

Package ‘drgee’

November 8, 2016

Type Package

Title Doubly Robust Generalized Estimating Equations

Version 1.1.6

Date 2016-11-07

Author Johan Zetterqvist <johan.zetterqvist@ki.se>, Arvid Sjölander <arvid.sjolander@ki.se> with contributions from Alexander Ploner.

Maintainer Johan Zetterqvist <johan.zetterqvist@ki.se>

Description Fit restricted mean models for the conditional association between an exposure and an outcome, given covariates. Three methods are implemented: O-estimation, where a nuisance model for the association between the covariates and the outcome is used; E-estimation where a nuisance model for the association between the covariates and the exposure is used, and doubly robust (DR) estimation where both nuisance models are used. In DR-estimation, the estimates will be consistent when at least one of the nuisance models is correctly specified, not necessarily both.

Depends stats, nleqslv, survival, R (>= 3.0.0), Rcpp, data.table

LinkingTo Rcpp, RcppArmadillo

License GPL-2 | GPL-3

NeedsCompilation yes

Encoding UTF-8

Repository CRAN

Date/Publication 2016-11-08 11:10:20

R topics documented:

drgee-package	2
drgee	2
drgeeData	6
findRoots	9
gee	10
getScoreResidualsFromClogit	11
robVcov	12

drgee-package

*Doubly Robust Generalized Estimating Equations***Description**

The main function is `drgee`, which estimates a parameter β in a model defined as $g\{E(Y|A, L)\} - g\{E(Y|A = 0, L)\} = \beta^T\{A \cdot X(L)\}$. By supplying (nuisance) models for $g\{E(Y|A = 0, L)\}$ and $E(A|L)$, a consistent estimate of β is obtained when at least one of these models is correctly specified.

In the function `drgee`, three estimation methods are implemented: O-estimation, where a model for $g\{E(Y|A = 0, L)\}$ is used; E-estimation, where a model for $E(A|L)$ is used; and DR-estimation where models for both $g\{E(Y|A = 0, L)\}$ and $E(A|L)$ are used.

The function `gee` is an implementation of standard GEE with independent working correlation matrix.

For conditional methods with clustered data, cluster-specific intercepts are assumed in all models.

The function `drgeeData` is used for extraction and manipulation of data, and is called by `drgee` and `gee`.

The function `RobVcov` is used to calculate standard errors of the estimates, given a vector of the residuals from estimating equations, the Jacobian, and a cluster-identifying variable.

The function `findRoots` solves a system of non linear equations.

Author(s)

Johan Zetterqvist, Arvid Sjölander with contributions from Alexander Ploner.

drgee

*Doubly Robust Generalized Estimating Equations***Description**

`drgee` performs O-estimation, E-estimation or DR-estimation given symbolic representations of an outcome nuisance model and an exposure nuisance model. For clustered data the nuisance models may have cluster-specific intercepts.

Usage

```
drgee(outcome, exposure,
      oformula, eformula, iaformula = formula(~1),
      olink = c("identity", "log", "logit"),
      elink = c("identity", "log", "logit"),
      data, estimation.method = c("dr", "o", "e"),
      cond = FALSE, clusterid, rootFinder = findRoots,
      ...)
```

Arguments

outcome	The outcome as variable or as a character string naming a variable in the data argument. If missing, the outcome is assumed to be the response of oformula.
exposure	The exposure as variable or as a character string naming a variable in the data argument. If missing, the exposure is assumed to be the response of eformula.
offormula	An expression or formula for the outcome nuisance model.
eformula	An expression or formula for the exposure nuisance model.
iaformula	An expression or formula where the RHS should contain the variables that "interact" (i.e. are supposed to be multiplied with) with the exposure in the main model. "1" will always added. Default value is no interactions, i.e. iaformula = formula(~1).
olink	A character string naming the link function in the outcome nuisance model. Has to be "identity", "log" or "logit". Default is "identity".
elink	A character string naming the link function in the exposure nuisance model. Has to be "identity", "log" or "logit". Default is "identity".
data	A data frame or environment containing the variables used. If missing, variables are expected to be found in the calling environment of the calling environment.
estimation.method	A character string naming the desired estimation method. Choose "o" for O-estimation, "e" for E-estimation or "dr" for DR-estimation. Default is "dr".
cond	A logical value indicating whether the nuisance models should have cluster-specific intercepts. Requires a clusterid argument.
rootFinder	A function to solve a system of non-linear equations. Default is findRoots.
clusterid	A cluster-defining variable or a character string naming a cluster-defining variable in the data argument. If it is not found in the data argument, it will be searched for in the calling frame. If missing, each observation will be considered to be a separate cluster. This argument is required when cond = TRUE.
...	Further arguments to be passed to the function rootFinder.

