Package: gompertztrunc (via r-universe)

August 26, 2024

Type Package

Title Conducting Maximum Likelihood Estimation with Truncated Mortality Data

Version 0.1.2

Maintainer Maria Osborne <mariaosborne@berkeley.edu>

Description Estimates hazard ratios and mortality differentials for doubly-truncated data without population denominators. This method is described in Goldstein et al. (2023) <doi:10.1007/s11113-023-09785-z>.

License GPL (>= 3)

URL https://caseybreen.github.io/gompertztrunc/, https://github.com/caseybreen/gompertztrunc

BugReports https://github.com/caseybreen/gompertztrunc/issues

Depends R (>= 3.5.0)

Imports broom, cowplot, data.table, dplyr, flexsurv, ggplot2, ggsci, grid, magrittr, modelr, rlang, stats, stringr, tibble, tidyr

Suggests knitr, rmarkdown, socviz, tidyverse

VignetteBuilder knitr

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.1

Repository https://caseybreen.r-universe.dev

RemoteUrl https://github.com/caseybreen/gompertztrunc

RemoteRef HEAD

RemoteSha 21d43e207df7222e46366c06babe542aa4e93c00

Contents

---

bunmd_demo *Demo BUNMD Data Set*

Description

A data set containing a sample of the CenSoc Berkeley Unified Numident Mortality Database (BUNMD) file, including age at death and select covariates.

Usage

```
bunmd_demo
```

Format

A data frame with 81,002 rows and 6 variables:

ssn Social Security number

bpl_string Country of birth

death_age Age at death (integer years)

byear Calendar year of birth

dyear Calendar year of death

age_first_application Age at first Social Security application

Details

The Berkeley Unified Numident Mortality Database (BUNMD) is a cleaned and harmonized version of the NARA Numident file, consisting of the most informative parts of the 60+ application, claim, and death files released by the National Archives.The full data set of nearly 50 million records is available at https://censoc.berkeley.edu/data/.

Source

Joshua R. Goldstein, Monica Alexander, Casey Breen, Andrea Miranda González, Felipe Menares, Maria Osborne, Mallika Snyder, and Ugur Yildirim. CenSoc Mortality File: Version 2.0. Berkeley: University of California, 2021. https://censoc.berkeley.edu/

`convert_hazards_to_ex` *Convert hazard ratios to life expectancy*

Description

Convert hazard ratios to differences in remaining life expectancy at a given age (defaults to age 65)

Usage

```
convert_hazards_to_ex(
  df,
  age = 65,
  upper_age = 120,
  M = 80,
  b = 0.075,
  use_model_estimates = FALSE
)
```

Arguments

`df`	Dataframe of results given by gompertz_mle() function
`age`	Age at which to calculate remaining life expectancy
`upper_age`	Maximal age to use in life table calculation
`M`	Gompertz parameter modal age at death
`b`	Gompertz mortality slope parameter
`use_model_estimates`	Use estimates of the Gompertz Parameters from the model, rather than defaults

Value

A dataframe of hazards ratios and corresponding e(x) estimates and confidence intervals

Examples

```
#model hazards as function of birthplace using bunmd_demo data
demo_dataset <- dplyr::filter(bunmd_demo, bpl_string %in% c("Poland", "England")) %>%
dplyr::sample_frac(0.1)

#run gompertz_mle()
bpl <- gompertz_mle(formula = death_age ~ bpl_string, left_trunc = 1988,
                    right_trunc = 2005, data = demo_dataset)
```

```
#convert to difference in life expectancy
convert_hazards_to_ex(df = bpl$results, use_model_estimates = FALSE)
```

`diagnostic_plot` *Create diagnostic plots*

Description

Compare empirical and modeled distribution of ages of death within a cohort. Only works with a single discrete covariate and a single cohort.

