Package: demogsurv (via r-universe)

August 22, 2024

Title Demographic analysis of DHS and other household surveys

Version 0.2.6

Description This package includes tools for calculating demographic indicators from household survey data. Initially developed for for processing and analysis from Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS). The package provides tools to calculate standard child mortality, adult mortality, and fertility indicators stratified arbitrarily by age group, calendar period, pre-survey time periods, birth cohorts and other survey variables (e.g. residence, region, wealth status, education, etc.). Design-based standard errors and sample correlations are available for all indicators via Taylor linearisation or jackknife.

Depends R (>= 3.2.0),

Imports survival (>= 2.39), survey (>= 2.33), reshape2 (>= 1.4.3), stats

Suggests foreign, rdhs, ggplot2, haven, knitr, rmarkdown, testthat, covr, mgcv

License GPL-3

Encoding UTF-8

LazyData true

VignetteBuilder knitr

RoxygenNote 7.1.2

Repository https://mrc-ide.r-universe.dev

RemoteUrl https://github.com/mrc-ide/demogsurv

RemoteRef master

RemoteSha 0389352e6cdd366f9b1324a0ffe837081c587d86

Contents

.mm_aggr . . . 2
calc_asfr . . . 2
calc_dhs_mx . . . 5
calc_nqx . . . 5
create_tips_data . . . 7
demog_pyears . . . 8
jackknife . . . 10
reshape_sib_data . . . 11
zzbr . . . 12
zzir . . . 12

Index **13**

`.mm_aggr` *Construct a model matrix for aggregating over age groups*

Description

Construct a model matrix for aggregating over age groups

Usage

```
.mm_aggr(mf, agegr)
```

Arguments

`mf`	Model frame for predicted rates
`agegr`	Numeric vector defining ages *in years* for splits.

`calc_asfr` *Calculate age-specific fertility rate (ASFR) and total fertility rate (TFR)*

Description

Calculate age-specific fertility rate (ASFR) and total fertility rate (TFR)

Usage

```
calc_asfr(
  data,
  by = NULL,
  agegr = 3:10 * 5,
  period = NULL,
  cohort = NULL,
  tips = c(0, 3),
  clusters = ~v021,
  strata = ~v024 + v025,
  id = "caseid",
  dob = "v011",
  intv = "v008",
  weight = "v005",
  varmethod = "lin",
  bvars = grep("^b3\\_[0-9]*", names(data), value = TRUE),
  birth_displace = 1e-06,
  origin = 1900,
  scale = 12,
  bhdata = NULL,
  counts = FALSE,
  clustcounts = FALSE
)
```

Arguments

data	A dataset (data.frame), for example a DHS individual recode (IR) dataset.
by	A formula specifying factor variables by which to stratify analysis.
agegr	Numeric vector defining ages *in years* for splits.
period	Numeric vector defining calendar periods to stratify analysis, use NULL for no periods.
cohort	Numeric vector defining birth cohorts to stratify analysis, use NULL for no cohort stratification.
tips	Break points for TIme Preceding Survey.
clusters	Formula or data frame specifying cluster ids from largest level to smallest level, '~0' or '~1' is a formula for no clusters.
strata	Formula or vector specifying strata, use 'NULL' for no strata.
id	Variable name for identifying each individual respondent (character string).
dob	Variable name for date of birth of each individual (character string).
intv	Variable name for interview date (character string).
weight	Formula or vector specifying sampling weights.
varmethod	Method for variance calculation. Currently "lin" for Taylor linearisation or "jk1" for unstratified jackknife, or "jkn", for stratified jackknife.
bvars	Names of variables giving child dates of birth. If bhdata is provided, then length(bvars) must equal 1.

birth_displace	Numeric value to displace multiple births date of birth by. Default is '1e-6'.
origin	Origin year for date arguments. 1900 for CMC inputs.
scale	Scale for dates inputs to calendar years. 12 for CMC inputs.
bhdata	A birth history dataset (data.frame) with child dates of birth in long format, for example a DHS births recode (BR) dataset.
counts	Whether to include counts of births & person-years ('pys') in the returned data.frame. Default is 'FALSE'.
clustcounts	Whether to return additional attributes storing cluster specific counts of births attr(val, 'events_clust'), person-years attr(val, 'pyears_clust') & number of clusters in each strata attr(val, 'strataid'). Only applicable when using jacknife varmethod 'jk1' or 'jkn'. 'strataid' is only included for 'jkn' varmethod. Default is 'FALSE'.