Details

drgee estimates the parameter β in a main model $g\{E(Y|A, L)\} - g\{E(Y|A = 0, L)\} = \beta^T \{A \cdot X(L)\}$, where Y is the outcome of interest, A is the exposure of interest, and L is a vector of covariates that we wish to adjust for. $X(L)$ is a vector valued function of L . Note that $A \cdot X(L)$ should be interpreted as a columnwise multiplication and that $X(L)$ will always contain a column of 1's. Given a specification of an outcome nuisance model $g\{E(Y|A = 0, L)\} = \gamma^T V(L)$ (where $V(L)$ is a function of L) O-estimation is performed. Alternatively, leaving $g\{E(Y|A = 0, L)\}$ unspecified and using an exposure nuisance model $h\{E(A|L)\} = \alpha^T Z(L)$ (where h is a link function and $Z(L)$ is a function of L) E-estimation is performed. When g is logit, the exposure nuisance model is required be of the form $\text{logit}\{E(A|Y = 0, L)\} = \alpha^T Z(L)$. In this case the exposure needs to binary.

Given both an outcome and an exposure nuisance model, DR-estimation can be performed. DR-estimation gives a consistent estimate of the parameter β when either the outcome nuisance model or the exposure nuisance model is correctly specified, not necessarily both.

Usage is best explained through an example. Suppose that we are interested in the parameter vector (β_0, β_1) in a main model $\text{logit}\{E(Y|A, L_1, L_2)\} - \text{logit}\{E(Y|A = 0, L_1, L_2)\} = \beta_0 A + \beta_1 A \cdot L_1$ where L_1 and L_2 are the covariates that we wish to adjust for. To adjust for L_1 and L_2 , we can use an outcome nuisance model $E(Y|A = 0, L_1, L_2; \gamma_0, \gamma_1) = \gamma_0 + \gamma_1 L_1$ or an exposure nuisance model $\text{logit}\{E(A|Y = 0, L_1, L_2)\} = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$ to calculate estimates of β_0 and β_1 in the main model. We specify the outcome nuisance model as `oformula=Y~L_1` and `olink = "logit"`. The exposure nuisance model is specified as `eformula = A~L_1+L_2` and `elink = "logit"`. Since the outcome Y and the exposure A are identified as the LHS of `oformula` and `eformula` respectively and since the outcome link is specified in the `olink` argument, the only thing left to specify for the main model is the (multiplicative) interactions $A \cdot X(L) = A \cdot (1, L_1)^T$. This is done by specifying `X(L)` as `iaformula = ~L_1`, since 1 is always included in $X(L)$. We can then perform O-estimation, E-estimation or DR-estimation by setting `estimation.method` to "o", "e" or "dr" respectively. O-estimation uses only the outcome nuisance model, and E-estimation uses only the exposure nuisance model. DR-estimation uses both nuisance models, and gives a consistent estimate of (β_0, β_1) if either nuisance model is correct, not necessarily both.

When `estimation.method = "o"`, the RHS of `eformula` will be ignored. The `eformula` argument can also be replaced by an exposure argument specifying what the exposure of interest is.

When `estimation.method = "e"`, the RHS of `oformula` will be ignored. The `oformula` argument can also be replaced by an outcome argument specifying what the outcome of interest is.

When `cond = TRUE` the nuisance models will be assumed to have cluster-specific intercept. These intercepts will not be estimated.

When E-estimation or DR-estimation is chosen with `olink = "logit"`, the exposure link will be changed to "logit". Note that this choice of outcome link does not work for DR-estimation when `cond = TRUE`.

