Package ‘gim’

June 12, 2020

Type Package

Title Generalized Integration Model

Version 0.33.1

Date 2020-06-12

Author Han Zhang, Kai Yu

Maintainer Han Zhang <zhangh.ustc@gmail.com>

Depends R (>= 3.4.0)

Imports numDeriv

Description Implements the generalized integration model, which integrates individual-level data and summary statistics under a generalized linear model framework. It supports continuous and binary outcomes to be modeled by the linear and logistic regression models. For binary outcome, data can be sampled in prospective cohort studies or case-control studies. Described in Zhang et al. (2020)<doi:10.1093/biomet/asaa014>.

License MIT + file LICENSE

URL https://github.com/zhangh12/gim

BugReports https://github.com/zhangh12/gim/issues

LazyData true

NeedsCompilation no

Suggests knitr, rmarkdown

VignetteBuilder knitr

Repository CRAN

Date/Publication 2020-06-12 09:10:24 UTC

R topics documented:

`gim-package` *Generalized Integration Model*

Description

`gim` implements the generalized integration model proposed in Zhang et al. (2018). `gim` integrates individual-level data and summary statistics under a generalized linear model framework. It supports continuous and binary outcomes to be modeled by the linear and logistic regression models. For binary outcome, data can be sampled in prospective cohort studies or case-control studies. `gim` uses ***different*** methods to model binary outcome under ***different*** designs.

Details

Package:	gim
Type:	Package
Version:	0.17.0
Date:	2019-02-25
License:	MIT + file LICENSE

Meta-analysis has become a powerful tool for enhanced inference by gathering evidence from multiple sources. It pools summary-level data, i.e. estiamtes of model coefficients, from different studies to improve estimating efficiency under the assumption that all participating studies are analyzed under the same statistical model. This assumption, however, is usually not true in practice as studies may adjust for different covariates according to specific purpose, or are conduct on partial observed covariates they collect. Meta-analysis can lead to biased estimates when this assumption is violated.

It is challenging to integrate external summary data calculated from different models with a newly conducted internal study in which individual-level data is collected. `gim` is a novel statistical inference framework based on the **G**eneralized **I**ntegration **M**odel, which effectively synthesizes internal and external information according to their variations for multivariate analysis. This new framework is versatile to incorporate various types of summary data from multiple sources. It can be showed that the `gim` estimate is theoretically more efficient than the internal data based maximum likelihood estimate, and the recently developed constraint maximum likelihood estimate that incorporates the outside information.

The `gim` function implemented in this package accounts for the sample sizes shared by different studies. Ignoring this sample overlap may lead to inflated false positive. `gim` requires estimates of coefficients in external working models, but do not rely on their standard errors for two reasons. (1) It is more convenient to request less information from users, especially when the unrequired information could be estimated with other given information. (2) The standard errors reported in literatures where users collect their external data maybe underestimated as a working model rather than a true underlying model was assumed.

The `gim` always requests a set of raw data in which both outcome and independent variables are available. This dataset is called the reference set or internal data. This requirement seems unconvenient as outcome can sometimes be expensive, and some other approaches may be applicable with a reference where only independent variables are collected. Theoretically, an slightly extended `gim` can become workable under the same circumstances, however, we suggest to be more careful to do so. A reference is used to estimate correlation between variables (including outcome) in the population of interest. In practice, there could be a difference in the population of your own study and external studies (from which summary information are collected), therefore, the correlation differ among studies. Conducting an analysis soly relying on a set of independent variables may be biased if the difference between studies is significant. In comparison, a full set of reference consisting of outcome can make the analysis more robust to the potential population difference.

The main function in this package is `gim`.

Author(s)

Han Zhang, Kai Yu

Maintainer: Han Zhang <han.zhang2@nih.gov>

References

Zhang, H., Deng, L., Schiffman, M., Qin, J., Yu, K. (2020) Generalized integration model for improved statistical inference by leveraging external summary data. Biometrika. asaa014, https://doi.org/10.1093/biomet/asaa01

`dat` *Data for example in* `gim`

Description

`dat` is a data frame used in the example of `gim`.

Usage

```
data("dat")
```

Format

A data frame with 4000 observations on the following 6 variables.

`y` a continuous outcome

`d` a binary outcome

`x1` a numeric variable

`x2` a numeric variable

`x3` a numeric variable

`x4` a character variable

Details

This is a dataset from which internal and external data are extracted for the example.

gim *Fitting Generalized Integration Models*

Description

gim is used to fit generalized integration models, which assume linear or logistic regression model on an (internal) individual-level data, while integrating auxiliary or summary information of relevant variables that are estimated from external data, on which different working models could be assumed. gim can work even if partial information from working models are available. Compared to conventional regression model, e.g., glm, that is based on internal data, the estimate of gim method gains additional power by making maximum use of all kinds of available data.

