Package 'staggR'

January 14, 2026

Type Package

Title Fit Difference-in-Differences Models with Staggered Interventions

Version 0.1.1

Description Fits linear difference-in-differences models in scenarios where intervention roll-outs are staggered over time. The package implements a version of an approach proposed by Sun and Abraham (2021) <doi:10.1016/j.jeconom.2020.09.006> to estimate cohort- and time-since-treatment specific difference-in-differences parameters, and it provides convenience functions both for specifying the model and for flexibly aggregating coefficients to answer a variety of research questions.

URL https://github.com/chse-ohsu/staggR

BugReports https://github.com/chse-ohsu/staggR/issues

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 4.1)

RoxygenNote 7.3.3

Suggests knitr, rmarkdown, sandwich

VignetteBuilder knitr

Imports ggplot2 (>= 3.0.0)

NeedsCompilation no

Author Kyle Hart [aut, cre, cph] (ORCID: <https://orcid.org/0000-0001-8964-204X>), Stephan Lindner [aut]

Maintainer Kyle Hart <hartky@ohsu.edu>

Repository CRAN

Date/Publication 2026-01-14 18:20:02 UTC

Contents

`ave_coeff` *Aggregate a specified set of terms and corresponding standard errors from an sdid model object*

Description

Aggregate a specified set of terms and corresponding standard errors from an sdid model object

Usage

```
ave_coeff(sdid, coefs)
```

Arguments

`sdid`	sdid object containing the model to summarize
`coefs`	Character vector containing the names of coefficients to aggregate. Can be specified using `select_period()` or `select_terms()`.

Value

data.frame

Examples

```
# First fit a model to generate a sdid object
sdid_hosp <- sdid(hospitalized ~ cohort + yr + age + sex + comorb,
                  df = hosp,
                  intervention_var  = "intervention_yr")

# Then request an average of a specified set of coefficients. Here we use the
# select_period() convenience function to automatically select all
# coefficients representing the post-intervention period.
ave_coeff(sdid_hosp, coefs = select_period(sdid_hosp, period = "post"))
```

```
# We could also specify the coefficients manually. Here we request the
# average effect for Cohort 5 in the post-intervention period.
ave_coeff(sdid_hosp, coefs = c("cohort_5:yr_2015", "cohort_5:yr_2016",
                               "cohort_5:yr_2017", "cohort_5:yr_2018",
                               "cohort_5:yr_2019", "cohort_5:yr_2020"))
```

`hosp` *Hospitalization data*

Description

A simulated data set of 15 counties, 11 of which implemented a policy intervention during 2015 - 2018 to reduce hospitalizations. The data set is longitudinal, with each row corresponding to an individual-year.

Usage

```
hosp
```

Format

`hosp`**:**

A data frame with 31,040 rows and 10 columns:

guid Character vector containing globally unique identifiers for individuals living in the 15 counties

county Character vector containing county names

intervention_dt Dates on which each county implemented their policy intervention to reduce hospitalizations

intervention_yr Character vector containing the year during which `intervention_dt` takes place

age Integer containing individuals' ages. Time-varying by year.

sex Character vector containing individuals' sexes. Not time-varying.

comorb Logical indicating whether each individual has comorbidities. Time-varying by year.

cohort Character vector identifying the intervention cohort to which each individual belongs. Takes values 0, 5, 6, 7, or 8, corresponding to counties that implemented the intervention not at all or during 2015, 2016, 2017, or 2018, respectively. Invariant within counties.

yr Character vector representing the observation year for each row.

hospitalized Integer indicating whether the individual was hospitalized during the current year.

Details

Consider a policy intervention designed to reduce inpatient hospitalizations in 15 counties. This longitudinal data set has one row per individual-year. Each individual is identified by a globally unique identifier (`guid`), and we have measures of the individuals' ages, sexes, and comorbidities, and a column indicating whether the individual was hospitalized during the current year.

The column intervention_yr tells us the year during which each county implemented the intervention. If intervention_yr is NA, we can conclude that the county never implemented the intervention. Among the 15 counties, 3 implemented the intervention in 2015; 2 counties implemented in 2016; 5 counties implemented in 2017; 1 county implemented in 2018; and 4 counties did not implement the intervention at all during the study period, which runs for 11 years, from 2010 through 2020.

---

hosp_agg *Aggregated hospitalization data*

Description

A simulated data set of 15 counties, 11 of which implemented a policy intervention during 2015 - 2018 to reduce hospitalizations. The data set is longitudinal and aggregated to county-year.

