Package 'cepumd'

July 22, 2025

Type Package

Title Calculate Consumer Expenditure Survey (CE) Annual Estimates

Version 2.1.0

Description Provides functions and data files to help CE Public-Use Microdata (PUMD) users calculate annual estimated expenditure means, standard errors, and quantiles according to the methods used by the CE with PUMD. For more information on the CE please visit <https://www.bls.gov/cex>. For further reading on CE estimate calculations please see the CE Calculation section of the U.S. Bureau of Labor Statistics (BLS) Handbook of Methods at <https://www.bls.gov/opub/hom/cex/calculation.htm>. For further information about CE PUMD please visit <https://www.bls.gov/cex/pumd.htm>.

License GPL (>= 3)

URL https://arcenis-r.github.io/cepumd/, https://github.com/arcenis-r/cepumd

BugReports https://github.com/arcenis-r/cepumd/issues

Depends R (>= 3.5.0)

Imports dplyr (>= 1.0.0), janitor, purrr, readr, readxl, rlang, stringr, tidyr, tidyselect (>= 1.2.0), utils

Suggests covr, knitr, rmarkdown, spelling, testthat (>= 2.1.0)

VignetteBuilder knitr

Encoding UTF-8

Language en-US

RoxygenNote 7.3.1

NeedsCompilation no

Author Arcenis Rojas [aut, cre, cph]

Maintainer Arcenis Rojas <arcenis.rojas@gmail.com>

Repository CRAN

Date/Publication 2024-03-18 17:50:02 UTC

Contents

`ce_hg` *Convert a CE heiarchical grouping file to a data frame*

Description

A CE heiarchical grouping ('HG') file shows the levels of aggregation for expenditure categories used to produce official CE expenditure estimates. This function reads in a CE HG file for the given year and HG type as data frame.

Usage

```
ce_hg(year, survey, hg_zip_path = NULL, hg_file_path = NULL)
```

Arguments

`year`	A year between 1996 and the last year of available CE PUMD.
`survey`	The type of HG file; one of "interview", "diary", or "integrated". Accepted as a character or symbol.
`hg_zip_path`	The path to a zip file containing HG files downloaded from the CE website. The structure of the zip file must be exactly as it is when downloaded to be useful to this function.
`hg_file_path`	The path to a single HG file that has already been extracted. If this argument is given 'hg_zip_path' is ignored.

Details

Interview and Diary HG files are available starting in 1997 and integrated files start in 1996. For consistency, this function and other `cepumd` functions only work with data starting in 1997.

The output will contain only expenditure UCCs and not UCCs related to household characteristics, income, assets, or liabilities. The scope of the functions in this package is limited to expenditures. Income, for example, is imputed and calculation of income means goes through a different process than do expenditure means. Please see User's Guide to Income Imputation in the CE

Value

A data frame containing the following columns:

- level - hierarchical level of the expenditure category
- title - the title of the expenditure category
- ucc - the Universal Classification Code (UCC) for the expenditure category
- survey - the survey instrument from which the data for a given UCC are sourced. This is most helpful when data for a type of expenditure are collected in both the Interview and the Diary.
- factor - the factor by which to multiply the expenditure in the calculation of estimated means / medians

Examples

```
## Not run:
# 'survey' can be entered as a string
ce_hg(2016, "integrated", "hg-files.zip")

# 'survey' can also be entered as a symbol
ce_hg(2016, integrated, "hg-files.zip")

## End(Not run)
```

`ce_mean` *Calculate a CE weighted mean*

Description

Calculate a weighted mean using the method used to produce official CE estimates.

Usage

```
ce_mean(ce_data)
```

Arguments

`ce_data`	A data frame containing at least a finlwt21 column, 44 replicate weight columns (wtrep01-44), a cost column, and a survey indicator column. All but the survey column must be numeric.