Robust variance for the estimated parameter is calculated using the function `robvcov`. A cluster robust variance is calculated when a character string naming a cluster variable is supplied in the `clusterid` argument.

For E-estimation when `cond = FALSE` and g is the identity or log link, see Robins et al. (1992).

For DR-estimation when `cond = TRUE` and g is the identity or log link, see Robins (1999). For DR-estimation when g is the logit link, see Tchetgen et al. (2010).

O-estimation can also be performed using the `gee` function.

Value

`drgee` returns an object of class `drgee` containing:

<code>coefficients</code>	Estimates of the parameters in the main model.
<code>vcov</code>	Robust variance for all main model parameters.
<code>coefficients.all</code>	Estimates of all estimated parameters.
<code>vcov.all</code>	Robust variance of the all parameter estimates.
<code>optim.object</code>	An estimation object returned from the function specified in the <code>rootFinder</code> , if this function is called for the estimation of the main model parameters.
<code>optim.object.o</code>	An estimation object returned from the function specified in the <code>rootFinder</code> , if this function is called for the estimation of the outcome nuisance parameters.

`optim.object.e` An estimation object returned from the function specified in the `rootFinder`, if this function is called for the estimation of the outcome nuisance parameters.

`call` The matched call.

`estimation.method` The value of the input argument `estimation.method`.

The class methods `coef` and `vcov` can be used to extract the estimated parameters and their covariance matrix from a `drgee` object. `summary.drgee` produces a summary of the calculations.

Author(s)

Johan Zetterqvist, Arvid Sjölander

References

- Orsini N., Belocco R., Sjölander A. (2013), Doubly Robust Estimation in Generalized Linear Models, *Stata Journal*, **13**, 1, pp. 185–205
- Robins J.M., Mark S.D., Newey W.K. (1992), Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders, *Biometrics*, **48**, pp. 479–495
- Robins JM (1999), Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models, *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, pp. 6–10
- Tchetgen E.J.T., Robins J.M., Rotnitzky A. (2010), On Doubly Robust Estimation in a Semiparametric Odds Ratio Model, *Biometrika*, **97**, 1, 171–180

See Also

[gee](#) for O-estimation, [findRoots](#) for nonlinear equation solving and [robVcov](#) for estimation of variance.

Examples

```
## DR-estimation when
## the main model is
##  $E(Y|A, L1, L2) - E(Y|A=0, L1, L2) = \beta_0 * A + \beta_1 * A * L1$ 
## and the outcome nuisance model is
##  $E(Y|A=0, L1, L2) = \gamma_0 + \gamma_1 * L1 + \gamma_2 * L2$ 
## and the exposure nuisance model is
##  $E(A|Y=0, L1, L2) = \text{expit}(\alpha_0 + \alpha_1 * L1 + \alpha_2 * L2)$ 

library(drgee)

expit<-function(x) exp(x)/(1+exp(x))

n<-5000

## nuisance
l1<-rnorm(n, mean = 0, sd = 1)
```

```

l2<-rnorm(n, mean = 0, sd = 1)

beta0<-1.5
beta1<-1
gamma0<--1
gamma1<--2
gamma2<-2
alpha0<-1
alpha1<-5
alpha2<-3

## Exposure generated from the exposure nuisance model
a<-rbinom(n,1,expit(alpha0 + alpha1*l1 + alpha2*l2))
## Outcome generated from the main model and the
## outcome nuisance model
y<-rnorm(n,
mean = beta0 * a + beta1 * a * l1 + gamma0 + gamma1 * l1 + gamma2 * l2,
sd = 1)

simdata<-data.frame(y,a,l1,l2)

## outcome nuisance model misspecified and
## exposure nuisance model correctly specified

## DR-estimation
dr.est <- drgee(oformula = formula(y~l1),
eformula = formula(a~l1+l2),
iaformula = formula(~l1),
olink = "identity", elink = "logit",
data = simdata, estimation.method = "dr")
summary(dr.est)

## O-estimation
o.est <- drgee(exposure = "a", oformula = formula(y~l1),
iaformula = formula(~l1), olink = "identity",
data = simdata, estimation.method = "o")
summary(o.est)

## E-estimation
e.est <- drgee(outcome = "y", eformula = formula(a~l1+l2),
iaformula = formula(~l1), elink="logit",
data = simdata, estimation.method = "e")
summary(e.est)

```

Description

Given a main model, an outcome nuisance model and an exposure nuisance model `drgeeData` extracts the model variables and matrices from a `data.frame` or an environment object. It also performs some data cleaning and error checking.

