Package ‘surveil’

November 11, 2021

Title  Time Series Models for Disease Surveillance
Version  0.1.0


License  GPL (>= 3)
Encoding  UTF-8
LazyData  true
RoxygenNote  7.1.1
Biarch  true
Depends  R (>= 3.5.0)
Imports  rstantools (>= 2.1.1), methods, Rcpp (>= 0.12.0), RcppParallel (>= 5.0.1), rstan (>= 2.18.1), tidybayes (>= 3.0.0), dplyr (>= 1.0.7), rlang (>= 0.4.0), tidyr (>= 1.1.0), ggplot2 (>= 3.0.0), gridExtra (>= 2.0), scales (>= 0.4.0), ggdist (>= 3.0.0)
LinkingTo  BH (>= 1.66.0), Rcpp (>= 0.12.0), RcppEigen (>= 0.3.3.3.0), RcppParallel (>= 5.0.1), rstan (>= 2.18.1), StanHeaders (>= 2.18.0)
Suggests  rmarkdown, knitr, testthat
SystemRequirements  GNU make
VignetteBuilder  knitr
surveil-package

The 'surveil' package

Description

References


---

apc

**Annual and cumulative percent change**

**Description**

Summarize annual and cumulative percent change in risk

**Usage**

```r
apc(x)
```

```
## S3 method for class 'surveil'
apc(x)
```

```
## S3 method for class 'stand_surveil'
apc(x)
```

**Arguments**

- **x** A fitted surveil model, or standardized rates (a stand_surveil object).

**Value**

An apc (list) object containing the following data frames:

- **apc** A data frame containing a summary of the posterior distribution for period-specific percent change. This contains the posterior mean (apc) 95 percent credible intervals (lwr and upr bounds).

- **cpc** A data frame containing a summary of the posterior distribution for the cumulative percent change in risk at each time period. This contains the posterior mean (cpc) and 95 percent credible interval (lwr and upr bounds).

- **apc_samples** MCMC samples from the posterior distribution for period percent change

- **cpc_samples** MCMC samples from the posterior distribution for cumulative percent change
cancer

US cancer incidence by age, 1999-2017

Description

Annual cancer cases (all sites) by age group for the United States.

Usage
cancer

Format

A data frame with the following columns:

Year  Year of diagnosis
Age   Age group
Count Number of cancer cases
Population Age-specific population estimates

Source


Examples
data(cancer)
head(cancer)
**group_diff**

**Measures of pairwise inequality**

**Description**

Calculate pairwise measures of health inequality from a fitted surveil time series model, with credible intervals and MCMC samples. Calculates absolute and fractional rate differences (RD and population attributable risk (PAR)), rate ratios, and excess cases.

**Usage**

`group_diff(fit, target, reference)`

**Arguments**

- **fit**: A fitted surveil time series model
- **target**: The name (character string) of the disadvantaged group that is the target of inference.
- **reference**: The name (character string) of the reference group to which target will be compared.

**Details**

For the following calculations, the terms reference and target refer to incidence rates for the respective groups; \( p \) is the size of the target population. (Target is the group that is the 'target' of our inferences, so that it is the numerator in rate ratios, etc.) The following measures are calculated by `group_diff`:

- **rate difference**
  
  \[ RD = \text{target} - \text{reference} \]

- **population attributable fraction**
  
  \[ PAR = \frac{RD}{\text{target}} = \frac{RR - 1}{RR} \]

- **rate ratio**
  
  \[ RR = \text{target}/\text{reference} \]

- **excess cases**
  
  \[ EC = RD \times p \]

As the math communicates, the PAR is the rate difference expressed as a fraction of total risk for the target population. This could also be read as the fraction of risk in the target population that would have been removed had the target rate equaled the reference rate (Menvielle et al. 2017).