Usage

```
hosp_agg
```

Format

hosp_agg**:**

A data frame with 31,040 rows and 10 columns:

yr Character vector representing the observation year for each row.

county Character vector containing county names

cohort Character vector identifying the intervention cohort to which each county belongs. Takes values 0, 5, 6, 7, or 8, corresponding to counties that implemented the intervention not at all or during 2015, 2016, 2017, or 2018, respectively. Invariant within counties.

intervention_yr Character vector containing the year during which intervention_dt takes place

pct_hospitalized Numeric vector containing the proportion of individuals in each county-year who were hospitalized.

n_enr Integer indicating the number of individuals living in each county during the curent year.

mean_age Numeric containing mean ages among individuals living in each county during the current year.

pct_fem Numeric containing the proportion of individuals in each county-year who are female.

pct_cmb Numeric containing the proportion of individuals in each county-year who have comorbidities.

Details

Consider a policy intervention designed to reduce inpatient hospitalizations in 15 counties. This longitudinal data set has one row per county-year and includes aggregated measures of individuals' ages, sexes, and comorbidities, and a column indicating proportion of individuals who were hospitalized during the current year.

The column intervention_yr tells us the year during which each county implemented the intervention. If intervention_yr is NA, we can conclude that the county never implemented the

intervention. Among the 15 counties, 3 implemented the intervention in 2015; 2 counties implemented in 2016; 5 counties implemented in 2017; 1 county implemented in 2018; and 4 counties did not implement the intervention at all during the study period, which runs for 11 years, from 2010 through 2020.

`id_tsi` *Identify time-since-intervention*

Description

`id_tsi()` identifies the number of time periods relative to the intervention for each observation. This information is used for plotting and for aggregating model coefficients with `ave_coeff()`.

Usage

```
id_tsi(df, cohort_var, time_var, intervention_var)
```

Arguments

`df`	Data frame containing the variables in the model.
`cohort_var`	Name of the variable in `df` that contains cohort assignments.
`time_var`	Name of the variable in `df` that contains time periods.
`intervention_var`	Name of the cohort-level variable in `df` that specifies which values in `time_var` correspond to the first post-intervention time period for each cohort.

Value

`tsi` Object containing a data frame showing time since intervention for each time period in the data frame for each cohort in the data frame.

Examples

```
# Generate a tsi object, containing a data frame showing the time since
# intervention (TSI value) for each time period in the data frame for each
# cohort.
id_tsi(hosp,
       cohort_var = "cohort",
       time_var = "yr",
       intervention_var = "intervention_yr")
```

`pick_time_refs` *Identify time period referents within each cohort.*

Description

Identify time period referents within each cohort.

Usage

```
pick_time_refs(
  df,
  cohort_var,
  cohort_ref,
  time_var,
  intervention_var = NULL,
  time_offset = -1
)
```

Arguments

`df`	A data frame containing the variables in the model.
`cohort_var`	String specifying the name of the column in `df` that defines the intervention cohorts.
`cohort_ref`	An optional string specifying the value of `cohort_var` to be used as the referent in the model. If not specified, the value is taken from the first observed value in `cohort_var`.
`time_var`	String specifying the name of the column in `df` that defines time periods over the study.
`intervention_var`	String specifying the name of the column in `df` that defines the intervention period. If values of `cohort_var` are named to match values of `time_var`, this parameter is not necessary.
`time_offset`	Integer specifying which time period relative to the intervention time period should be used as the referent for each cohort. Defaults to -1, which corresponds to the time period immediately preceding intervention.

Value

list

Examples

```
pick_time_refs(hosp, "cohort", "0", "yr", "intervention_yr")
```

prep_data *Prepare a data frame to work with sdid() function*

Description

Prepare a data frame to work with sdid() function

Usage

```
prep_data(df, cohort_var, cohort_ref = NULL, time_var, time_ref = NULL)
```

Arguments

df	A data frame containing the variables in the model.
cohort_var	String specifying the name of the column in df that defines the intervention cohorts.
cohort_ref	An optional string specifying the value of cohort_var to be used as the referent in the model. If not specified, the value is taken from the first observed value in cohort_var.
time_var	String specifying the name of the column in df that defines time periods over the study.
time_ref	An optional string specifying the value of time_var to be used as the referent in the model.

Value

data.frame

Examples

```
dta_prepped <- prep_data(hosp,
                         cohort_var = "cohort",
                         cohort_ref = "0",
                         time_var = "yr",
                         time_ref = "2010")
head(dta_prepped)
```

sdid *Fit a staggered difference-in-differences model*

Description

Fits a linear staggered difference-in-differences model, following the Abraham and Sun (2018) approach. It facilitates optional weighting and user-specified variance-covariance function.

