1 Implementation

1.1 G-Estimation For End-of-Study Outcomes

The main working functions of **gesttools** are indexed "gest" and "gestmult", which perform g-estimation for end-of-study and time-varying (multiple) outcome studies respectively. There are 2 functions to perform g-estimation for end-of-study outcomes. The function `gest` performs g-estimation for end-of-study outcomes for a binary or continuous exposure, and `gest.cat` an equivalent function for categorical exposures of 3 or more categories. The functions `gestmult` and `gestmult.cat` are equivalent functions for time-varying outcome data.

```
gest(data,Yn,An,Ybin,Abin,Lny,Lnp,z=NULL,type=NA,timevarying,Cn=NA,
LnC=NA,...)
gest.cat(data,Yn,An,Ybin,Lny,Lnp,z=NULL,type=NA,timevarying,Cn=NA,
LnC=NA,...)
gestmult(data,Yn,An,Ybin,Abin,Lny,Lnp,z=NULL,type=NA,timevarying,
Cn=NA,LnC=NA,cutoff=NA,...)
gestmult.cat(data,Yn,An,Ybin,Lny,Lnp,z=NULL,type=NA,timevarying,
Cn=NA,LnC=NA,cutoff=NA,...)
```

- `data`: the data to be analysed. These data must be set up in a specific format, described in the details section below.
- `Yn`, `An`: Name of the outcome and exposure variable written in quotations (""). When the outcome or exposure is binary it must be written as a numeric variable holding values 0 or 1, with 0 indicating unexposed and 1 indicating exposed.
- `Ybin`, `Abin`: True or False indicator of whether Y is continuous or binary/count data and if A is binary or continuous.
- `Lny`: Vector of names each name to be given within quotes of the covariates in the adjusted outcome model.
- `Lnp`: Vector of names each name to be given within quotes of the covariates in the propensity score model.
- `z`: vector specifying the form of z_t in the SNMM. Allows manual control of the SNMM type to fit (see examples).
- `type`: Value from 1 to 4 determining which SNMM specification (*i.e.* type) to fit. When fitting SNMM types 2 or 4, effect modification is allowed between the exposure and the covariate in the first element of Lny.
- `timevarying`: True or False indicator of whether the model allows a time-varying effect of the exposure.
- `Cn`: Name of the censoring indicator in quotations (if applicable). Must be written as a numeric 0,1 variable with 1 indicating censored.

- `Lnc`: Vector of names (each name to be given within quotes) of the covariates in the censoring score model.
- `cutoff`: A number c from 1 to T, which which will stop the algorithm once H_c is calculated (and ψ is estimated) for an end-of-study outcome, or once H_c $H_{s(s-c)}$ is calculated for multiple outcomes .

Each function will output a vector of causal effects labeled as they are in the outcome model fitted by `geem()`. If the effect is time varying, the causal effects will be labelled according to the exposure time they relate to. For an end-of-study outcome these are labeled t, t = $1, \ldots, T$, and for a multiple outcome study these are labeled s – t, t = $(1, \ldots, T)$.

Data Setup and Details

Data

The data must be in long format, that is individual data at each time point are to be stored on separate rows. The data must be ordered by individual identifier and within individual by ascending time, and these variables must be labelled "id" and "time", respectively. The time points must be labelled 1 to T (i.e. they must not start at 0).

Each row with time=t should contain the concurrent values for exposure and covariates (i.e. A_t and L_t), and the values for the outcome and censoring indicator at the next time period (i.e. Y_{t+1} and C_{t+1}). Note that when t represents some period of time, rather than a specific time point, A_t denotes the exposure measured at the start of time period t, and Y_{t+1} is the outcome measured at the end of time period t. For an end-of-study outcome, the outcome Y_{t+1}, should be repeated on each row.

