Package 'mdhs'

February 6, 2020

Type Package

Title Compute Death Probability Statistics from DHS Sibling Histories

Date January 20, 2020

Version 0.1.0

Author Kene David Nwosu

Maintainer Kene David Nwosu <kendavidn@gmail.com>

Description Provides functions to extract sibling survival data from DHS individual recode files and compute death probability statistics based on these data.
The user can define specific cohorts, ages and years for which they would like to obtain death probabilities.
For example, a user can answer the question 'what was the early adult death probability (death between 20 and 35) for cohorts between 1970 and 1980 in the Northern District of Nigeria?'

License GPL-2

Encoding UTF-8

LazyData true

Depends R (>= 3.5.0)

Imports stringr (>= 1.4.0),
reshape2 (>= 1.4.3),
dplyr (>= 0.8.3)

RoxygenNote 7.0.2

R topics documented:

death_prob_cohort *Compute Death Probabilities for a Specified Cohort at a Specified Age*

Description

Calculates the probability of death for a user-specified cohort and a user-specified age. The formula used can be found at tinyurl.com/demograsource.

For example, the question 'what was the probability of death at age 1 for the cohort born in 1995?' can be answered with this function. This would be computed as the number of people in that cohort who died at age 1 divided by the number that reached their 1st birthday.

The function requires a data frame that contains variables for:

- date of birth in century month code

- survival status

- survival time, as can be generated by the mdhs_obs_time function in this package.

Usage

```
death_prob_cohort(
  data,
  age,
  cohort,
  surv_stat = "mm2",
  DOB_CMC = "mm4",
  obs_time = "obs_time"
)
```

Arguments

data	The relevant data frame
age	The age for which the statistic is desired
cohort	The cohort for which the statistic is desired
surv_stat	String name for the column containing survival status. Requires that deceased is coded as 0, and alive is coded as 1. Defaults to 'mm2', the original DHS variable name.
DOB_CMC	String name for column containing dates of birth in century month code. Defaults to 'mm4', the original DHS variable name.
obs_time	String name for the column containing observation times, as can be generated by the mdhs_obs_time function. Defaults to 'obs_time'

Value

A single numerical value

See Also

death_prob_cohort_range

Examples

```
# compute probability of death at age 0 for the cohort born in 2000
death_prob_cohort(malawisib, 0, 2000)

# same example as above, with each parameter more explicitly defined
death_prob_cohort(data = malawisib, age =0, cohort = 2000,
DOB_CMC = 'mm4', surv_stat = 'mm2', obs_time = 'obs_time')
```

`death_prob_cohort_range` *Compute Death Probabilities for a Specified Cohort Range over a Specified Age Range*

Description

This function salculates the probability of death for a user-specified cohort and age range. It simply iterates the 'death_prob_cohort' function over the range of values provided by the user.

For example, the question 'what was the probability of death during teenage (13-19) years for the cohorts born between 1995 and 1997?' can be answered with this function. This would be an averaged value of:

-The death probability at age 13 for the 1995 cohort (The number of people in the 1995 cohort cohort who died at age 13 divided by the number that reached their 1st birthday.)

-The death probability at age 14 for the 1995 cohort

-...

-The death probability at age 19 for the 1995 cohort

-The death probability at age 13 for the 1996 cohort

-The death probability at age 14 for the 1996 cohort

-...

-The death probability at age 19 for the 1996 cohort

Requires a data frame that contains columns for date of birth in century month code, survival status and observation time, as can be generated by the mdhs_obs_time function in this package.