Details

Events and person-years are calculated using normalized weights. Unweighted aggregations may be output by specifying weights=NULL (default) or weights=~1.

The assumption is that all dates in the data are specified in the same format, typically century month code (CMC). The period argument is specified in calendar years (possibly non-integer).

Default values for agegr, period, and tips parameters returns age-specific fertility rates over the three-years preceding the survey, the standard fertility indicator produced in DHS reports.

Value

A data.frame consisting of estimates and standard errors. The full covariance matrix of the estimates can be retrieved by vcov(val).

See Also

demog_pyears()

Examples

```
data(zzir)

## Replicate DHS Table 5.1
## ASFR and TFR in 3 years preceding survey by residence
calc_asfr(zzir, ~1, tips=c(0, 3))
reshape2::dcast(calc_asfr(zzir, ~v025, tips=c(0, 3)), agegr ~ v025, value.var = "asfr")
calc_tfr(zzir, ~v025)
calc_tfr(zzir, ~1)

## Replicate DHS Table 5.2
## TFR by resdience, region, education, and wealth quintile
calc_tfr(zzir, ~v102)  # residence
calc_tfr(zzir, ~v101)  # region
calc_tfr(zzir, ~v106)  # education
calc_tfr(zzir, ~v190)  # wealth
calc_tfr(zzir)         # total
```

```
## Calculate annual TFR estimates for 10 years preceding survey
tfr_ann <- calc_tfr(zzir, tips=0:9)

## Sample covariance of annual TFR estimates arising from complex survey design
cov2cor(vcov(tfr_ann))

## Alternately, calculate TFR estimates by calendar year
tfr_cal <- calc_tfr(zzir, period = 2004:2015, tips=NULL)
tfr_cal

## sample covariance of annual TFR estimates arising from complex survey design
## Generate estimates split by period and TIPS
cov2cor(vcov(tfr_cal))

calc_tfr(zzir, period = c(2010, 2013, 2015), tips=0:5)

## ASFR estimates by birth cohort
asfr_coh <- calc_asfr(zzir, cohort=c(1980, 1985, 1990, 1995), tips=NULL)
reshape2::dcast(asfr_coh, agegr ~ cohort, value.var = "asfr")
```

`calc_dhs_mx` *Calculate age-specific mortality rates in period preceding survey.*

Description

Should replicate mortality rates reported in DHS reports.

Usage

```
calc_dhs_mx(sib, period = c(0, 84))
```

Arguments

`sib`	A dataset as 'data.frame'.
`period`	Interval 'period' defined in the months before the survey.

`calc_nqx` *Calculate the probability of dying between age x and x+n (nqx)*

Description

Default arguments are configured to calculate under 5 mortality from a DHS Births Recode file.

Usage

```
calc_nqx(
  data,
  by = NULL,
  agegr = c(0, 1, 3, 5, 12, 24, 36, 48, 60)/12,
  period = NULL,
  cohort = NULL,
  tips = c(0, 5, 10, 15),
  clusters = ~v021,
  strata = ~v024 + v025,
  weight = "v005",
  dob = "b3",
  dod = "dod",
  death = "death",
  intv = "v008",
  varmethod = "lin",
  origin = 1900,
  scale = 12
)
```

Arguments

data	A dataset (data.frame), for example a DHS births recode (BR) dataset.
by	A formula specifying factor variables by which to stratify analysis.
agegr	Numeric vector defining ages *in years* for splits.
period	Numeric vector defining calendar periods to stratify analysis, use NULL for no periods.
cohort	Numeric vector defining birth cohorts to stratify analysis, use NULL for no cohort stratification.
tips	Break points for TIme Preceding Survey.
clusters	Formula or data frame specifying cluster ids from largest level to smallest level, '~0' or '~1' is a formula for no clusters.
strata	Formula or vector specifying strata, use 'NULL' for no strata.
weight	Formula or vector specifying sampling weights.
dob	Variable name for date of birth (character string).
dod	Variable name for date of death (character string).
death	Variable name for event variable (character string).
intv	Variable name for interview date (character string).
varmethod	Method for variance calculation. Currently "lin" for Taylor linearisation or "jk1" for unstratified jackknife, or "jkn", for stratified jackknife.
origin	Origin year for date arguments. 1900 for CMC inputs.
scale	Scale for dates inputs to calendar years. 12 for CMC inputs.