**Value**

A list, also of class "surveil_diff", with the following elements:

- **summary**: A tibble with a summary of posterior distributions (mean and 95 percent cred. intervals) for the target group incidence rate, the RD, RR, PAR, and excess cases.
cumulative_cases  Summary of the posterior distribution for the cumulative number of excess cases and the PAR (mean and 95 percent cred. intervals)
groups  Character string with target and reference population names
samples  A data frame of MCMC samples for each quantity of interest (target and reference rates, RD, RR, PAR, and EC, as well as Trend_Cases = Rate * Population). Indexed by time.
cum_samples  MCMC samples of the cumulative number of excess cases.

Author(s)
Connor Donegan (Connor.Donegan@UTSouthwestern.edu)

Source

See Also
plot.surveil_diff print.surveil_diff theil

Examples

data(msa)
houston <- msa[grep("Houston", msa$MSA), ]
fit <- stan_rw(houston, time = Year, group = Race, iter = 1500)
gd <- group_diff(fit, "Black or African American", "White")
print(gd, scale = 100e3)
plot(gd, scale = 100e3)

msa  

Colorectal cancer incidence by Texas MSA, 1999-2017, ages 50-79

Description
Annual counts of colorectal cancer (cancer of colon or rectum), ages 50-79, for Texas’s top four metropolitan statistical areas (MSAs), with population at risk estimates, by race-ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic/Latino).

Usage
msa
Format

A tibble with the following attributes:

- **Year**  Year of diagnosis
- **Race**  Race-ethnicity designation
- **MSA**  Metropolitan statistical area
- **Count**  Number of CRC cases
- **Population**  Age-specific population estimate

Source


Examples

```r
data(msa)
head(msa)
```

---

**plot.surveil**  
*Methods for fitted surveil models*

Description

Print and plot methods for surveil model results

Usage

```r
## S3 method for class 'surveil'
print(x, scale = 1, ...)

## S3 method for class 'surveil'
plot(
  x,
  scale = 1,
  style = c("mean_qi", "lines"),
  facet = FALSE,
  base_size = 14,
  palette = "Dark2",
  M = 250,
  alpha,
  lwd,
  fill = "gray80",
)```
size = 1.5,
...
)

Arguments

x A fitted surveil model
scale Scale the rates by this amount; e.g., scale = 100e3 will print rates per 100,000 at risk.
... For the plot method, additional arguments will be passed to `theme`; for the print method, additional arguments will be passed to `print.data.frame`.
style If `style = "mean_qi"`, then the posterior mean and 95 percent credible interval will be plotted; if `style = "lines"`, then M samples from the joint probability distribution of the annual rates will be plotted.
facet If `facet = TRUE`, `facet_wrap` will be used instead of differentiating by line color.
base_size Passed to `theme_classic()` to control size of plot components (text).
palette For multiple groups, choose the color palette. For a list of options, see `scale_color_brewer`. The default is palette = "Dark2". Not used if `facet = TRUE`.
M If `style = "lines"`, then M is the number of samples from the posterior distribution that will be plotted; the default is M = 250.
alpha Numeric value from zero to one. When `style = "lines"`, this controls transparency of lines; passed to `geom_line`. For `style = "mean_qi"`, this controls the transparency of the shaded credible interval; passed to `geom_ribbon`.
lwd Numeric value indicating linewidth. Passed to `geom_line`
fill Color for the shaded credible intervals; only used when `style = "mean_qi"`.
size Positive numeric value. For `style = "mean_qi"`, this controls the size of the points representing crude rates. To exclude these points from the plot altogether, use `size = 0`.

Value

The plot method returns a `ggplot` object; the print method returns nothing but prints a summary of results to the R console.

Author(s)

Connor Donegan (Connor.Donegan@UTSouthwestern.edu)

See Also

`stan_rw`
Examples

```r
data(msa)
houston <- msa[grep("Houston", msa$MSA), ]
fit <- stan_rw(houston, time = Year, group = Race, iter = 1500)

print(fit)

