Package 'DiDforBigData'

July 21, 2025

Title A Big Data Implementation of Difference-in-Differences Estimation with Staggered Treatment

Version 1.0

Description Provides a big-data-friendly and memory-efficient difference-in-differences estimator for staggered (and non-staggered) treatment contexts.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.3

Depends data.table, sandwich

Suggests ggplot2, knitr, rmarkdown, scales, parallel, fixest, progress

VignetteBuilder knitr

URL https://setzler.github.io/DiDforBigData/

BugReports https://github.com/setzler/DiDforBigData/issues

NeedsCompilation no

Author Bradley Setzler [aut, cre, cph]

Maintainer Bradley Setzler <bradley.setzler@gmail.com>

Repository CRAN

Date/Publication 2023-04-03 15:50:02 UTC

Contents

DiD *Combine DiD estimates across cohorts and event times.*

Description

Estimate DiD for all possible cohorts and event time pairs (g,e), as well as the average across cohorts for each event time (e).

Usage

```
DiD(
  inputdata,
  varnames,
  control_group = "all",
  base_event = -1,
  min_event = NULL,
  max_event = NULL,
  Esets = NULL,
  return_ATTs_only = TRUE,
  parallel_cores = 1
)
```

Arguments

inputdata	A data.table.
varnames	A list of the form varnames = list(id_name, time_name, outcome_name, cohort_name), where all four arguments of the list must be a character that corresponds to a variable name in inputdata.
control_group	There are three possibilities: control_group="never-treated" uses the never-treated control group only; control_group="future-treated" uses those units that will receive treatment in the future as the control group; and control_group="all" uses both the never-treated and the future-treated in the control group. Default is control_group="all".
base_event	This is the base pre-period that is normalized to zero in the DiD estimation. Default is base_event=-1.
min_event	This is the minimum event time (e) to estimate. Default is NULL, in which case, no minimum is imposed.
max_event	This is the maximum event time (e) to estimate. Default is NULL, in which case, no maximum is imposed.
Esets	If a list of sets of event times is provided, it will loop over those sets, computing the average ATT_e across event times e. Default is NULL.
return_ATTs_only	Return only the ATT estimates and sample sizes. Default is TRUE.
parallel_cores	Number of cores to use in parallel processing. If greater than 1, it will try to run library(parallel), so the "parallel" package must be installed. Default is 1.

Value

A list with two components: results_cohort is a data.table with the DiDge estimates (by event e and cohort g), and results_average is a data.table with the DiDe estimates (by event e, average across cohorts g). If the Esets argument is specified, a third component called results_Esets will be included in the list of output.

Examples

```
# simulate some data
simdata = SimDiD(sample_size=200, ATTcohortdiff = 2)$simdata

# define the variable names as a list()
varnames = list()
varnames$time_name = "year"
varnames$outcome_name = "Y"
varnames$cohort_name = "cohort"
varnames$id_name = "id"

# estimate the ATT for all cohorts at event time 1 only
DiD(simdata, varnames, min_event=1, max_event=1)
```

DiDge *Estimate DiD for a single cohort (g) and a single event time (e).*

Description

Estimate DiD for a single cohort (g) and a single event time (e).

Usage

```
DiDge(
  inputdata,
  varnames,
  cohort_time,
  event_postperiod,
  base_event = -1,
  control_group = "all",
  return_data = FALSE,
  return_ATTs_only = TRUE
)
```

Arguments

inputdata	A data.table.
varnames	A list of the form varnames = list(id_name, time_name, outcome_name, cohort_name), where all four arguments of the list must be a character that corresponds to a variable name in inputdata.

`cohort_time` The treatment cohort of reference.

`event_postperiod`
Number of time periods after the cohort time at which to estimate the DiD.

`base_event` This is the base pre-period that is normalized to zero in the DiD estimation. Default is base_event=-1.

`control_group` There are three possibilities: control_group="never-treated" uses the never-treated control group only; control_group="future-treated" uses those units that will receive treatment in the future as the control group; and control_group="all" uses both the never-treated and the future-treated in the control group. Default is control_group="all".