Value

A 1-row dataframe containing the following columns:

- agg_exp - The estimated aggregate expenditure
- mean_exp - The estimated mean expenditure
- se - The estimated standard error of the estimated mean expenditure
- cv - The coefficient of variation of the estimated mean expenditure

Note

Estimates produced using PUMD, which is topcoded by the CE and has some records suppressed to protect respondent confidentiality, will not match the published estimates released by the CE in most cases. The CE's published estimates are based on confidential data that are not topcoded nor have records suppressed. You can learn more at CE Protection of Respondent Confidentiality

See Also

`ce_quantiles()` `ce_prepdata()`

Examples

```
# Download the HG file keeping the section for expenditures on utilities
## Not run:
utils_hg <- ce_hg(2017, interview) |>
  ce_uccs("Utilities, fuels, and public services", uccs_only = FALSE)

## End(Not run)

# Download and prepare interview data
## Not run:
utils_interview <- ce_prepdata(
  2017,
  interview,
  uccs = ce_uccs(utils_hg, "Utilities, fuels, and public services"),
  zp = NULL,
  integrate_data = FALSE,
  hg = utils_hg,
  bls_urbn
)

## End(Not run)

# Calculate the mean expenditure on utilities
## Not run: ce_mean(utils_interview)

# Calculate the mean expenditure on utilities by urbanicity
## Not run:
utils_interview |>
  tidyr::nest(-bls_urbn) |>
  mutate(mean_utils = purrr::map(data, ce_mean)) |>
  select(-data) |>
  unnest(mean_utils)

## End(Not run)
```

`ce_prepdata` *Prepare CE data for calculating an estimated mean or median*

Description

Reads in the family characteristics (FMLI/-D) and expenditure tabulation (MTBI/EXPD) files and merges the relevant data for calculating a weighted mean or median.

Usage

```
ce_prepdata(
  year,
  survey,
  hg,
  uccs,
  ...,
  int_zp = NULL,
  dia_zp = NULL,
  recode_variables = FALSE,
  dict_path = NULL,
  own_codebook = NULL
)
```

Arguments

`year`	A year between 1997 and the last year of available CE PUMD.
`survey`	One of either interview, diary, or integrated as a character or symbol.
`hg`	A data frame that has, at least, the title, level, ucc, and factor columns of a CE HG file. Calling `ce_hg()` will generate a valid HG file.
`uccs`	A character vector of UCCs corresponding to expenditure categories in the hierarchical grouping (HG) for a given year and survey.
`...`	Variables to include in the dataset from the family characteristics file. This is intended to allow the user to calculate estimates for subsets of the data.
`int_zp`	String indicating the path of the Interview data zip file(s) if already stored. If the file(s) does not exist its corresponding zip file will be stored in that path. The default is `NULL` which causes the zip file to be stored in temporary memory during function operation.
`dia_zp`	Same as `int_zp` above, but for Diary data.
`recode_variables`	A logical indicating whether to recode all coded variables except 'UCC' using the codes in the CE's excel dictionary which can be downloaded from the CE Documentation Page
`dict_path`	A string indicating the path where the CE PUMD dictionary is stored if already stored. If the file does not exist and `recode_variables = TRUE` the dictionary

will be stored in this path. The default is NULL which causes the zip file to be stored in temporary memory during function operation. Automatically changed to NULL if a valid input for own_codebook is given.

own_codebook — An optional data frame containing a user-defined codebook containing the same columns as the CE Dictionary "Codes " sheet. If the input is not a data frame or does not have all of the required columns, the function will give an error message. See details for the required columns.

Details

CE microdata include 45 weights. The primary weight that is used for calculating estimated means and medians is finlwt21. The 44 replicate weights are computed using Balanced Repeated Replication (BRR) and are used for calculating weighted standard errors.

"Months in scope" refers to the proportion of the data collection quarter for which a CU reported expenditures. For the Diary survey the months in scope is always 3 because the expenditure data collected are meant to be reported for the quarter in which they are collected. The Interview Survey, on the other hand, is a quarterly, rolling, recall survey and the CU's report expenditures for the 3 months previous to the month in which the data are collected. For example, if a CU was interviewed in February 2017, then they would be providing data for November 2016, December 2016, and January 2017. If one is calculating a weighted estimated mean for the 2017 calendar year, then only the January 2017 data would be "in scope."