## plot probability distribution for disease risk
plot(fit, style = "lines")
plot(fit, facet = TRUE, scale = 100e3)

## as a ggplot, you can customize the output
library(ggplot2)
plot(fit) + theme_bw()
```

---

plot.theil

Methods for Theil's index

Description

Printing and plotting methods for Theil’s inequality index

Usage

```r
## S3 method for class 'theil'
plot(
x, 
style = c("mean_qi", "lines"),
M = 250,
col = "black",
fill = "black",
alpha,
lwd,
base_size = 14,
scale = 100,
labels = x$summary$time,
...
)

## S3 method for class 'theil_list'
plot(
x, 
style = c("mean_qi", "lines"),
M = 250,
```
col = "black",
fill = "black",
alpha,
lwd,
between_title = "Between",
within_title = "Within",
total_title = "Total",
scale = 100,
plot = TRUE,
base_size = 14,
... 
)

## S3 method for class 'theil'
print(x, scale = 100, digits = 3, ...)

## S3 method for class 'theil_list'
print(x, scale = 100, digits = 3, ...)

Arguments

x An object of class thiel’ or theil_list, as returned by calling theilon a list of fitted surveil’ models

style If style = "mean_qi", then the posterior mean and 95 percent credible interval will be plotted; if style = "lines", then M samples from the joint probability distribution will be plotted.

M If style = "lines", then M is the number of samples from the posterior distribution that will be plotted; the default is M = 250.

col Line color

fill Fill color

alpha For style = "mean_qi", this controls the transparency for the credible interval (passed to geom_ribbon) and defaults to alpha = 0.5; for style = "lines", this controls the transparency of the lines and defaults to alpha = 0.7.

lwd Line width; for style = "mean_qi", the default is lwd = 1; for style = "lines", the default is lwd = 0.05.

base_size Passed to theme_classic to control size of plot elements (e.g., text)

cscale Scale Theil’s index by scale

labels x-axis labels (time periods)

... additional arguments

between_title Plot title for the between geography component of Theil’s T; defaults to "Between”.

within_title Plot title for the within geography component of Theil’s T; defaults to "Within”.

total_title Plot title for Theil’s index; defaults to "Total”.

plot If FALSE, return a list of ggplots. Not used when style = "lines”.

digits number of digits to print (passed to print.data.frame)
**Value**

**plot.theil:**
The plot method returns an object of class `ggplot`.

**plot.theil_list:**
If `style = "lines"`, the plot method for `theil_list` objects returns a `ggplot` with facets for each component of inequality (between-areas, within-areas, and total). For `style = "mean_qi"`, the plot method returns either a list of plots (all of class `ggplot`) or, when `plot = TRUE`, it will draw them to current plotting device using `grid.arrange`.

**print.theil:**
The print returns nothing and method prints a summary of results to the R console.

**See Also**

`theil`

---

**print.apc**

Methods for APC objects

**Description**

Methods for APC objects

**Usage**

```r
## S3 method for class 'apc'
print(x, digits = 1, max = 10, ...)

## S3 method for class 'apc'
plot(
x, 
cumulative = FALSE, 
style = c("mean_qi", "lines"), 
M = 250, 
col = "black", 
fill = "black", 
alpha, 
lwd, 
base_size = 14, 
... 
)
```
Arguments

- **x**: An `apc` object returned by `apc`
- **digits**: Print this many digits (passed to `print.data.frame`)
- **max**: Maximum number of time periods (rows) to print
- **...**: Additional arguments; for the print argument, these will be passed to `print.data.frame`. For the plot method, these will be passed to `theme`.
- **cumulative**: Plot cumulative percent change? Defaults to `cumulative = FALSE`.
- **style**: If `style = "mean_qi"`, then the posterior mean and 95 percent credible interval will be plotted; if `style = "lines"`, then $M$ samples from the joint probability distribution will be plotted.
- **$M$$**: If `style = "lines"`, then $M$ is the number of samples from the posterior distribution that will be plotted; the default is $M = 250$.
- **col**: Line color
- **fill**: Fill color for the 95 percent credible interval
- **alpha**: For `style = "mean_qi"`, this controls the transparency for the credible interval (passed to `geom_ribbon`) and defaults to `alpha = 0.5`; for `style = "lines"`, this controls the transparency of the lines and defaults to `alpha = 0.7`.
- **lwd**: Line width
- **base_size**: Size of plot attributes, passed to `theme_classic`

Value

- **print**: The print method does not have a return value, but prints a summary of results to the R console.

- **Plot**: The plot method returns a `ggplot`.

See Also

- `apc`

---

**print.stand_surveil**  
*Methods for standardized rates*

**Description**

Print and plot methods for `stand_surveil` (standardized rates obtained from a fitted `surveil` model)
Usage