Usage

```
diagnostic_plot(
  data,
  object,
  covar,
  death_var = "death_age",
  byear_var = "byear",
  xlim = c(65, 110)
)
```

Arguments

`data`	data used to create gompertz_mle object
`object`	gompertz_mle object
`covar`	covariate of interest
`death_var`	death age variable
`byear_var`	birth year/cohort variable
`xlim`	x-limits for figure

Value

a ggplot object

Examples

```
# Create a single-cohort data set
numident_c1920 <- numident_demo %>% dplyr::filter(byear == 1920) %>%
dplyr::mutate(finished_hs = as.factor(educ_yrs >= 12))

# Run gompertz_mle()
gradient <- gompertztrunc::gompertz_mle(formula = death_age ~ finished_hs,
left_trunc = 1988, right_trunc = 2005, data = numident_c1920)
```

```
# Create diagnostic histogram plot using model outcome
gompertztrunc::diagnostic_plot(object = gradient, data = numident_c1920,
covar = "finished_hs", xlim = c(60, 95))
```

`diagnostic_plot_hazard` *Create diagnostic plot (hazard scale)*

Description

Compare empirical and model-based estimated hazard rates within a cohort. Only works with a single discrete covariate and a single cohort. Will plot hazards for to 9 levels/values of the discrete covariate.

Usage

```
diagnostic_plot_hazard(
  data,
  object,
  covar,
  death_var = "death_age",
  byear_var = "byear",
  xlim = c(65, 110)
)
```

Arguments

`data`	data.frame of observed data for gompertz_mle
`object`	gompertz_mle object
`covar`	covariate of interest
`death_var`	death age variable
`byear_var`	birth year/cohort variable
`xlim`	x-limits for figure

Details

This function assumes that no population denominators exist with which to calculate hazards. Therefore, the "observed" hazards produced are not truly empirical values. Instead, it relies partially on the modeled parameters to compute life table values.

To find these quasi-observed hazards, the modeled Gompertz distribution is used to calculate *l(x_min)*; i.e., the number of survivors to the earliest observable age at death in the data. This is done for each category/level of the specified covariate. Then, the number of observed deaths at each age is used to infer the number of survivors to each subsequent age and the death rate at each age.

Value

a ggplot object

Examples

```
# Create a single-cohort data set
numident_c1920 <- numident_demo %>% dplyr::filter(byear == 1920) %>%
dplyr::mutate(finished_hs = as.factor(educ_yrs >= 12))

# Run gompertz_mle()
gradient <- gompertztrunc::gompertz_mle(formula = death_age ~ finished_hs,
left_trunc = 1988, right_trunc = 2005, data = numident_c1920)

# Create diagnostic hazards plot using model outcome
gompertztrunc::diagnostic_plot_hazard(object = gradient, data = numident_c1920,
covar = "finished_hs", xlim = c(60, 95))
```

---

get.par.start *Get starting values for parameters*

Description

Uses linear modeling to compute initial values for MLE optimizer

Usage

```
get.par.start(formula, data)
```

Arguments

formula	the estimation formula
data	data matrix with y, u, l, and covariates, including cohort

Value

Named vector of initial parameter estimates

`gompertztrunc_simu` *Simulate Gompertzian death distribution*

Description

Simulate Gompertzian death distribution

Usage

```
gompertztrunc_simu(
  n,
  formula,
  coefs,
  dummy = NULL,
  sigma = NULL,
  seed = NULL,
  a0 = 10^-4,
  b = 1/10,
  verbose = FALSE
)
```

Arguments

`n`	sample size
`formula`	estimation formula
`coefs`	named vectors of coefficients and corresponding true values
`dummy`	vector flags for each coefficient
`sigma`	standard deviation for each variable
`seed`	random seed to duplicate data
`a0`	Gompertz alpha parameter
`b`	Gompertz b parameter
`verbose`	print internal check if true

Value

dataframe of simulated death ages and covariate values

Examples

```
gompertztrunc_simu(n=1000, formula = death_age ~ sex + ambient_temp,
coefs = c('sex'=-0.8, 'ambient_temp'=0.3), dummy=c(TRUE,FALSE))
```

gompertz_mle *Gompertz MLE function*

Description

Fits a Gompertz distribution with proportional hazards to doubly-truncated mortality data using maximum likelihood estimation.

Usage

```
gompertz_mle(
  formula,
  left_trunc = 1975,
  right_trunc = 2005,
  data,
  byear = byear,
  dyear = dyear,
  lower_age_bound = NULL,
  upper_age_bound = NULL,
  weights = NULL,
  start = NULL,
  death_age_data_type = "auto",
  maxiter = 10000
)
```

Arguments

formula	the estimation formula
left_trunc	left truncation year
right_trunc	right truncation year
data	a data frame containing variables in the model
byear	vector of birth years
dyear	vector of death years
lower_age_bound	lowest age at death to include (optional)
upper_age_bound	highest age at death to include (optional)
weights	an optional vector of individual weights
start	an optional vector of starting values for the optimizer. must be a numeric vector that exactly matches the output of `get.par.start(formula, data)` in length and element names.
death_age_data_type	option for handling of continuous and discrete death age variable (not yet implemented)
maxiter	maximum number of iterations for optimizer