Crucially there must exist a row for each individual at each time period t, for t = $1, \ldots, T$, i.e. data must have a rectangular form. Even when an individual is censored before the end of follow-up, the user must provide additional records, with missing values entered for all variables other than id, time and the censoring indicator, with the latter taking value 1 in correspondence to the last record with data and all the subsequent ones. These additional records are needed to calculate counterfactuals, and as such there must be an equal number of data rows for all T.

Details

Strictly speaking the user should include the same covariates in the propensity and adjusted outcome models, that is `Lnp`≡`Lny`. However in practice this can sometime lead to issues in fitting the adjusted outcome model, either due to collinearity, or in the case of sparse binary outcome data, insufficient data to estimate the parameters. In this case we recommend to remove (some of the) covariates from the outcome model (`Lny`) but to keep them in the propensity score model. Causal effect estimates will remain unbiased provided that the propensity score model is correctly specified.

If the outcome is time-varying the algorithms become increasingly slow as T becomes large. For example, when T = 3, there are 3 + 2 + 1 = 6 counterfactuals H_{st} to estimate, but when T = 10 there are 10 + 9 + ... + 1 = 55 to estimate. The `cutoff`

option states the value c (with a choice from 1 to T) that controls the number of times step 5 in the algorithm described in section 4 is repeated, thus permitting g-estimation when T is large without unreasonably long computation time being needed. More so, it allows the user to specify that the exposure has an effect on the outcome only up to c time periods after.

1.2 Choice of Structural Nested Mean Models

The package allows users to specify the form of the SNMM they wish to fit either through an input argument `type`, or manual specification of z_{st} through the arguments `z` and `timevarying`. Note that the argument `type` will override the arguments `z` and `timevarying`.

- Type 1: This requires setting `type=1`, or `z=1` and `timevarying=FALSE`.
- Type 2: This requires setting `type=2` or `z=`$c(1,"L^*")$ and `timevarying=FALSE`, where L^* is the first covariate in the list defined by `Ln`. Note that this variable must be either continuous, or binary, in the latter case held as a numeric $0, 1$ variable. (Effect modification by ordinal variables are not supported.)
- Type 3: This requires setting `type=3` or `z=1` and `timevarying=TRUE`.
- Type 4: This requires setting `type=4` or `z=`$c(1,"L^*")$ and `timevarying=TRUE`.

By specifying `z`, other SNNM types, such as effect modification by multiple covariates is possible.

1.3 Bootstrap Function

Standard errors for the causal effect estimates are obtained by bootstrapping the data using the function "gest.boot".

```
gest.boot(data,gestfunc,Yn,An,Ybin,Abin,Lny,Lnp,
z,type=NA,timevarying=FALSE,Cn=NA,LnC=NA,cutoff=NA,bn,alpha,...)
```

- `func`: Name of the g-estimation function to use, for example `gest`
- `data`, `Yn`, `An`, `Abin`, `Lny`, `Lnp`, `z`, `type`, `timevarying`, `Cn`, `LnC`, `cutoff`: Same arguments as in g-estimation functions
- `bn`: The number of bootstrapped datasets to be generated
- `alpha`: The desired α level

The function will output a two-sided $1 - \alpha\%$ confidence interval for every causal estimate comprising ψ. The function will assume enough bootstrap samples are used such that an asymptotic normal confidence interval is appropriate. Both standard normal and Bonferroni corrected intervals for multiple comparisons are provided. Intervals for each causal effect are labelled in the same way as in the g-estimation functions.