```r
## S3 method for class 'stand_surveil'
print(x, scale = 1, digits = 3, ...)

## S3 method for class 'stand_surveil'
plot(
  x,
  scale = 1,
  style = c("mean_qi", "lines"),
  M = 250,
  base_size = 14,
  col = "black",
  fill = "gray80",
  alpha,
  lwd,
  ...
)
```

Arguments

- `x`: An object of `stand_surveil` obtained by calling `standardize` on a fitted `surveil` model.
- `scale`: Scale the rates by this amount; e.g., `scale = 100e3` will print rates per 100,000 at risk.
- `digits`: Number of digits to print.
- `...`: Additional arguments.
- `style`: If `style = "mean_qi"`, then the posterior means and 95 percent credible intervals will be plotted; if `style = "lines"`, then `M` samples from the joint posterior distribution will be plotted.
- `M`: Number of samples to plot when `style = "lines"`
- `base_size`: Passed to `theme_classic()` to control size of plot components (text).
- `col`: Line color.
- `fill`: Fill color for the 95 percent credible intervals.
- `alpha`: For `style = "mean_qi"`, this controls the transparency for the credible interval (passed to `geom_ribbon`) and defaults to `alpha = 0.5`; for `style = "lines"`, this controls the transparency of the lines and defaults to `alpha = 0.7`.
- `lwd`: Line width; for `style = "mean_qi"`, the default is `lwd = 1`; for `style = "lines"`, the default is `lwd = 0.05`.

Details

Calling `standardize` on a fitted `surveil` model will create a new object that contains the `surveil` model results as well standardized rates. This new `stand_surveil` object has its own methods for printing and plotting.
print.stand_surveil:
Any additional arguments (…) will be passed to print.data.frame

plot.stand_surveil:
Any additional arguments (…) will be passed to ’theme.

Value

print.stand_surveil:
The print method returns nothing but prints a summary of results to the console.

plot.stand_surveil:
The plot method returns an object of class ggplot.

See Also

standardize stan_rw

---

## Prior distributions

**Description**

Prior distributions

**Usage**

normal(location = 0, scale, k = 1)

lkj(eta)

**Arguments**

- location: Location parameter (numeric)
- scale: Scale parameter (positive numeric)
- k: Optional; number of groups for which priors are needed. This is a shortcut to avoid using the rep function to repeat the same prior for each group, as in: normal(location = rep(0, times = 3), scale = rep(1, times = 3). To provide distinct priors for each group, simply specify them individually, as in normal(location = c(-5,-6,-8), scale = c(2,2,2)).
- eta: The shape parameter for the LKJ prior
The prior distribution functions are used to set the values of prior parameters. Users can control the values of the parameters, but the distribution (model) itself is fixed. The first log-rate \( \eta[t], t=1 \) and the scale parameters \( \sigma \) are assigned Gaussian (normal) prior distribution. (The scale parameter, \( \sigma \), is constrained to be positive, making it a half-normal prior.) For correlated time series, the correlation matrix is assigned the LKJ prior.

**Parameterizations:**
For details on how any distribution is parameterized, see the Stan Language Functions Reference document: [https://mc-stan.org/users/documentation/](https://mc-stan.org/users/documentation/).

**LKJ prior:**
The LKJ prior for correlation matrix has a single parameter, \( \eta (\eta > 0) \). If \( \eta=1 \), then you are placing a uniform prior on any \( K \)-by-\( K \) correlation matrix. For \( \eta > 1 \), there is a higher probability on the identity matrix, such that as \( \eta \) increases beyond 1, you are expressing greater scepticism towards large correlations. If \( 0 < \eta < 1 \), then you will be expressing scepticism towards correlations of zero and favoring non-zero correlations. See Stan documentation: [https://mc-stan.org/docs/2_27/functions-reference/lkj-correlation.html](https://mc-stan.org/docs/2_27/functions-reference/lkj-correlation.html).

**Value**
An object of class `prior` which will be used internally by `surveil` to set parameters of prior distributions.

**Source**

**Examples**