Examples

```
data(zzbr)
zzbr$death <- zzbr$b5 == "no"  # b5: child still alive ("yes"/"no")
zzbr$dod <- zzbr$b3 + zzbr$b7 + 0.5

## Calculate 5q0 from birth history dataset.
## Note this does NOT exactly match DHS calculation.
## See calc_dhs_u5mr().
u5mr <- calc_nqx(zzbr)
u5mr

## Retrieve sample covariance and correlation
vcov(u5mr)  # sample covariance
cov2cor(vcov(u5mr))  # sample correlation

## 5q0 by sociodemographic characteristics
calc_nqx(zzbr, by=~v102) # by urban/rural residence
calc_nqx(zzbr, by=~v190, tips=c(0, 10)) # by wealth quintile, 0-9 years before
calc_nqx(zzbr, by=~v101+v102, tips=c(0, 10)) # by region and residence

## Compare unstratified standard error estiamtes for linearization and jackknife
calc_nqx(zzbr, varmethod = "lin")  # unstratified design
calc_nqx(zzbr, strata=NULL, varmethod = "lin")  # unstratified design
calc_nqx(zzbr, strata=NULL, varmethod = "jk1")  # unstratififed jackknife
calc_nqx(zzbr, varmethod = "jkn")  # stratififed jackknife

## Calculate various child mortality indicators (neonatal, infant, etc.)
calc_nqx(zzbr, agegr=c(0, 1)/12)  # neonatal
calc_nqx(zzbr, agegr=c(1, 3, 5, 12)/12) # postneonatal
calc_nqx(zzbr, agegr=c(0, 1, 3, 5, 12)/12) # infant (1q0)
calc_nqx(zzbr, agegr=c(12, 24, 36, 48, 60)/12) # child (4q1)
calc_nqx(zzbr, agegr=c(0, 1, 3, 5, 12, 24, 36, 48, 60)/12) # u5mr (5q0)

## Calculate annaul 5q0 by calendar year
calc_nqx(zzbr, period=2005:2015, tips=NULL)
```

`create_tips_data` *Create episode dataset split by period, age group, and time preceding survey indicator (TIPS)*

Description

Create episode dataset split by period, age group, and time preceding survey indicator (TIPS)

Usage

```
create_tips_data(
  dat,
```

```
  period = do.call(seq.int, as.list(range(dat$intvy) + c(-16, 1))),
  agegr = 3:12 * 5,
  tips = 0:15,
  dobvar = "sibdob",
  dodvar = "sibdod"
)
```

Arguments

dat	A dataset as 'data.frame'.
period	Numeric vector defining calendar periods to stratify analysis, use 'NULL' for no periods.
agegr	Numeric vector defining ages *in years* for splits.
tips	Break points for TIme Preceding Survey.
dobvar	Variable name for date of birth (character string).
dodvar	Variable name for date of death (character string).

`demog_pyears` *Events and person-years from episode data for demographic analysis*

Description

This is a wrapper for the `pyears` function in the `survival` package with convenient stratifications for demographic analyses.

Usage

```
demog_pyears(
  formula,
  data,
  period = NULL,
  agegr = NULL,
  cohort = NULL,
  tips = NULL,
  origin = 1900,
  scale = 12,
  dob = "(dob)",
  intv = "(intv)",
  tstart = "tstart",
  tstop = "tstop",
  event = "event",
  weights = NULL
)
```

Arguments

formula	a formula object. The response variable will be a vector of follow-up times for each subject, or a Surv object containing the survival time and an event indicator. The predictors consist of optional grouping variables separated by + operators (exactly as in survfit), time-dependent grouping variables such as age (specified with tcut), and optionally a ratetable term. This latter matches each subject to his/her expected cohort.
data	a data frame in which to interpret the variables named in the formula, or in the subset and the weights argument.
period	Numeric vector defining calendar periods to stratify analysis, use NULL for no periods.
agegr	Numeric vector defining ages *in years* for splits.
cohort	Numeric vector defining birth cohorts to stratify analysis, use NULL for no cohort stratification.
tips	Break points for TIme Preceding Survey.
origin	Origin year for date arguments. 1900 for CMC inputs.
scale	a scaling for the results. As most rate tables are in units/day, the default value of 365.25 causes the output to be reported in years.
dob	Variable name for date of birth (character string).
intv	Variable name for interview date (character string).
tstart	Variable name for the start of follow up time, example is date of birth. Default is 'tstart'.
tstop	Variable name for the end of follow up time, examples include interview date or date of death. Default is 'tend'.
event	Variable name for the event indicator, example is birth or death. Default is 'event'.
weights	case weights.