Usage

```
death_prob_cohort_range(
  data,
  lower_age,
  upper_age,
  lower_cohort,
  upper_cohort,
  surv_stat = "mm2",
  DOB_CMC = "mm4",
  obs_time = "obs_time"
)
```

Arguments

data	The data frame containing sibling histories
lower_age	The lower boundary of the age interval for which the statistic is desired
upper_age	The upper boundary of the age interval for which the statistic is desired
lower_cohort	The lower boundary of the cohort interval for which the statistic is desired
upper_cohort	The upper boundary of the cohort interval for which the statistic is desired
surv_stat	String name for the column containing survival status. Requires that deceased is coded as 0, and alive is coded as 1. Defaults to 'mm2', the original DHS variable name.
DOB_CMC	String name for column containing dates of birth in century month code. Defaults to 'mm4', the original DHS variable name.
obs_time	String name for the column containing observation times, as generated by the mdhs_obs_time function. Defaults to 'obs_time'

Value

A dataframe with death probabilities. Ages are arranged along the columns, while cohorts(birth years) are arranged along the rows.

See Also

death_prob_cohort

Examples

```
# compute probability of death between ages 0 and 5 for cohorts born between 2000 and 2005
death_prob_cohort_range(malawisib, 0,5,2000,2005)

# same example as above, with each parameter more explicitly defined
death_prob_cohort_range(data = malawisib, lower_age = 0, upper_age = 5,
lower_cohort = 2000, upper_cohort = 2005,
DOB_CMC = 'mm4', surv_stat = 'mm2', obs_time = 'obs_time')
```

---

death_prob_year *Compute Death Probabilities for a Specified Year at a Specified Age*

Description

Calculates the probability of death for a user-specified year and a user-specified age.

For example, the question 'what was the probability of death at age 1 in the year 1995?' can be answered with this function. The computation of this not straightforward, since one individual may be aged 1 for a single day in 1995, while another may be aged 1 for the entire year of 1995. The formula used here is based on that found in Chapter 2 of Demography: Measuring and Modeling Population Processes by Preston et.al (tinyurl.com/demograsource).

Requires a data frame that contains the variables date of birth in century month code, survival status and observation time, as generated by the mdhs_obs_time function in this package.

Usage

```
death_prob_year(
  data,
  age,
  year,
  surv_stat = "mm2",
  DOB_CMC = "mm4",
  obs_time = "obs_time"
)
```

Arguments

data	The relevant data frame
age	The age for which the statistic is desired
year	The year for which the statistic is desired
surv_stat	String name for the column containing survival status. Requires that deceased is coded as 0, and alive is coded as 1. Defaults to 'mm2', the original DHS variable name.
DOB_CMC	String name for column containing dates of birth in century month code. Defaults to 'mm4', the original DHS variable name.
obs_time	String name for the column containing observation times, as can be generated by the mdhs_obs_time function. Defaults to 'obs_time'

Value

A single numerical value

See Also

death_prob_year_range

Examples

```
# compute probability of death at age 0 in the year 2000
death_prob_year(malawisib, 0, 2000)

# same example as above, with each param more explicitly defined
death_prob_year(data = malawisib, age = 0, year = 2000,
DOB_CMC = 'mm4', surv_stat = 'mm2', obs_time = 'obs_time')
```

---

death_prob_year_range *Compute Death Probabilities for a Specified Year Range and a Specified Age Range*

Description

Calculates the probability of death for a user-specified year and age range

For example, the question 'what was the probability of death for teenagers (13-19) in the time between 1995 and the end of 1996?' can be answered with this function. This would be an averaged value of:

-The death probability at age 13 in 1995 (as computed by the death_prob_year function)

-The death probability at age 14 in 1995

-...

-The death probability at age 19 in 1995; and then

-The death probability at age 13 in 1996

-The death probability at age 14 in 1996

-...

-The death probability at age 19 in 1996

Requires a data frame that contains the variables date of birth in century month code, survival status and observation time, as can be generated by the mdhs_obs_time function in this package.