Usage

```
gim(formula, family, data, model, nsample = NULL,
    ncase = NULL, nctrl = NULL, ref = NULL, ...)
```

Arguments

formula
: an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted on the given dataset. More details of model specification are illustrated in ’Details’ and ’Examples’.

family
: a character. "gaussian" for linear regression. For binary outcome fitted by logistic regression, use "binomial" for random sample, or "case-control" for case-control data. gim employs different methods to make inference on random sample and case-control data. If your data are collected in case-control studies, do not use "binomial", otherwise inference may be problematic.

data
: a data frame containing all variables that are specified in formula and model. Incomplete lines will be discarded.

model
: a list describing auxiliary information and working models that are used to generate such information. See ’Details’ and ’Examples’ for more details.

nsample
: a matrix specifying the number of samples shared in datasets that are used to fit the working models given in model. Specify this argument when family is "gaussian" or "binomial", otherwise NULL. See ’Details’ and ’Examples’ for more details.

ncase
: a matrix specifying the number of cases shared in datasets that are used to fit the working models given in model. Specify this argument when family is "case-control", otherwise NULL. ncase and nctrl should be specified simultaneously. See ’Details’ and ’Examples’ for more details.

nctrl
: a matrix specifying the number of controls shared in datasets that are used to fit the working models given in model. Specify this argument when family is case-control, otherwise NULL. See ’Details’ and ’Examples’ for more details.

ref
: a data frame containing the covariates specified in formula and model. It is the reference sample for modeling summary statistics in model. This assumes that ref is sampled from the external population. By default it is NULL which means

that the internal and external populations are the same, therefore gim will use data as the reference. Outcome could be absent or missing in ref because gim will anyway ignore it. See 'Details' for more details.

... for test purpose, use its default value.

Details

formula formula is the model to be used to fit a conventional regression model if no additional information is available. It could be very general as long as it is acceptable to the glm or lm functions. It can eliminate the intercept, y ~ .-1, or involve arithmetic expressions, e.g., log(x), or other operators like * for interactions as.factor(x1)*I(x2 > 0).

model Summary information are calculated on data of external studies, but we do not have access to their raw data. Instead, estimates from working model fitted on external data are given (e.g., reported in literature). The argument model is a list, each component contains information of a working model. Specifically, a component is also a list of two entries form and info, where form is a formula representing the fitted working model, and info is a data frame with two columns var and bet, the names of variables and their estimates from the working model, respectively. Usually the estimate of intercept of a working model is unavailable as people fit but do not reporte it. If user is able to provide such an estimate, the name in column var must be "(Intercept)". See below for an example.

Note that multiple working models could be fitted on the same external data, in that case, the summary information of each working model should be given in model separately. For example, on an external dataset, if two models y ~ x1 and y ~ x2 are fitted, then the estimates of x1 and x2 should be given as two components in model. This happens as many research groups can study the same datasets from different angles.

data gim requires an internal dataset data in which individual-level samples are available. Statistically, this data is critical to provide information of correlation between covariates. This data is also known as the reference data in the literatures. Since general formula is supported in gim, it is important to provide variables in data so that R can find columns of all variables parsed from formulas in formula and model. Read vignettes (upcoming) for more examples about how to create a proper data for gim. We will also release a function to help users with this. gim will discard incomplete lines in data.

nsample Some of summary information can be calculated from datasets that share samples. Ignoring this will lead to underestimated standard error. For example, if a dataset is studied by two different models, the estimates from these two models are not independent but highly correlated. Therefore, this correlation must be properly handled when calculating the standard error of gim estimate, from which a hypothesis testing is conducted. nsample is a squared matrix of dimension p, which is equal to the length of model. Thus, the (i,i) entry in nsample is the number of samples used in fitting the working model specified in model[[i]]$formula, while the (i,k) entry is the number of samples that are involved in fitting working models model[[i]]$formula and model[[k]]$formula. For example, if two working models, e.g., y ~ x1 and y ~ x2 are fitted on the same dataset of 100 samples, then nsample is a matrix of all entries being 100. Read example below and vignettes (upcoming) for more examples.

ncase and nctrl Specify these two arguments when data are sampled from case-control studies. Refer to nsample for their formats.

ref By default, ref is NULL if it is not specified explicitly. This assumes that the internal and external populations are the same, and gim will assign data to ref implicitly. If this assumption holds, and you have additional covariates data (no outcome), e.g. `add.ref`, that also comes from the internal population, you can specified `ref` as `rbind(data,add.ref)` where the column of missing outcome in `add.ref` is set as `NA`. You can also rbind `data` and `add.ref`, with outcome in `data` being deleted. If the external population is different from the internal population, you have to assign `add.ref` to `ref` as reference.