Usage

```
sdid(
  formula,
  df,
  weights = NULL,
  cohort_var = NULL,
  cohort_ref = NULL,
  cohort_time_refs = NULL,
  time_var = NULL,
  time_ref = NULL,
  intervention_var,
  .vcov = stats::vcov,
  ...
)
```

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. The details of model specification are given under 'Details'.
df	A data frame containing the variables in the model.
weights	An optional vector of weights to be passed to `stats::lm()` to be used in the fitting process. Should be NULL or a numeric vector.
cohort_var	Name of the variable in `df` that contains cohort assignments. If NULL, this is assumed to be the first column named in the right hand side of `formula`.
cohort_ref	Value of `cohort_var` that serves as the referent for main effects for cohorts. If NULL, this is assumed to the be the first value in the set of values for `cohort_var`.
cohort_time_refs	A list, whose elements are named to match levels of `cohort_var`, specifying the value of `time_var` that serves as the referent for each time interaction with values of `cohort_var`. See 'Details.'
time_var	Name of the variable in `df` that contains time periods. If NULL, this is assumed to be the second column named in the right hand side of `formula`.
time_ref	Value of `time_var` that serves as the referent for main effects for time periods. If NULL, this is assumed to the be the first value in the set of values for `time_var`.

`intervention_var`
: Name of the cohort-level variable in `df` that specifies which values in `time_var` correspond to the first post-intervention time period for each cohort.

`.vcov`
: Function to be used to estimate the variance-covariance matrix. Defaults to stats::vcov.

`...`
: Additional arguments to be passed to `.vcov`.

Details

Fitting a staggered difference-in-differences model requires deliberate attention to two specific independent variables:

- The intervention cohort column assigns a cohort name to all individuals or groups having the the intervention during the same time period. For example, if the longitudinal data is at the year level, ranging from 2010 to 2020, and it contains 15 counties, 3 of whom implemented the intervention of interest in 2015, those 3 counties would be assigned to the same cohort. Similarly, if 2 more counties implemented the intervention in 2016, those 2 counties would be assigned to the next cohort.
- The time period column assigns each observation to a time period at the most granular level of the longitudinal data. In the example described above, these values would correspond to the years 2010, ..., 2020.

To specify a model, a formula is passed following the format `response ~ cohort_var + time_var + covariates`. This, however, is not the formula use to fit the model; `sdid()` expands this formula to include main effects and every possible interaction between `cohort_var` and `time_var`, excluding referents for identification:

- Referents for main effects are either the first levels `cohort_var` and `time_var` or the referents specified in `cohort_ref` and `time_ref`.
- Referents for cohort-time interactions are either the factor level of `time_var` that immediately precedes the value of `intervention_var` within each cohort or the referents specified in `cohort_time_refs`.

`sdid()` also accommodates aggregated data through the `weights` argument.

Value

Returns an object of class `sdid`, which is a list containing the following components:

mdl : The `lm` object returned from the call to `stats::lm()` in `sdid()`

formula : A list object containing both the original formula specified in the call to `sdid()` and the generated formula, with all cohort-time interactions, passed to `stats::lm()` to fit the model

vcov : The variance-covariance matrix used to estimate standard errors

tsi : The time-since-intervention dataset used to enumerate time periods relative to the intervention period for each cohort

obs_cnt : Counts of observations within each cohort-time interaction cohort : A list object containing details about cohorts. `var` contains the name of the column in `df` that identifies cohorts; `ref` contains the value of the cohort column that functions as the referent for main effects; and

time_refs contains the referent time values within each cohort for each set of cohort-time interactions.

time : A list object containing var, which is the name of the column in df identified by the sdid() argument time_var, and ref, the referent value of time_var for main effects.

intervention_var : Name of the column in df that contains the time period during which each cohort implemented the intervention of interest

covariates : A character vector containing the terms in formula other than those corresponding to cohorts and time periods

References

Abraham S, Sun L. Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. MIT; 2018.

Examples

```
# Fit a staggered difference-in-differences model
sdid_hosp <- sdid(hospitalized ~ cohort + yr + age + sex + comorb,
                  df = hosp,
                  intervention_var  = "intervention_yr")
summary(sdid_hosp)
```

---

select_period *Retrieve a list of interaction terms from a sdid model representing the pre-intervention period*

Description

Retrieve a list of interaction terms from a sdid model representing the pre-intervention period

Usage

```
select_period(sdid, period = "post", cohorts = NULL)
```

Arguments

sdid	A sdid object
period	One of 'pre' or 'post', to return the pre-intervention or post-intervention coefficients respectively
cohorts	A character vector containing cohort levels to include in the term selection. If cohorts is omitted, all available cohorts will be selected

Value

character vector

Examples

```
# Fit a staggered difference-in-differences model
sdid_hosp <- sdid(hospitalized ~ cohort + yr + age + sex + comorb,
                  df = hosp,
                  intervention_var  = "intervention_yr")

# Select coefficients corresponding to the PRE-intervention period for cohort 5
coef_selection_pre <- select_period(sdid_hosp,
                                    period = "pre",
                                    cohorts = "5")
coef_selection_pre

# Pass the set of coefficients to `ave_coeff` to aggregate the effect of the
# intervention
ave_coeff(sdid_hosp, coefs = coef_selection_pre)

# Select coefficients corresponding to the POST-intervention period for cohort 5
coef_selection_post <- select_period(sdid_hosp,
                                     period = "post",
                                     cohorts = "5")
coef_selection_post

# Pass the set of coefficients to `ave_coeff` to aggregate the effect of the
# intervention
ave_coeff(sdid_hosp, coefs = coef_selection_post)
```