Usage

```
drgeeData(outcome, exposure,
          oformula, eformula, iaformula = formula(~1),
          olink = c("identity", "log", "logit"),
          elink = c("identity", "log", "logit"),
          data, estimation.method = c("dr", "o", "e"),
          cond = FALSE, clusterid)
```

Arguments

<code>outcome</code>	The outcome as a variable or as a character string naming a variable in the data argument. If it is not found in the data argument, it will be searched for in the calling frame. If missing, the outcome is assumed to be the response of <code>oformula</code> .
<code>exposure</code>	The exposure as a variable or as a character string naming a variable in the data argument. If it is not found in the data argument, it will be searched for in the calling frame. If missing, the outcome is assumed to be the response of <code>eformula</code> .
<code>oformula</code>	An expression or formula for the outcome nuisance model. The outcome is identified as the response in this formula.
<code>eformula</code>	An expression or formula for the exposure nuisance model. The exposure is identified as the response in this formula.
<code>iaformula</code>	An expression or formula where the RHS should contain the variables that "interact" (i.e. are supposed to be multiplied with) with the exposure in the main model to create the terms associated with the parameters of interest. "1" will always added. Default value is no interactions, i.e. <code>formula(~1)</code> .
<code>olink</code>	A character string naming the link function in the outcome nuisance model. Have to be "identity", "log" or "logit". Default is "identity".
<code>elink</code>	A character string naming the link function in the exposure nuisance model. Have to be "identity", "log" or "logit". Default is "identity". When <code>olink="logit"</code> this is replaced by "logit".
<code>data</code>	A data frame or environment containing the variables in <code>iaformula</code> , <code>oformula</code> and <code>eformula</code> . If missing, variables are expected to be found in the calling frame.
<code>estimation.method</code>	A character string naming the desired estimation method. Choose "o" for O-estimation, "e" for E-estimation or "dr" for DR-estimation. Default is "dr".
<code>cond</code>	A logical value indicating whether the nuisance models should have cluster-specific intercepts. If <code>cond=TRUE</code> the design matrices for the nuisance models do not have an intercept. Requires a <code>clusterid</code> argument.

`clusterid` A cluster-defining variable or a character string naming a cluster-defining variable in the `data` argument. If it is not found in the `data` argument, it will be searched for in the calling frame. If missing, each observation will be considered to be a separate cluster. This argument is required when `cond = TRUE`.

Details

`drgeeData` is called by `drgee` and `gee` to extract data from a `data.frame` or environment object. The data can then be used to for O-estimation, E-estimation or DR-estimation. `drgeeData` uses `model.frame` and `model.matrix` to remove incomplete observations and to convert factors to dummy variables. It also performs check the supplied data for errors or inconsistencies.

The class method `summary.drgeeData` produces strings for the formulas with terms referring to the columns in the produced design matrices.

Value

`drgee.data` returns an object of class `drgeeData` containing

<code>orig.order</code>	The original order of the observations.
<code>y</code>	The outcome matrix.
<code>a</code>	The exposure matrix.
<code>x</code>	The matrix of of interactions defined in <code>iaformula</code> . This matrix will always contain a column with 1's.
<code>ax</code>	The matrix of elementwise product(s) of <code>a</code> and each column in <code>x</code> .
<code>v</code>	The matrix of terms in the outcome nuisance model.
<code>z</code>	The matrix of terms in the exposure nuisance model.
<code>yx</code>	The matrix of elementwise product(s) of <code>y</code> and each column in <code>x</code> .
<code>id</code>	A factor defining clusters. For independent observations, the number of levels equals the number of complete observations.
<code>clustname</code>	A string for the name of the cluster defining variable.
<code>y.names</code>	A string for the name of the outcome.
<code>a.names</code>	A string for the name of the exposure.
<code>x.names</code>	A string vector for the variable names in <code>x</code> .
<code>ax.names</code>	A string vector for the variable names in <code>ax</code> .
<code>v.names</code>	A string vector for the variable names in <code>v</code> .
<code>z.names</code>	A string vector for the variable names in <code>z</code> .
<code>yx.names</code>	A string vector for the variable names in <code>yx</code> .
<code>olink</code>	A character string naming the link function in the outcome nuisance model.
<code>elink</code>	A character string naming the link function in the exposure nuisance model.
<code>cond</code>	A logical value indicating whether cluster-specific intercepts should be assumed. If <code>TRUE</code> , there is no column for the intercept in <code>v</code> and <code>z</code> . Outcome concordant will also be removed.
<code>oterms</code>	The terms object corresponding to the outcome nuisance model.
<code>eterms</code>	The terms object corresponding to the exposure nuisance model.