```r
# note there are three groups in the data, each requires a prior
prior <- list()
prior$eta_1 <- normal(location = -6, scale = 4, k = 3)
## by default, location = 0
prior$sigma <- normal(scale = 1, k = 3)
prior$omega <- lkj(2)

dfw <- msa[grep("Dallas", msa$MSA), ]
fit <- stan_rw(dfw, time = Year, group = Race, prior = prior, iter = 1200)
plot(fit)
```
standardize

---

### standard

**2000 U.S. standard million population**

---

#### Description

2000 U.S. standard million population

#### Usage

```r
standard
```

#### Format

An object of class `data.frame` with 19 rows and 3 columns.

#### Source


#### Examples

```r
data(standard)
head(standard)
```

---

### standardize

**Standardized rates**

---

#### Description

Convert surveil model results to age standardized rates using a fixed age distribution

#### Usage

```r
standardize(x, label, standard_pop)
```

#### Arguments

- **x**: A fitted `surveil` model
- **label**: Labels (character strings) for the age groups that correspond to the values of `stand_pop`. The labels must match the grouping variable used to fit the model (i.e., `all(label %in% names(x$data$cases))` must be true).
- **standard_pop**: Standard population values corresponding to the age groups specified by `label`
**Value**

A list, also of class "stand_surveil", containing a summary data frame (mean and 95 percent credible intervals) (named `summary`), a data frame of MCMC samples of standardized rates per time period (named `samples`), and the user-provided labels and standard population sizes (named `label` and `standard_pop`). In addition, all of the items from the user-provided surveil model are automatically appended to the list.

**See Also**

`stan_rw` `plot.stand_surveil` `print.stand_surveil`

**Examples**

```r
data(cancer)
data(standard)

head(standard)
head(cancer)

fit <- stan_rw(cancer,
    time = Year,
    group = Age
)

stands <- standardize(fit,
    label = standard$age,
    standard_pop = standard$standard_pop)

print(stands)
plot(stands, style = "lines")
```

---

**stan_rw**

*Time series models for mortality and disease incidence*

**Description**

Model time-varying disease risk given time series of case (or death) counts and population at risk.

**Usage**

```r
stan_rw(
    data,
    group,
    time,
    cor = FALSE,
    prior = list(),
    chains = 4,
)```
cores = 1,
iter = 3000,
refresh = 1500,
control = list(adapt_delta = 0.98),
...
)

Arguments

data A data.frame containing the following columns:

Count Number of cases or deaths; this column must be named 'Count'.
Population Size of population at risk; this column must be named 'Population'.
time Time period indicator. (Provide the (unquoted) column name using the
time argument.)
group Optional grouping variable. (Provide the (unquoted) column name using
the group argument.)

group If data is aggregated by demographic group, provide the (unquoted) name of the
column in data containing the grouping structure, such as age brackets or race-
ethnicity. E.g., if data has column names Year, Race, Cases, and Population,
then you would provide group = Race.
time Specify the (unquoted) name of the time variable in data, as in time = Year.
This variable must be numeric-alike (i.e., as.numeric(data$time) will not
fail).
cor For correlated random walks use cor = TRUE; default value is FALSE. Note this
only applies when the group argument is used.
prior Optionally provide a named list with prior parameters. If any of the following
items are missing, default priors will be assigned and printed to the console.