1.4 Example 1: gest and gestmult

A simulated dataset with a continuous end-of-study outcome and binary exposure was constructed following the structure of figure 1. The data are simulated with 10,000 individuals and T = 3 as follows

- Unmeasured covariate: $U \sim N(0,1)$
- Covariate $L_t \sim N(1 + L_{t-1} + \alpha_t A_{t-1} + U)$ $t = 1,2,3$
- Exposure: $A_t \sim Bin(1, expit(1 + 0.1 * L_t + 0.1 * A_{t-1}))$ $t = 1,2,3.$
- End-of-study outcome: $Y_4 \sim N(1 + \gamma_1 A_1 + \gamma_2 A_2 + \gamma_3 A_3 + L_1 + L_2 + L_3 + U, 1)$

where $A_{t-1} = 0$ when $t = 1$ and $expit(x) = exp(x)(1 + exp(x))^{-1}$ is the inverse logit function. By setting $(\alpha_2, \alpha_3) = (1/3, 1/2)$ and $(\gamma_1, \gamma_2, \gamma_3) = (1/3, 1/2, 1)$, the true causal effect for the four SNNM specifications are

- Type 1: $\psi = 1$
- Type 2: $\psi = (1, 0)$
- Type 3: $\psi = (1, 1, 1)$
- Type 4:

$$\psi = \begin{pmatrix} 0,1 \\ 0,1 \\ 0,1 \end{pmatrix}.$$

This data are generated in R using the code found in the Appendix and are labeled "dl". A snippet of the data can be seen below, with the dataset available in "SimulatedExamples.R".

```
    id         Y          U time A          L
1.1  1  7.236854 -0.5767339    1 1  0.6834219
1.2  1  7.236854 -0.5767339    2 1  1.6403011
1.3  1  7.236854 -0.5767339    3 1  2.3059240
2.1  2 -5.471200 -1.3704997    1 1 -2.0019739
2.2  2 -5.471200 -1.3704997    2 0 -2.9648525
2.3  2 -5.471200 -1.3704997    3 1 -2.3242440
```

Here we demonstrate the `gest` function for SNMM types 1, 2 and 3, as well as the `gest.boot` function. For this data, the propensity score model includes covariates U and L, that is `Lnp=`c("L","U") and the outcome model includes L, thus `Lny=`c("L"). G-estimation for SNMM type 1 is therefore

```
>data<-dl
>#SNMM type 1
>gest(data=data,Yn="Y",An="A",Ybin=FALSE,Abin=TRUE,
>Lny=c("L","U"),Lnp=c("L"),z=c(1),type=1,
>timevarying=FALSE,Cn=NA,LnC=NA)
```

```
$psi
       A
1.027724
```

We fit SNMM type 2 as

```
>#SNMM type 1
>gest(data=data,Yn="Y",An="A",Ybin=FALSE,Abin=TRUE,
>Lny=c("L","U"),Lnp=c("L"),z=c(1,"L"),type=2,
>timevarying=FALSE,Cn=NA,LnC=NA)
\end{verbatimF}
\begin{verbatim}
$psi
        A       A:L
0.9677974 0.0277276
```

Here `A` is the overall causal effect of exposure on outcome when $L = 0$, and `A:L` is the effect modification due to L, that is the change in causal effect of exposure for each unit increase in L. Now we fit SNMM type 3.

```
>#SNMM type 3
>gest(data=data,Yn="Y",An="A",Ybin=FALSE,Abin=TRUE,
>Lny=c("L","U"),Lnp=c("L"),z=c(1),type=3,
>timevarying=TRUE,Cn=NA,LnC=NA)

$psi
    t=1.A     t=2.A     t=3.A
1.0024918 0.9826677 1.0182055
```