Author(s)

Johan Zetterqvist, Arvid Sjölander

See Also

[drgee](#), [model.frame](#) and [model.matrix](#).

 findRoots

Non Linear Equation System Solving

Description

A wrapper around `nleqslv` from the **nleqslv** package to solve a non linear system of equations.

Usage

```
findRoots(beta.init, eq.func, d.eq.func = NULL, arg.list, ...)
```

Arguments

<code>beta.init</code>	An initial guess for the zero.
<code>eq.func</code>	A function of two variables for which the zero are sought. Its first argument <code>beta</code> should be a vector over which the zeros are sought and the second argument <code>arg.list</code> a list of additional arguments.
<code>d.eq.func</code>	A function to return the Jacobian of <code>eq.func</code> taking the same arguments as <code>eq.func</code> . Supplying this function can speed up calculations. Default is <code>NULL</code> .
<code>arg.list</code>	The second argument to <code>eq.func</code> and <code>d.eq.func</code>
<code>...</code>	A list of additional arguments to be passed to <code>nleqslv</code>

Details

`findRoots` calculates zeros fo the function `eq.func` and is the default equation solving function in `drgee`. It is supplied as a separate function in order to allow users to use other equation solvers by writing their own wrapper with the same interface as `findRoots`.

Value

The value is a list containing the following arguments:

<code>roots</code>	The zero(s) of the function <code>eq.func</code> .
<code>optim.object</code>	The optimization object returned from <code>nleqslv</code> .

Author(s)

Johan Zetterqvist, Arvid Sjölander

See Also

[nlqslv](#) in package **nlqslv**.

 gee

Generalized Estimating Equations

Description

gee performs estimation of parameters in a restricted mean model using standard GEEs with independent working correlation matrix. For clustered data, cluster-robust standard errors are calculated. When cond=TRUE, cluster-specific intercepts are assumed.

Usage

```
gee(formula, link = c("identity", "log", "logit"), data, cond = FALSE,
    clusterid, rootFinder = findRoots, ...)
```

Arguments

formula	An expression or formula representing the expected outcome conditional on co-variates.
link	A character string naming the link function to use. Has to be "identity", "log" or "logit". Default is "identity".
data	A data frame or environment containing the variables appearing in formula. If missing, the variables are expected to be found in the environment of the formula argument.
cond	A logical value indicating whether cluster-specific intercepts should be assumed. Requires a clusterid argument.
clusterid	A cluster-defining variable or a character string naming a cluster-defining variable in the data argument. If it is not found in the data argument, it will be searched for in the calling frame. If missing, each observation will be considered to be a separate cluster. This argument is required when cond = TRUE.
rootFinder	A function to solve a system of non linear equations. Default is findRoots.
...	Further arguments to be passed to the function rootFinder.

Details

Estimates parameters in a regression model, defined by formula. When cond=FALSE, the estimated coefficients are identical to those obtained with glm, but since no distributional assumptions are made, a robust variance is calculated. When cond=TRUE and link is "identity" or "log", the coefficients will be calculated using conditional estimating equations as described in Goetgeluk and Vansteelandt (2008). When cond=TRUE and link="logit", the coefficients will be calculated by conditional logistic regression (with robust standard errors).