eta_1 The first value of log-risk in each series must be assigned a Gaussian prior
probability distribution. Provide the location and scale parameters for each
demographic group in a list, where each parameter is a k-length vector.
For example, with k=2 demographic groups, the following code will assign
priors of normal(-6.5,5) to the starting values of both series: prior = list(eta_1 = nor-
mal(location = -6.5, scale = 5, k = 2). Note, eta is the log-rate, so centering
the prior for eta_1 on -6.5 is similar to centering the prior rate on exp(-
6.5)*100,000 = 150 cases per 100,000 person-years at risk. Note, however,
that the translation from log-rate to rate is non-linear.
sigma Each demographic group has a scale parameter assigned to its log-rate.
This is the scale of the annual deviations from the previous year's log-rate.
The scale parameters are assigned independent half-normal prior distribu-
tions (these half normal distributions are restricted to be positive-valued
only).
omega If cor = TRUE, an LKJ prior is assigned to the correlation matrix, Omega.

chains Number of independent MCMC chains to initiate (passed to sampling).
cores The number of cores to use when executing the Markov chains in parallel (passed
to sampling).
iter   Total number of MCMC iterations. Warmup draws are automatically half of iter.
refresh How often to print the MCMC sampling progress to the console.
control A named list of parameters to control Stan’s sampling behavior. The most common parameters to control are adapt_delta, which may be raised to address divergent transitions, and max_treedepth. For example, control = list(adapt_delta = 0.99, max_treedepth = 13), may be a reasonable specification to address a divergent transitions or maximum treedepth warning from Stan.

... Other arguments passed to sampling.

Details

For time $t = 1,...,n$, the models have Poisson likelihoods for the case counts, given log-risk $\eta_t$ and population at time $P_t$; the log-risk is modeled using the first-difference (or random-walk) prior:

$$y_t \sim \text{Poisson}(P_t \times \exp(\eta_t))$$
$$\eta_t \sim \text{Normal}(\eta_{t-1}, \sigma)$$
$$\eta_1 \sim \text{Normal}(-6, 5) (-\infty, 0)$$
$$\sigma \sim \text{Normal}(0, 1) (0, \infty)$$

This style of model has been discussed in Bayesian (bio)statistics for quite some time. See Clayton (1996).

The above model can be used for multiple distinct groups; in that case, each group will have its own independent time series model. However, it is also possible to add a correlation structure to that set of models. Let $Y_t$ be a $k$-length vector of observations for each of $k$ groups at time $t$ (the capital letter now indicates a vector), then:

$$Y_t \sim \text{Poisson}(P_t \times \exp(Eta_t))$$
$$Eta_t \sim \text{MVNormal}(Eta_{t-1}, Sigma)$$
$$Eta_1 \sim \text{Normal}(-6, 5) (-\infty, 0)$$
$$Sigma = \text{diag}(sigma) \times Omega \times \text{diag}(sigma)$$
$$Omega \sim \text{LKJ}(2)$$
$$sigma \sim \text{Normal}(0, 1) (0, \infty)$$

where $Omega$ is a correlation matrix and $\text{diag}(sigma)$ is a diagonal matrix with scale parameters on the diagonal. This was adopted from Brandt and Williams (2007); for the LKJ prior, see the Stan Users Guide and Reference Manual.

Value

The function returns a list, also of class surveil, containing the following elements:

- **summary** A data.frame with posterior means and 95 percent credible intervals, as well as the raw data (Count, Population, time period, grouping variable if any, and crude rates).
- **samples** A stanfit object returned by `sampling`. This contains the MCMC samples from the posterior distribution of the fitted model.
- **cor** Logical value indicating if the model included a correlation structure.
stan_rw

- **time** A list containing the name of the time-period column in the user-provided data and a data.frame of observed time periods and their index.
- **group** If a grouping variable was used, this will be a list containing the name of the grouping variable and a data.frame with group labels and index values.