Details

Note that event must be a binary variable per the internals of the pyears() function. The function could be updated to work around this stipulation.

See Also

pyears, tcut

jackknife *Jackknife covariance calculation*

Description

Calculate the covariance matrix for a vector of estimates of the form `fn(L * x/n)` using unstratified (JK1) or stratified (JKn) jackknife calculation removing a single cluster at a time. The calculation assumes infinite population sampling.

Usage

```
jackknife(x, n, strataid = NULL, L = diag(nrow(x)), fn = function(x) x)
```

Arguments

`x`	v x k matrix specifying weighted numerator for each of k PSUs (across columns)
`n`	v x k matrix specifying weighted denominator for each PSU (across columns)
`strataid`	integer or factor vector consisting of id for each strata. Optional, length should be number of columns of x if supplied.
`L`	q x v matrix defining a linear transform
`fn`	function to transorm ratio x/n.

Details

If `strataid` is provided, then the stratified (JKn) covariance is calculated, while if `strataid = NULL` then the unstratified (JK1) covariance is calculated. The latter corresponds to the unstratified jackknife covariance reported in DHS survey reports. The calculations are equivalent for `strataid = rep(1, ncol(x))`.

Value

a data frame with `q` rows consisting of estimates calculated as `fn(L * rowSums(x) / rowSums(n))`, standard error, and 95% CIs calculated on the untransformed scale and then transformed. The covariance matrix is returned as the `"var"` attribute and can be accessed by `vcov(val)`.

References

Pedersen J, Liu J (2012) Child Mortality Estimation: Appropriate Time Periods for Child Mortality Estimates from Full Birth Histories. PLoS Med 9(8): e1001289. https://doi.org/10.1371/journal.pmed.1001289.

reshape_sib_data *Convert respondent-level sibling history data to one row per sibling*

Description

Convert respondent-level sibling history data to one row per sibling

Usage

```
reshape_sib_data(
  data,
  widevars = grep("^v", names(data), value = TRUE),
  longvars = grep(sibvar_regex, names(data), value = TRUE),
  idvar = "caseid",
  sib_vars = sub("(.*)_.*", "\\1", longvars),
  sib_idvar = "mmidx",
  sibvar_regex = "^mm[idx0-9]"
)
```

Arguments

data	A dataset as data.frame.
widevars	Character vector of respondent-level variable names to include.
longvars	Character vector of variables corresponding to each sibling.
idvar	Vector of variable names uniquely identifying each respondent.
sib_vars	Vector of same length as longvars giving variable names in long dataset.
sib_idvar	Variable name uniquely identifying each sibling record. Should appear amongst sib_vars.
sibvar_regex	Optionally, a regular expression to identify variable names for longvars from names of data.

Examples

```
data(zzir)

zzsib <- reshape_sib_data(zzir)
zzsib$death <- factor(zzsib$mm2, c("dead", "alive")) == "dead"
zzsib$sex <- factor(zzsib$mm1, c("female", "male"))  # drop mm2 = 3: "missing"
calc_nqx(zzsib, by=~sex, agegr=seq(15, 50, 5), tips=c(0, 7), dob="mm4", dod="mm8")
```

zzbr *DHS Model Births Recode Dataset*

Description

An example DHS births recode dataset with each row representing a child ever born to individuals eligible for the women's questionnaire. This data is not from any actual DHS survey.

Usage

```
zzbr
```

Format

A data frame with 23,666 rows and 105 variables. To keep the model dataset small only variables starting with 'caseid', 'v0', 'v1', or 'b', have been included.

Source

https://dhsprogram.com/data/Download-Model-Datasets.cfm

zzir *DHS Model Individual Recode Dataset*

Description

An example DHS individual recode dataset with each row representing an individual eligible for the women's questionnaire. This data is not from any actual DHS survey.

Usage

```
zzir
```

Format

A data frame with 8,348 rows and 819 variables. To keep the model dataset small only variables starting with 'caseid', 'v0', 'v1', 'v2', 'b', or 'mm' have been included.

Source

https://dhsprogram.com/data/Download-Model-Datasets.cfm

Index

∗ **datasets**
    zzbr, 12
    zzir, 12
.mm_aggr, 2

calc_asfr, 2
calc_dhs_mx, 5
calc_nqx, 5
calc_tfr *(*calc_asfr*)*, 2
create_tips_data, 7

demog_pyears, 8
demog_pyears(), *4*

jackknife, 10

pyears, *8*, *9*

reshape_sib_data, 11

tcut, *9*

zzbr, 12
zzir, 12