CE data are reported quarterly, but the sum of the weights (finlwt21) is for all CU's is meant to represent the total number of U.S. CU's for a given year. Since a calculating a calendar year estimate requires the use of 4 quarters of data and the sum of the weights in each quarter equals the number of households in the U.S. for a given year, adding up the sums of the weights in the 4 quarters of data would yield a total number of households that is approximately 4 times larger than the actual number of households in the U.S. in the corresponding year.

Since some UCC's can appear in both surveys, for the purposes of integration, the CE has a source selection procedure by which to choose which source data will be taken from for a given UCC. For example, of the 4 UCC's in the "Pets" category in 2017 two were sourced for publication from the Diary and two from the Interview. Please download the CE Source Selection Document for a complete listing: https://www.bls.gov/cex/ce_source_integrate.xlsx.

Family characteristic variables added through "..." will be read in as character data type.

Value

A data frame containing the following columns:

- newid - A consumer unit (CU), or household, identifier
- finlwt21 - CU weight variable
- wtrep01 through wtrep44 - CU replicate weight variables (see details)
- ... - Any family characteristics variables that were kept
- mo_scope - Months in scope (see details)
- popwt - An adjusted weight meant to account for the fact that a CUs value of finlwt21 is meant to be representative of only 1 quarter of data (see details)
- ucc - The UCC for a given expenditure

- ref_yr - The year in which the corresponding expenditure occurred
- ref_mo - The month in which the corresponding expenditure occurred
- cost - The value of the expenditure (in U.S. Dollars)
- survey - An indicator of which survey the data come from: "I" for Interview and "D" for Diary.

Examples

```
## Not run:
# The following workflow will prepare a dataset for calculating integrated
# pet expenditures for 2021 keep the "sex_ref" variable in the data to
# potentially calculate means by sex of the reference person.

# First generate an HG file
my_hg <- ce_hg(2021, integrated, "CE-HG-Inter-2021.txt")

# Store a vector of UCC's in the "Pets" category
pet_uccs <- ce_uccs(my_hg, "Pets")

# Store the diary data (not run)
pets_dia <- ce_prepdata(
  year = 2021,
  survey = integrated,
  uccs = pet_uccs,
  integrate_data = FALSE,
  hg = my_hg,
  dia_zip = "diary21.zip"
  sex_ref
)

## End(Not run)
```

`ce_quantiles` *Calculate a CE weighted quantiles*

Description

Calculate a CE weighted quantiles

Usage

```
ce_quantiles(ce_data, probs = 0.5)
```

Arguments

`ce_data`	A data frame containing at least a finlwt21 column and a cost column. Both columns must be numeric.
`probs`	A numeric vector of probabilities between 0 and 1 for which to compute quantiles. Default is 0.5 (median).

Value

A two-column data frame in which the first column contains the probabilities for which quantiles were calculated and their corresponding quantiles in the second column.

See Also

ce_mean()

Examples

```
## Not run:
# Download the HG file keeping the section for expenditures on utilities
utils_hg <- ce_hg(2017, interview) |>
  ce_uccs("Utilities, fuels, and public services", uccs_only = FALSE)

# Download and prepare interview data
utils_interview <- ce_prepdata(
  2017,
  interview,
  uccs = ce_uccs(utils_hg, "Utilities, fuels, and public services"),
  zp = NULL,
  integrate_data = FALSE,
  hg = utils_hg,
  bls_urbn
)

# Calculate the 25%, 50%, and 75% utilities expenditure quantiles
ce_quantiles(utils_interview)

# Calculate the 25%, 50%, and 75% utilities expenditure quantiles by
# urbanicity
utils_interview |>
  tidyr::nest(-bls_urbn) |>
  mutate(quant_utils = purrr::map(data, ce_quantiles, c(0.25, 0.5, 0.75))) |>
  select(-data) |>
  unnest(quant_utils)

## End(Not run)
```