**Author(s)**
Connor Donegan (Connor.Donegan@UTSouthwestern.edu)

**Source**

**Examples**

```r
library(rstan)
data(msa)
austin <- msa[grep("Austin", msa$MSA), ]

fit <- stan_rw(austin,
    time = Year,
    group = Race,
    iter = 1500)

## MCMC diagnostics
rstan::stan_mcse(fit$samples)
rstan::stan_rhat(fit$samples)
print(fit$samples)

## print the surveil object
print(fit)
head(fit$summary)

## plot time trends
plot(fit, style = 'lines')

## age-specific rates and cumulative percent change
data(cancer)
fit <- stan_rw(cancer, time = Year, group = Age, iter = 2000)
fit_apc <- apc(fit)
plot(fit_apc, cumulative = TRUE)

# age-standardized rates
data(standard)
fit_stands <- standardize(fit, label = standard$age, standard_pop = standard$standard_pop)
```
print(fit_stands)
plot(fit_stands)
fit_stands_apc <- apc(fit_stands)
plot(fit_stands_apc)

---

**surveil_diff**

*Methods for surveil_diff objects*

**Description**

Methods for surveil_diff objects

print surveil_diff objects for analyses of inequality

**Usage**

```r
## S3 method for class 'surveil_diff'
plot(
x,
style = c("mean_qi", "lines"),
M = 250,
col = "black",
fill = "gray80",
lwd,
alpha,
plot = TRUE,
scale = 1e+05,
PAR = TRUE,
base_size = 14,
...)

## S3 method for class 'surveil_diff'
print(x, scale = 1, ...)
```

**Arguments**

- `x`: Object of class surveil_diff, as returned by calling group_diff on a fitted surveil model
- `style`: If `style = "mean_qi"`, then the posterior mean and 95 percent credible interval will be plotted; if `style = "lines"`, then `M` samples from the joint probability distribution of the annual rates will be plotted.
- `M`: If `style = "lines"`, then `M` is the number of samples from the posterior distribution that will be plotted; the default is `M = 250`.
- `col`: Line color
- `fill`: Fill color for credible intervals, passed to geom_ribbon
Theil's entropy-based index of inequality

**Usage**

```r
theil(x)
theil2(Count, Population, rates, total = TRUE)
```

**Arguments**

- `x` A fitted `surveil` model, from `stan_rw`; or, a list of fitted `surveil` models, where each model represents a different geographic area (e.g., states).
- `Count` Case counts, integers
- `Population` Population at risk, integers
rates  If Count is not provided, then rates must be provided (Count = rates * Population).

If total = TRUE, Theil's index will be returned. Each unit contributes to Theil's index; if total = FALSE, all of the elements that sum to Theil's index will be returned.

Details
Theil's index is a good index of inequality in disease and mortality burdens when multiple groups are being considered. It provides a summary measure of inequality across a set of demographic groups that may be tracked over time (and/or space). Also, it is interesting because it is additive, and thus admits of simple decompositions.

The index measures discrepancies between a population's share of the disease burden, \( \omega \), and their share of the population, \( \eta \). A situation of zero inequality would imply that each population's share of cases is equal to its population share, or, \( \omega = \eta \). Each population’s contribution to total inequality is calculated as:

\[
T_i = \omega_i \times \log(\omega_i/\eta_i),
\]

the log-ratio of case-share to population-share, weighted by their share of cases. Theil’s index for all areas is the sum of each area’s \( T_i \):

\[
T = \sum_{i=1}^n T_i.
\]

Theil's T is thus a weighted mean of log-ratios of case shares to population shares, where each log-ratio (which we may describe as a raw inequality score) is weighted by its share of total cases. The index has a minimum of zero and a maximum of \( \log(N) \), where \( N \) is the number of units (e.g., number of states).