Note that the causal effects are now labeled by time, with the effect labeled `t=1.A` is the effect of exposure at $t = 1$ on outcome at time $t = 4$. We now demonstrate the bootstrap function `gest.boot` for SNMM type 3, using 100 bootstraps and alpha set at 0.05

```
>#SNMM type 3 bootstrap
>gest.boot(data=data,gestfunc=gest,Yn="Y",An="A",Ybin=FALSE,Abin=TRUE,
>Lny=c("L","U"),Lnp=c("L"),z=c(1),type=3,
>timevarying=TRUE,Cn=NA,LnC=NA,cutoff=NA,bn=100,alpha=0.05)

$original
    t=1.A     t=2.A     t=3.A
1.0024918 0.9826677 1.0182055

$conf
             low      upp
t=1.A 0.8990022 1.105981
t=2.A 0.9087869 1.056549
t=3.A 0.9330595 1.103352
```

```
$conf.Bonferroni
           lowb     uppb
t=1.A 0.8760854 1.128898
t=2.A 0.8924266 1.072909
t=3.A 0.9142047 1.122206

$mean
    t=1.A     t=2.A     t=3.A
1.0013768 0.9808164 1.0247623

$s.e
     t=1.A      t=2.A      t=3.A
0.05280178 0.03769502 0.04344266
```

The causal effects for the original (non-bootstrapped) data is given by `$original`, with `$mean` and `$s.e` giving the average causal effects of the bootstrapped samples, and the between bootstrap standard error respectively. Finally `$conf` and `$conf.Bonferroni` give the standard and Bonferroni corrected bootstrap confidence intervals for each causal effect.

Additionally we can demonstrate the `gestmult` function on this data, by supposing that our end-of-study outcome, repeated on each row, is actually a time-varying outcome. We will test SNMM type 3.

```
>#SNMM type 3 gestmult
>gestmult(data=data,Yn="Y",An="A",Ybin=FALSE,Abin=TRUE,
>Lny=c("L","U"),Lnp=c("L"),z=c(1),type=3,
>timevarying=TRUE,Cn=NA,LnC=NA,cutoff=NA)

$psi
   s-1.A    s-2.A    s-3.A
1.046011 1.050415 1.000294
```

Although the above causal effects are not valid as Y is not in fact a repeated outcome (note that the causal effects are correct due to the way the data are simulated), this provides a demonstration of the output of `gestmult`. The effect labelled `s-1.A` is the casual effect of exposure at time $s-1$ on Y_s, that is the effect of exposure on the subsequent outcome.

1.5 Example 2:gest.cat and gestmult.cat

As a final example we demonstrate `gest.cat` and `gestmult.cat` which performs g-estimation in the case of a categorical exposure variable. As in example 1 we generate a dataset with n = 10000 and T = 3, where we set A as a three category variable with levels "a","b" and "c", with category "a" the reference category. We define a function

$\zeta(A)$ where

$$\zeta(A) = \begin{cases} 0: \text{ if } A = "a" \\ 1: \text{ if } A = "b" \\ 2: \text{ if } A = "c" \end{cases}$$

which defines the coefficient of each category of A at a given time with respectively the next value of A and L, as well as with the final outcome Y. The exposure A is now sampled from a multinomial distribution with probabilities

- $P(A_t = "a") = \frac{2}{5} * expit(1 + 0.1 * L_t + \zeta(A_{t-1}))$
- $P(A_t = "b") = \frac{1}{5} * expit(1 + 0.1 * L_t + \zeta(A_{t-1}))$
- $P(A_t = "c") = \frac{2}{5} * expit(1 + 0.1 * L_t + \zeta(A_{t-1}))$

for $t = 1, 2, 3$ where $\zeta(A_0) = 0$, and U, L and the end-of-study outcome Y_4 are sampled in a similar way as in example 1

- Unmeasured covariate: $U \sim N(0, 1)$
- Covariate $L_t \sim N(1 + L_{t-1} + \zeta(A_{t-1}) + U)$ $t = 1, 2, 3$
- End-of-study outcome: $Y_4 \sim N(1+\zeta(A_1)+\zeta(A_2)+\zeta(A_3)+L_1+L_2+L_3+U, 1)$.

In the case the true causal effects for SNMM types 1 and 3 are

- Type 1: $\psi_b = 2$, $\psi_c = 4$
- Type 3: $\psi_b = (3, 2, 1)$, $\psi_c = (6, 4, 2)$.

