodel,
                                intvars = intvars,
                                interventions = interventions,
                                restrictions = restrictions,
                                int_descript = int_descript,
                                histories = histories, histvars = histvars,
                                basecovs = c('L3'), nsimul = nsimul,
                                seed = 1234)

gform_basic

```

 static

Static Intervention

Description

This function implements a static intervention (i.e., either constant treatment or no treatment over all time points) for the specified intervention variable in the data table `newdf`.

Usage

```
static(newdf, pool, intvar, intvals, time_name, t)
```

Arguments

<code>newdf</code>	Data table containing the simulated data at time t .
<code>pool</code>	Data table containing the simulated data at times before t .
<code>intvar</code>	Character string specifying the name of the variable to be intervened on in each round of the simulation.

threshold	<i>Threshold Intervention</i>
-----------	-------------------------------

Description

This function implements a threshold intervention (i.e., once treatment bypasses a certain threshold, it remains at that threshold until end of follow-up) for the specified intervention variable in the data table `newdf`.

Usage

```
threshold(newdf, pool, intvar, intvals, time_name, t)
```

Arguments

<code>newdf</code>	Data table containing the simulated data at time t .
<code>pool</code>	Data table containing the simulated data at times before t .
<code>intvar</code>	Character string specifying the name of the variable to be intervened on in each round of the simulation.
<code>intvals</code>	A list of length 2. The first entry is lower bound of the threshold, and the second entry is the upper bound.
<code>time_name</code>	Character string specifying the name of the time variable in <code>pool</code> and <code>newdf</code> .
<code>t</code>	Integer specifying the current time index.

Value

No value is returned. The data table `newdf` is modified in place.

Examples

```
## Estimating the effect of threshold interventions on the mean of a binary
## end of follow-up outcome

id <- 'id_num'
time_name <- 'time'
covnames <- c('cov1', 'cov2', 'treat')
outcome_name <- 'outcome'
histories <- c(lagged, cumavg)
histvars <- list(c('treat', 'cov1', 'cov2'), c('cov1', 'cov2'))
covtypes <- c('binary', 'zero-inflated normal', 'normal')
covparams <- list(covmodels = c(cov1 ~ lag1_treat + lag1_cov1 + lag1_cov2 + cov3 +
  time,
  cov2 ~ lag1_treat + cov1 + lag1_cov1 + lag1_cov2 +
  cov3 + time,
  treat ~ lag1_treat + cumavg_cov1 +
  cumavg_cov2 + cov3 + time))
```

```

ymodel <- outcome ~ treat + cov1 + cov2 + lag1_cov1 + lag1_cov2 + cov3
intvars <- list('treat', 'treat')
interventions <- list(list(c(static, rep(0, 7))),
                      list(c(threshold, 1, Inf)))
int_descript <- c('Never treat', 'Threshold - lower bound 1')
nsimul <- 10000
ncores <- 2

gform_bin_eof <- gformula_binary_eof(obs_data = binary_eofdata, id = id,
                                   time_name = time_name,
                                   covnames = covnames,
                                   outcome_name = outcome_name,
                                   covtypes = covtypes,
                                   covparams = covparams,
                                   ymodel = ymodel,
                                   intvars = intvars,
                                   interventions = interventions,
                                   int_descript = int_descript,
                                   histories = histories, histvars = histvars,
                                   basecovs = c("cov3"), seed = 1234,
                                   parallel = TRUE, nsamples = 5,
                                   nsimul = nsimul, ncores = ncores)

gform_bin_eof

```

vcov.gformula

Variance-covariance method for objects of class "gformula"

Description

This function extracts the variance-covariance matrices of the parameters of the fitted models for the time-varying covariates, outcome, and competing event (if applicable).

Usage

```

## S3 method for class 'gformula'
vcov(object, ...)

```

Arguments

object	Object of class "gformula".
...	Other arguments.

Value

If `bootdiag` was set to `FALSE` in `gformula`, this function returns a list of the variance-covariance matrices of the parameters of the fitted models to the observed data set. If bootstrapping was used and `bootdiag` was set to `TRUE` in `gformula`, this function returns a list described as follows. The first

Index

* datasets

- basicdata, 2
- basicdata_nocomp, 3
- binary_eofdata, 4
- censor_data, 6
- continuous_eofdata, 8
- continuous_eofdata_pb, 9

- basicdata, 2
- basicdata_nocomp, 3
- binary_eofdata, 4

- carry_forward, 4
- censor_data, 6
- coef.gformula, 6
- continuous_eofdata, 8
- continuous_eofdata_pb, 9
- cumavg (lagged), 39

- gformula, 7, 9, 24, 30, 37, 50, 55, 56
- gformula_binary_eof, 19, 42
- gformula_continuous_eof, 25, 44
- gformula_survival, 31, 46, 47
- ggarrange, 42, 44, 47

- lagavg (lagged), 39
- lagged, 39

- plot.gformula_binary_eof, 15, 24, 41
- plot.gformula_continuous_eof, 15, 43
- plot.gformula_survival, 15, 36, 45
- print.gformula_binary_eof, 15, 24
- print.gformula_binary_eof
 - (print.gformula_survival), 48
- print.gformula_continuous_eof, 15, 30
- print.gformula_continuous_eof
 - (print.gformula_survival), 48
- print.gformula_survival, 15, 36, 48
- print.summary.gformula
 - (print.gformula_survival), 48

- setDTthreads, 13, 22, 28, 35
- simple_restriction, 51
- static, 52
- summary.gformula
 - (print.gformula_survival), 48

- threshold, 54

- vcov.gformula, 55