Theil’s index, which is based on Shannon’s information theory, can be extended to measure inequality across multiple groups nested within non-overlapping geographies (e.g., states).

Value

theil2:
If total = TRUE (the default), theil2 returns Theil’s index as a numeric value. Else, theil2 returns a vector of values that sum to Theil’s index.

theil.surveil:
A named list with the following elements:

summary  A data.frame summarizing the posterior probability distribution for Theil’s T, including the mean and 95 percent credible interval for each time period

samples  A data.frame with MCMC samples for Theil’s T

theil.list:
A list (also of class theil_list) containing a summary data frame and a tbl_df containing MCMC samples for Theil’s index at each time period.
The summary data frame includes the following columns:

time  time period
Theil mean for Theil’s index; equal to the sum of Theil_between and Theil_within.
Theil_between  The between-areas component to Theil’s inequality index
Theil_within  The within-areas component to Theil’s inequality index

Additional columns contain the upper and lower limits of the 95 percent credible intervals for each component of Theil’s index.
The data frame of samples contains the following columns:
time  Time period indicator
draw  An id for each MCMC sample; note that samples are from the joint distribution
Theil_between  The between-geographies component of Thiel’s index
Theil_within  The within-geographies component of Theil’s index
Theil  Theil’s inequality index (T = Between + Within).

Source


See Also

plot.theil print.theil plot.theil_list

Examples

houston <- msa[grep("Houston", msa$MSA), ]
fit <- stan_rw(houston, time = Year, group = Race, iter = 1200)
theil_dfw <- theil(fit)
plot(theil_dfw)

Count <- c(10, 12, 3, 111)
Pop <- c(1000, 1200, 4000, 9000)
theil2(Count, Pop)
theil2(Count, Pop, total = FALSE)
Description

Widely Application Information Criteria (WAIC) for model comparison

Usage

waic(fit, pointwise = FALSE, digits = 2)

Arguments

- **fit**: An `surveil` object
- **pointwise**: Logical (defaults to FALSE); if `pointwise = TRUE`, a vector of values for each observation will be returned.
- **digits**: Round results to this many digits.

Value

A vector of length 3 with WAIC, a rough measure of the effective number of parameters estimated by the model Eff_pars, and log predictive density Lpd. If `pointwise = TRUE`, results are returned in a `data.frame`.

Source


Examples

data(msa)
austin <- msa[grep("Austin", msa$MSA),]
austin.w <- austin[grep("White", austin$Race),]fit <- stan_rw(austin.w, time = Year, iter = 1200)waic(fit)
Index

* datasets
  cancer, 4
  msa, 6
  standard, 16

apc, 3, 12

Cancer, 4

facet_wrap, 8

geom_line, 8

geom_ribbon, 8, 10, 12, 13

grid.arrange, 11, 22

group_diff, 5

lkj (priors), 14

msa, 6

normal (priors), 14

plot.apc, 4

plot.apc (print.apc), 11

plot.stand_surveil, 17

plot.stand_surveil
  (print.stand_surveil), 12

plot.surveil, 7

plot.surveil_diff, 6

plot.surveil_diff (surveil_diff), 21

plot.theil, 9, 24

plot.theil_list, 24

plot.theil_list (plot.theil), 9

print.apc, 4, 11

print.data.frame, 8, 10, 12, 14

print.stand_surveil, 12, 17

print.surveil (plot.surveil), 7

print.surveil_diff, 6

print.surveil_diff (surveil_diff), 21

print.theil, 24

print.theil (plot.theil), 9

print.theil_list (plot.theil), 9

priors, 14

sampling, 18, 19

scale_color_brewer, 8

stan_rw, 4, 8, 14, 17, 17, 22

standard, 16

standardize, 4, 13, 14, 16

surveil (surveil-package), 2

surveil-package, 2

surveil_diff, 21

theil, 6, 11, 22

theil2 (theil), 22

theme, 8, 12, 14

theme_classic, 12

waic, 25