Value

Returns a named list consisting of the following components (See `stats::optim()` for additional details):

`starting_values` list of starting values of parameters

`optim_fit` A list consisting of:

- `par` best estimation of parameter values
- `value` log likelihood
- `counts` number of calls to function and gradient
- `convergence` returns 0 if the model converged, for other values see `stats::optim()`
- `message` any other information returned by optimizer
- `hessian` Hessian matrix

`results` A table of estimates and upper/lower bounds of the 95 percent confidence interval for the estimates. Confidence interval computed as 1.96*standard_error.

Examples

```
## Not run:
#model hazards as function of birthplace using bunmd_demo file
gompertz_mle(formula = death_age ~ bpl_string, left_trunc = 1988, right_trunc = 2005,
data = bunmd_demo)

## End(Not run)
```

---

hazard_ratio_to_le *Translate a single hazard ratio to remaining life expectancy*

Description

Translate a single hazard ratio to effect on remaining life expectancy at a specified age, using a Gompertz mortality schedule as the baseline

Usage

```
hazard_ratio_to_le(lower, upper, hr, M = 80, b = 0.1)
```

Arguments

`lower`	age at which to compute change in remaining life expectancy
`upper`	upper age bound for life table calculations
`hr`	hazard ratio
`M`	Gompertz modal age at death parameter
`b`	Gompertz mortality slope parameter

Value

hazard ratio converted to effect on life expectancy

negLL_function *Gompertz Negative Log Likelihood Function*

Description

Computes negative log likelihood for optimizer

Usage

```
negLL_function(par, y, X, y.left, y.right, wt)
```

Arguments

par	a vector of parameter estimates
y	a vector of death ages
X	a model matrix
y.left	left truncation age
y.right	right truncation age
wt	weight

Value

The negative log likelihood of parameter estimates given observed data

numident_demo *Demo Numident Data Set*

Description

A data set containing a sample of the CenSoc-Numident file, including age at death and select covariates.

Usage

```
numident_demo
```

Format

A data frame with 62,899 rows and 30 variables:

histid Historical unique identifier
byear Year of birth
bmonth Month of birth
dyear Year of death
dmonth Month of death
death_age Age at death (years)
weight CenSoc weight
zip_residence ZIP Code of residence at time of death
pernum Person number in sample unit
perwt IPUMS person weight
age Age in 1940
sex Sex in 1940
bpl Place of birth
mbpl Mother's place of birth
fbpl Father's place of birth
educd Educational attainment (detailed)
empstatd Employment status (detailed)
hispan Hispanic/Spanish/Latino origin
incnonwg Had non-wage/salary income over $50
incwage Wage and salary income
marst Marital status
nativity Foreign birthplace or parentage
occ Occupation
occscore Occupational income score
ownershp Ownership of dwelling (tenure)
race Race
rent Monthly contract rent
serial Household serial number
statefip State of residence 1940
urban Urban/rural status
educ_yrs Years of education attained

Details

The CenSoc-Numident dataset links the 1940 census to the National Archives' public release of the Social Security Numident file. The prelinked demo version of the file has 63 thousand mortality records and 20 mortality covariates from the 1940 census (~1 percent of the complete CenSoc-Numident dataset). Both demo and full versions of the data are available at https://censoc.berkeley.edu/data/.

Source

Joshua R. Goldstein, Monica Alexander, Casey Breen, Andrea Miranda González, Felipe Menares, Maria Osborne, Mallika Snyder, and Ugur Yildirim. CenSoc Mortality File: Version 2.0. Berkeley: University of California, 2021. https://censoc.berkeley.edu/.

Steven Ruggles, Sarah Flood, Ronald Goeken, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 12.0 (dataset). Minneapolis, MN: IPUMS, 2022. doi:10.18128/D010.V12.0.

`sim_data` *Simulated mortality data set*

Description

A data set containing simulated age at death and covariates according to a truncated Gompertz distribution with proportional hazards

Usage

```
sim_data
```

Format

A data frame with 6732 rows and 6 variables:

aod Age at death, in integer years

byear Calendar year of birth

dyear Calendar year of death

temp Temperature

sex Sex (0 = male, 1 = female)

isSouth Live in south (0 = FALSE, 1 = TRUE)

Index