`return_data` If true, this returns the treated and control differenced data. Default is FALSE.

`return_ATTs_only`
Return only the ATT estimates and sample sizes. Default is TRUE.

Value

A single-row data.table() containing the estimates and various statistics such as sample size. If `return_data=TRUE`, it instead returns a list in which the `data_prepost` entry is the previously-mentioned single-row data.table(), and the other argument `data_prepost` contains the constructed data that should be provided to OLS.

Examples

```
# simulate some data
simdata = SimDiD(sample_size=200)$simdata

# define the variable names as a list()
varnames = list()
varnames$time_name = "year"
varnames$outcome_name = "Y"
varnames$cohort_name = "cohort"
varnames$id_name = "id"

# estimate the ATT for cohort 2007 at event time 1
DiDge(simdata, varnames, cohort_time=2007, event_postperiod=1)

# change the base period to -3
DiDge(simdata, varnames, base_event=-3, cohort_time=2007, event_postperiod=1)

# use only the never-treated control group
DiDge(simdata, varnames, control_group = "never-treated", cohort_time=2007, event_postperiod=1)
```

`SimDiD` *DiD data simulator with staggered treatment.*

Description

Simulate data from the model Y_it = alpha_i + mu_t + ATT*(t >= G_i) + epsilon_it, where i is individual, t is year, and G_i is the cohort. The ATT formula is ATTat0 + EventTime*ATTgrowth + *cohort_counter*ATTcohortdiff, where cohort_counter is the order of treated cohort (first, second, etc.).

Usage

```
SimDiD(
  seed = 1,
  sample_size = 100,
  cohorts = c(2007, 2010, 2012),
  ATTat0 = 1,
  ATTgrowth = 1,
  ATTcohortdiff = 0.5,
  anticipation = 0,
  minyear = 2003,
  maxyear = 2013,
  idvar = 1,
  yearvar = 1,
  shockvar = 1,
  indivAR1 = FALSE,
  time_covars = FALSE,
  clusters = FALSE,
  markets = FALSE,
  randomNA = FALSE,
  missingCohorts = NULL
)
```

Arguments

seed	Set the random seed. Default is seed=1.
sample_size	Number of individuals. Default is sample_size=100.
cohorts	Vector of years at which treatment onset occurs. Default is cohorts=c(2007,2010,2012).
ATTat0	Treatment effect at event time 0. Default is 1.
ATTgrowth	Increment in the ATT for each event time after 0. Default is 1.
ATTcohortdiff	Incrememnt in the ATT for each cohort. Default is 0.5.
anticipation	Number of years prior to cohort to allow 50% treatment effects. Default is anticipation=0.
minyear	Minimum calendar year to include in the data. Default is minyear=2003.
maxyear	Maximum calendar year to include in the data. Default is maxyear=2013.
idvar	Variance of individual fixed effects (alpha_i). Default is idvar=1.
yearvar	Variance of year effects (mu_i). Default is yearvar=1.
shockvar	Variance of idiosyncratic shocks (epsilon_it). Default is shockvar=1.
indivAR1	Each individual's shocks follow an AR(1) process. Default is FALSE.

time_covars	Add 2 time-varying covariates, called "X1" and "X2". Default is FALSE.
clusters	Add 10 randomly assigned clusters, with cluster-specific AR(1) shocks. Default is FALSE.
markets	Add 10 randomly assigned markets, with market-specific shocks that are systematically greater for markets that are treated earlier. Default is FALSE.
randomNA	If TRUE, randomly assign the outcome variable with missing values (NA) in some cases. Default is FALSE.
missingCohorts	If set to a particular cohort (or vector of cohorts), all of the outcomes for that cohort at event time -1 will be set to missing. Default is NULL.

Value

A list with two data.tables. The first data.table is simulated data with variables (id, year, cohort, Y), where Y is the outcome variable. The second data.table contains the true ATT values, both at the (event,cohort) level and by event averaging across cohorts.

Examples

```
# simulate data with default options
SimDiD()
```

Index