Value

`gim` returns an object of class "`gim`". The function `summary` can be used to print a summary of the results. We will support the use of `anova` in later versions.

The generic accessor functions `coef`, `confint`, and `vcov` can be used to extract coefficients, confidence intervals, and variance-covariance of estimates from the object returned by `gim`.

An object of class "`gim`" is a list containing the following components:

`coefficients`	a named vector of coefficients
`vcov`	the variance-covariance matrix of estimates, including the intercept
`sigma2`	estimated variance of error term in a linear model. Only available for the `gaussian` family
`call`	the matched call
`V.bet`	the variance-covariance matrix of external estimate `bet` in `model`

Author(s)

Han Zhang

References

Zhang, H., Deng, L., Schiffman, M., Qin, J., Yu, K. (2020) Generalized integration model for improved statistical inference by leveraging external summary data. Biometrika. asaa014, https://doi.org/10.1093/biomet/asaa01

Examples

```
## An artificial dataset is lazyloaded to illustrate the concept of GIM method
## It contains:
## A continuous outcome y.
## Four covariates x1, x2, x3, x4 (character).
## A binary outcome d

head(dat)

## internal data of 500 samples
dat0 <- dat[1:500, ]

## three external datasets.
## dat2 and dat3 share some samples
dat1 <- dat[501:1500, c('y', 'x1', 'x2')]
dat2 <- dat[1501:2500, c('y', 'x1', 'x3', 'x4')]
dat3 <- dat[2001:3000, c('y', 'x3', 'x4')]
```

```
## four working models are fitted
form1 <- 'y ~ I(x1 < 0) + I(x2 > 0)'
form2 <- 'y ~ x3 + x4'
form3 <- 'y ~ I(x4 == ”a”)'
form4 <- 'y ~ sqrt(x3)'

## two working models are fitted on dat3
## thus nsample is a 4x4 matrix
nsample <- matrix(c(1000, 0, 0, 0,
                    0, 1000, 500, 500,
                    0, 500, 1000, 1000,
                    0, 500, 1000, 1000),
                  4, 4)

fit1 <- summary(lm(form1, dat1))$coef
fit2 <- summary(lm(form2, dat2))$coef
fit3 <- summary(lm(form3, dat3))$coef ## <-- dat3 is used twice
fit4 <- summary(lm(form4, dat3))$coef ## <-- dat3 is used twice

options(stringsAsFactors = FALSE)
model <- list()
## partial information is available
model[[1]] <- list(form = form1,
                   info = data.frame(var = rownames(fit1)[2],
                                     bet = fit1[2, 1]))

## intercept is provided, but miss estimate of a covariate
model[[2]] <- list(form = form2,
                   info = data.frame(var = rownames(fit2)[1:2],
                                     bet = fit2[1:2, 1]))

model[[3]] <- list(form = form3,
                   info = data.frame(var = rownames(fit3)[2],
                                     bet = fit3[2, 1]))

model[[4]] <- list(form = form4,
                   info = data.frame(var = rownames(fit4)[2],
                                     bet = fit4[2, 1]))

form <- 'y ~ I(x1 < 0) + I(x1 > 1) + x2 * x4 + log(x3) - 1'
fit <- gim(form, 'gaussian', dat0, model, nsample)

summary(fit)
coef(fit)
confint(fit)

# one can compare the gim estimates with those estimated from internal data
fit0 <- lm(form, dat0)
summary(fit0)

# by default, covariates in dat is used as reference in gim
# which assumes that the external and internal populations are the same
fit1 <- gim(form, 'gaussian', dat0, model, nsample, ref = dat0)
```

```
all(coef(fit) == coef(fit1)) # TRUE

# if additional reference is available,
# and it comes from the internal population from which dat is sampled
# gim can use it
add.ref <- dat[3001:3500, ]
add.ref$y <- NA ## <-- outcome is unavailable in reference
ref <- rbind(dat0, add.ref)
fit2 <- gim(form, 'gaussian', dat0, model, nsample, ref = ref)

# if the external population is different from the internal population
# then reference for summary data specified in model needs to be provided
ext.ref <- dat[3501:4000, ] ## <-- as an example, assume ext.ref is different
                            ##    from dat0
fit3 <- gim(form, 'gaussian', dat0, model, nsample, ref = ext.ref)
```

Index