We will therefore demonstrate SNMM types 1 and 3, as well as `gestmult.cat`. G-estimation for SNMM type 1 sets

```
>#SNMM type 1
>gest.cat(data=data,Yn="Y",An="A",Ybin=FALSE,
>Lny=c("L","U"),Lnp=("L"),z=c(1),type=1,
>timevarying=FALSE,Cn=NA,LnC=NA)

$psi
      Ab       Ac
2.050298 4.014344
```

The effect labelled `Ab` is the causal effect of exposure to category "b" (versus "a") on the outcome.

```
>#SNMM type 3
>gest.cat(data=data,Yn="Y",An="A",Ybin=FALSE,
>Lny=c("L","U"),Lnp=("L"),z=c(1),type=3,
>timevarying=TRUE,Cn=NA,LnC=NA)

$psi
   t=1.Ab    t=1.Ac    t=2.Ab    t=2.Ac    t=3.Ab    t=3.Ac
3.0166774 5.9124846 2.0104973 3.9934634 0.9735185 2.0545611
```

The effect labelled `t=1.Ab` is the causal effect of exposure to category "b" (versus "a") at time $t = 1$ on the outcome. We can also run `gest.boot`

```
gest.boot(data=data,gestfunc=gest.cat,Yn="Y",An="A",Ybin=FALSE,
Lny=c("L","U"),Lnp=c("L"),z=c(1),type=3,
timevarying=TRUE,Cn=NA,LnC=NA,cutoff=NA,bn=100,alpha=0.05)
```

```
$original
   t=1.Ab    t=1.Ac    t=2.Ab    t=2.Ac    t=3.Ab    t=3.Ac
3.0166774 5.9124846 2.0104973 3.9934634 0.9735185 2.0545611

$conf
             low      upp
t=1.Ab 2.8694088 3.163946
t=1.Ac 5.8068199 6.018149
t=2.Ab 1.9282046 2.092790
t=2.Ac 3.9255048 4.061422
t=3.Ab 0.8666756 1.080361
t=3.Ac 1.9689491 2.140173

$conf.Bonferroni
            lowb     uppb
t=1.Ab 2.8184429 3.214912
t=1.Ac 5.7702520 6.054717
t=2.Ab 1.8997252 2.121269
t=2.Ac 3.9019861 4.084941
t=3.Ab 0.8297001 1.117337
t=3.Ac 1.9393210 2.169801

$mean
  t=1.Ab   t=1.Ac   t=2.Ab   t=2.Ac   t=3.Ab   t=3.Ac
3.014307 5.912895 2.007832 3.993829 0.970815 2.050453

$s.e
    t=1.Ab     t=1.Ac     t=2.Ab     t=2.Ac     t=3.Ab     t=3.Ac
0.07513842 0.05391158 0.04198684 0.03467337 0.05451266 0.04368037

Warning message:
In geem(terms(lmH1), family = family, id = dtcom$id, data = dtcom,  :
  Did not converge
```

We note the convergence warning above, which indicates that in one of the bootstrapped datasets, the outcome model could not be fit due to collinearity. Any bootstrap result which does not converge is removed from the list used to generate confidence intervals. If many bootstrapped datasets do not converge, this may indicate a sparse dataset with little outcome data, or that removing covariates from the outcome model may be necessary. Finally, we demonstrate `gestmult.cat` for SNMM type 3

```
>gestmult.cat(data=data,Yn="Y",An="A",Ybin=FALSE,
>Lny=c("L","U"),Lnp=("L"),z=c(1),type=3,
>timevarying=TRUE,Cn=NA,LnC=NA,cutoff=NA)

$psi
  s-1.Ab   s-1.Ac   s-2.Ab   s-2.Ac   s-3.Ab   s-3.Ac
2.080766 4.118024 2.519934 4.892746 3.008773 5.840919
```

where `s-1A.b` is the effect of exposure at category "b" (compared to "a") on the subsequent outcome.