Usage

```
death_prob_year_range(
  data,
  lower_age,
  upper_age,
  lower_year,
  upper_year,
  surv_stat = "mm2",
  DOB_CMC = "mm4",
  obs_time = "obs_time"
)
```

Arguments

data	The relevant data frame
lower_age	The lower boundary of the age interval for which the statistic is desired
upper_age	The upper boundary of the age interval for which the statistic is desired
lower_year	The lower boundary of the year interval for which the statistic is desired
upper_year	The upper boundary of the year interval for which the statistic is desired
surv_stat	String name for the column containing survival status. Requires that deceased is coded as 0, and alive is coded as 1. Defaults to 'mm2', the original DHS variable name.
DOB_CMC	String name for column containing dates of birth in century month code. Defaults to 'mm4', the original DHS variable name.
obs_time	String name for the column containing observation times, as generated by the mdhs_obs_time function. Defaults to 'obs_time'

Value

A dataframe with death probabilities. Ages are arranged along the columns, while years are arranged along the rows.

See Also

death_prob_year

Examples

```
# compute probability of death between ages 0 and 5 for cohorts born between 2000 and 2005
death_prob_year_range(malawisib, 0,5,2000,2005)

# same example as above, with each parameter more explicitly defined
death_prob_year_range(data = malawisib, lower_age = 0, upper_age = 5,
lower_year = 2000, upper_year = 2005,
DOB_CMC = 'mm4', surv_stat = 'mm2', obs_time = 'obs_time')
```

---

malawidhs *Malawi DHS Individual Recode Files from Four Surveys*

Description

Contains data from 77,349 respondents from the Malawi 2004 through 2015 DHS surveys.

The dataset contains the following variables, needed to compute mortality statistics:

caseid: Unique identifier for each respondent

v008: Date of interview in century month code

The following sibling variables, with 20 subvariables (mm1_01, mm1_02...mm1_20) for each of a maximum of 20 possible siblings per respondent:

mm1: Sex of sibling (1 = male, 2 = female)

mm2: Survival status of sibling (0 = deceased, 1 = alive)

mm4: Sibling's date of birth in century month code

mm8: Sibling's date of death in century month code

Usage

```
malawidhs
```

Format

dataframe

`malawisib` *Sibling Data Reshaped from 'malawidhs' dataframe*

Description

Contains data for the 436,371 siblings of individuals in the 'malawidhs' data frame that is also included in this package.

The dataset contains the following variables:

caseid: Unique identifier for each respondent

sib_id: Unique identifier for each sibling of each respondent

v008: Date of interview in century month code

mm1: Sex of sibling (1 = male, 2 = female)

mm2: Survival status of sibling (0 = deceased, 1 = alive)

mm4: Siblings' dates of birth in century month code

Usage

```
malawisib
```

Format

dataframe

`mdhs_merge` *Merge a Sibling Recode File with Select Variables from the DHS Individual Recode*

Description

This is simply a wrapper for the 'merge' function in base R with handy defaults. It allows the user to easily select which of the variables from the DHS individual recode should be included.

Usage

```
mdhs_merge(sib_df, respondent_df, respondent_vars = c("v008"), id = "caseid")
```

Arguments

`sib_df`	The DHS sibling recode file, as generated with the mdhs_reshape function
`respondent_df`	The corresponding DHS individual respondent recode dataframe
`respondent_vars`	String names of the columns in the individual recode file to be merged. Defaults to 'v008', as this is the only additional variable needed to calculate sibling mortality statistics.
`id`	The name of the ID variable in the dataset, entered as a string. Defaults to 'caseid'.

Value

Returns the merged dataframe

Examples

```
# first generate the sibling recode dataframe
sib_recode <- mdhs_reshape(data = malawidhs, sib_cols = c(1:5))
# then merge it with the individual recode file
merged_file <- mdhs_merge(sib_df = sib_recode, respondent_df = malawidhs)
```

---

mdhs_obs_time *Compute Observation Time for Survival Analysis using DHS Century Month Codes*

Description

Generates a 'Time Under Observation' variable from DHS data, accounting for right censoring. For individuals who are alive at the time of the survey, observation time is their age, in months, at the start of the survey. For individuals who are deceased at the time of the survey, observation time is their age, in months, at the time of death.