Value

gee return an object of class gee containing:

coefficients	Estimates of the parameters.
vcov	Robust variance of the estimates.
call	The matched call.
y	The outcome vector.
x	The design matrix. For conditional methods there is no column corresponding to the intercept.
optim.object	An estimation object returned from the function specified in the rootFinder, if this function is called.
res	The residuals from the estimating equations.
d.res	The derivative of the residuals from the estimating equations.

The class methods `coef` and `vcov` can be used to extract the estimated parameters and their covariance matrix from a gee object. `summary.drgee` produces a summary of the calculations.

Author(s)

Johan Zetterqvist, Arvid Sjölander

References

Goetgeluk S., & Vansteelandt S. (2008), Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, **64**(3), pp. 772–780.

See Also

`glm`

getScoreResidualsFromClogit

Get the (conditional) score residuals from conditional logistic regression

Description

A function to obtain the conditional score residuals from conditional logistic regression as well as the sum of derivatives of these residuals with respect to the parameters evaluated at the estimated parameters.

Usage

```
getScoreResidualsFromClogit(fit, coefs, y, x, id)
```

Arguments

fit	A fitted object. Could be any object that has a class.function coefficients, but is intended to be used with clogit in the survival package.
coefs	An estimate of the parameters. If the argument fit is supplied, this will be replaced by coefficients(fit).
y	The observed outcomes.
x	The design matrix.
id	The cluster identification variable.

Details

getScoreResidualsFromClogit calculates the residuals from the conditional score equations used in conditional logistic regression as well as the sums of their derivatives with respect to the parameters.

This is useful if one wants to use a calculate the sandwich estimate of a variance where the uncertainty in the estimation of a conditional odds ratio needs to be taken into account.

Value

getScoreResidualsFromClogit return a list containing:

U	An $n \times p$ matrix, where n is the number of observations and p is the length of the estimated parameter.
dU.sum	An $p \times p$ matrix containing the sums of the derivatives. The rows correspond to the estimating equations and the columns correspond to the partial derivatives.

Author(s)

Johan Zetterqvist, Arvid Sjölander

See Also

[robVcov](#).

 robVcov

Robust Variance Calculation

Description

robVcov calculates the asymptotic variance for Z-estimators.

Usage

```
robVcov(u, d.u, id=NULL)
```

Arguments

u	A $n \times q$ matrix of the estimating equations evaluated at the estimated model parameters, where n is the number of observations and q is the number of estimating equations.
d.u	The mean of the jacobian of u evaluated at the solution to the estimating equations, with rows corresponding to the estimating equations and columns corresponding to the model parameters. The number of model parameters is assumed to equal the number of estimating equations such that $d.u$ is a $q \times q$ square matrix.
id	A factor with levels corresponding the clusters in the data. Default is NULL in which case all observations are considered to be independent.

Details

For robust variance estimation, see van der Vaart (2000).

For clustered data, the rows in u are added clusterwise resulting in a cluster robust estimate of the variance.

Value

The estimated covariance matrix.

Author(s)

Johan Zetterqvist, Arvid Sjölander

References

van der Vaart, A.W. (2000), *Asymptotic Statistics*, Cambridge University Press, pp. 52–53.

Index

*Topic **cluster**

drgee, 2
drgee-package, 2
drgeeData, 6
gee, 10
robVcov, 12

*Topic **models**

drgee, 2
drgee-package, 2
drgeeData, 6
gee, 10

*Topic **optimize**

findRoots, 9
getScoreResidualsFromClogit, 11

*Topic **regression**

drgee, 2
drgee-package, 2
drgeeData, 6
gee, 10

coef.drgee (drgee), 2

coef.gee (gee), 10

drgee, 2, 9

drgee-package, 2

drgeeData, 6

findRoots, 5, 9

gee, 5, 9, 10

getScoreResidualsFromClogit, 11

model.frame, 9

model.matrix, 9

nleqslv, 10

print.drgee (drgee), 2

print.gee (gee), 10

print.summary.drgee (drgee), 2

print.summary.drgeeData (drgeeData), 6

print.summary.gee (gee), 10

robVcov, 5, 12, 12

summary.drgee (drgee), 2

summary.drgeeData (drgeeData), 6

summary.gee (gee), 10

vcov.drgee (drgee), 2

vcov.gee (gee), 10