`select_terms` *Retrieve a list of interaction terms from a sdid model to be passed on for aggregation*

Description

Retrieve a list of interaction terms from a sdid model to be passed on for aggregation

Usage

```
select_terms(sdid, coefs = NULL, selection = NULL)
```

Arguments

`sdid`	A sdid object
`coefs`	Optional list of specific terms from `mdl` to be selected
`selection`	List object containing values for named elements `cohorts`, `times`, and `tsi`. `cohorts` contains a character vector of cohort levels to include in the term selection; `times` contains a character vector of time period levels to include in the term selection; and `tsi` contains a vector of integers representing the number of units of time relative to each cohort's intervention to include in the term selection. If `cohorts` is omitted, all available cohorts will be selected. One of `times` or `tsi` must be specified. If both are specified, `times` is ignored.

Value

character vector

Examples

```
# Fit a staggered difference-in-differences model
sdid_hosp <- sdid(hospitalized ~ cohort + yr + age + sex + comorb,
                          df = hosp,
                          intervention_var  = "intervention_yr")

# Select coefficients corresponding to all intervention cohorts in 2018
terms_2018 <- select_terms(sdid = sdid_hosp,
                                  selection = list(times = "2018"))
terms_2018

# Pass the set of coefficients to `ave_coeff` to aggregate the effect of the
# intervention
ave_coeff(sdid_hosp, coefs = terms_2018)

# Select coefficients corresponding to added risk of hospitalization associated with
# the intervention in the year 2018, but only for the first two cohorts (5 and 6)
terms_2018_cohorts56 <- select_terms(sdid = sdid_hosp,
                                     selection = list(cohorts = c("5", "6"),
                                                      times = "2018"))

# Pass the set of coefficients to `ave_coeff` to aggregate the effect of the
# intervention
ave_coeff(sdid_hosp, coefs = terms_2018_cohorts56)
```

`summary.sdid_mdl` *Summarize an sdid model*

Description

Summarize an sdid model

Usage

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

Arguments

`object`	A `sdid_mdl` object.
`...`	Passed through.

Value

An object of class `summary.sdid_mdl`.

Examples

```
# Fit a staggered difference-in-differences model
sdid_hosp <- sdid(hospitalized ~ cohort + yr + age + sex + comorb,
                  df = hosp,
                  intervention_var  = "intervention_yr")
# Summarize the results
summary(sdid_hosp)
```

`ts_plot` *Generates time-series plots, optionally faceted by groups*

Description

Generates time-series plots, optionally faceted by specified groups. The resulting object can be customized using `ggplot2` functions and themes.

Usage

```
ts_plot(
  formula = NULL,
  y = NULL,
  group = NULL,
  time_var = NULL,
  intervention_var = NULL,
  df,
  tsi = NULL,
  weights = NULL,
  ncol = 2
)
```

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. The details of model specification are given under 'Details'.
`y`	Name of the variable in `df` that contains the outcome of interest. If NULL, this is assumed to be the column named in the left-hand side of `formula`.
`group`	Name of the variable in `df` that contains cohort assignments or other groups by which the plot should be faceted. If NULL, this is assumed to be the first column named in the right-hand side of `formula`. If no `formula` is specified, the resulting plot will aggregate all results into a single panel.
`time_var`	Name of the variable in `df` that contains time periods. If NULL, this is assumed to be the second column named in the right-hand side of `formula`.
`intervention_var`	Name of the cohort-level variable in `df` that specifies which values in `time_var` correspond to the first post-intervention time period for each cohort. If NULL, vertical lines indicating the intervention period will be omitted from the plot.

df	A data frame containing the variables in the model.
tsi	An object of class tsi, created by tsi(), that defines the number of time periods relative to the intervention time period for each cohort observation.
weights	An optional vector of weights to be passed to lm() to be used in the fitting process. Should be NULL or a numeric vector.
ncol	Number of columns in the faceted plot

Value

Returns an object of class "ggplot"

Examples

```
# Use a formula to specify the setup of the time-series plot. Here we set
# hospitalized as the outcome, faceted by county, with yr on the X axis.
ts_plot(hospitalized ~ county + yr,
        df = hosp,
        intervention_var = "intervention_yr")

# We can specify the same plot without using a formula.
ts_plot(y = "hospitalized",
        group = "county",
        time_var = "yr",
        df = hosp,
        intervention_var = "intervention_yr")
```

Index