Usage

```
mdhs_obs_time(
  data,
  surv_stat = "mm2",
  DOB_CMC = "mm4",
  DOD_CMC = "mm8",
  interview_date_CMC = "v008"
)
```

Arguments

data	The relevant data frame
surv_stat	String name for the column containing survival status. Requires that deceased is coded as 0, and alive is coded as 1. Defaults to 'mm2' as in the DHS Individual Recode File.
DOB_CMC	String name for the column containing dates of birth in century month code. Defaults to 'mm4', which is the corresponding variable name in the DHS individual recode files.
DOD_CMC	String name for the column containing dates of death in century month code. Defaults to 'mm8', which is the corresponding variable name in the DHS individual recode files.
interview_date_CMC	String name for the column containing dates of interview in century month code. Defaults to 'v008', which is the corresponding variable name in the DHS individual recode files.

Value

A vector of observation times

Examples

```
# first generate the sibling recode dataframe
sib_recode <- mdhs_reshape(data = rwandadhs, sib_cols = c(1,2,4,8))
# then merge it with the individual recode file
merged_file <- mdhs_merge(sib_df = sib_recode, respondent_df = rwandadhs)
# then create an observation time column in the merged data frame
merged_file$obs_time <-
   mdhs_obs_time(merged_file, surv_stat = 'mm2' , DOB_CMC = 'mm4', DOD_CMC = 'mm8',
   interview_date_CMC = 'v008' )
```

`mdhs_reshape` *Reshape a DHS individual recode file into a one-sibling-per-row file*

Description

This function takes a DHS individual recode file, with one respondent per row, and transforms it into a sibling recode file, with one sibling per row. The user can choose to indicate which of the 15 sibling variables should be included in the reshaped file.

Usage

```
mdhs_reshape(data, sib_cols = c(1, 2, 4, 8), id = "caseid")
```

Arguments

`data`	The DHS individual recode file.
`sib_cols`	A vector of numbers, e.g.c(1:8), to indicate which sibling variables to reshape. Defaults to c(1,2,4,8), because "mm1", "mm2", "mm4" and "mm8" are the DHS variables needed for the mortality analysis functions in this package.
`id`	The string name of the ID variable in the dataset, entered as a string. Defaults to 'caseid'.

Value

Returns a dataframe with one sibling per row

Examples

```
mdhs_reshape(data = malawisib, sib_cols = c(1,2,4,8))
```

prep_for_lexis *Prepare Output from death_prob Function for Lexis Plot*

Description

The function reshapes and rearranges the input data frame into a 'long' format, ready to be turned into a ggplot geom_tile visualization.

Usage

```
prep_for_lexis(data)
```

Arguments

data	The relevant data frame

rwandadhs *Rwanda DHS Individual Recode Files from Four Surveys*

Description

Contains data from 23,918 respondents from the Rwanda 2004 through 2014 DHS surveys.

The dataset contains the following variables, needed to compute mortality statistics:

caseid: Unique identifier for each respondent

v008: Date of interview in century month code

The following sibling variables, with 20 subvariables (mm1_01, mm1_02...mm1_20) for each of a maximum of 20 possible siblings per respondent:

mm1: Sex of sibling (1 = male, 2 = female)

mm2: Survival status of sibling (0 = deceased, 1 = alive)

mm4: Sibling's date of birth in century month code

mm8: Sibling's date of death in century month code

Usage

```
rwandadhs
```

Format

dataframe

rwandasib *Sibling Data Reshaped from 'rwandadhs' dataframe*

Description

Contains data for the 144,612 siblings of individuals in the 'rwandadhs' data frame that is also included in this package.

The dataset contains the following variables:

caseid: Unique identifier for each respondent

sib_id: Unique identifier for each sibling of each respondent

v008: Date of interview in century month code

mm1: Sex of sibling (1 = male, 2 = female)

mm2: Survival status of sibling (0 = deceased, 1 = alive)

mm4: Siblings' dates of birth in century month code

Usage

```
rwandasib
```

Format

dataframe

Index