---

ce_uccs *Find UCCs for expenditure categories*

Description

Find UCCs for expenditure categories

Usage

```
ce_uccs(hg, expenditure = NULL, ucc_group = NULL, uccs_only = TRUE)
```

Arguments

hg	A data frame that has, at least, the title, level, and ucc columns of a CE HG file.
expenditure	A string that is an expenditure category contained in a CE HG file (exact match required). Either expenditure or ucc_group is required. The default is NULL.
ucc_group	A string indicating an expenditure category by UCC group in a CE HG file (exact match required). Either expenditure or ucc_group is required. The default is NULL.
uccs_only	A logical indicating whether to return only the expenditure category's component ucc's. If TRUE (default), a vector of UCC's will be returned. If FALSE, a dataframe will be returned containing the section of the HG file containing the expenditure category and its component sub- categories

Details

If both a valid expenditure and valid ucc_group are input, ucc_group will be used.

Value

A vector of Universal Classification Codes (UCC's) corresponding to the lowest hierarchical level for that category.

Examples

```
## Not run:
# First generate an HG file
my_hg <- ce_hg(2021, interview, hg_file_path = "CE-HG-Inter-2021.txt")

# Store a vector of UCC's in the "Pets" category
pet_uccs <- ce_uccs(my_hg, "Pets")
pet_uccs
# [1] "610320" "620410" "620420"

## End(Not run)
```

---

get_survey_files *Generate tables of the necessary survey data files*

Description

Generate tables of the necessary survey data files

Usage

```
get_survey_files(year, survey, file_yrs, qtrs, zp_file)
```

Arguments

year	A year between 1996 and the last year of available CE PUMD.
survey	One of either interview, diary, or integrated as a character or symbol.
file_yrs	The substrings of years for which to pull data, i.e., for some years files have to be pulled from across different files.
qtrs	The quarters to be included in the analysis for a given year.
zp_file	Character indicating the zip file containing the CE PUMD for a given year

Details

This is a hidden file called only by exported package functions.

---

read.expd *Read in and modify EXPD files*

Description

Read in and modify EXPD files

Usage

```
read.expd(fp, zp, year, uccs, integrate_data, hg)
```

Arguments

fp	File to extract from zip file
zp	Zip file path
year	Year
uccs	Vector of UCC's to filter for
integrate_data	Whether to prepare data for integrated estimates
hg	Hierarchical grouping data

Details

This is a hidden file called only by exported package functions.

read.fmld *Read in and modify FMLD files*

Description

Read in and modify FMLD files

Usage

```
read.fmld(fp, zp, ...)
```

Arguments

fp	File to extract from zip file
zp	Zip file path
...	<dynamic-dots> Additional variables to keep (intended for grouping)

Details

This is a hidden file called only by exported package functions.

read.fmli *Read in and modify FMLI files*

Description

Read in and modify FMLI files

Usage

```
read.fmli(fp, zp, year, ...)
```

Arguments

fp	File to extract from zip file
zp	Zip file path within ce_dir
year	Year
...	<dynamic-dots> Additional variables to keep (intended for grouping)

Details

This is a hidden file called only by exported package functions.

read.mtbi *Read in and modify MTBI files*

Description

Read in and modify MTBI files

Usage

```
read.mtbi(fp, zp, year, uccs, integrate_data, hg)
```

Arguments

fp	File to extract from zip file
zp	Zip file path
year	Year
uccs	Vector of UCC's to filter for
integrate_data	Whether to prepare data for integrated estimates
hg	Hierarchical grouping data

Details

This is a hidden file called only by exported package functions.

recode_ce_variables *Recode variables in interview and diary data*

Description

Recode variables in interview and diary data

Usage

```
recode_ce_variables(srvy_data, code_file, srvy)
```

Arguments

srvy_data	A data frame containing either Interview or Diary data that has been prepped
code_file	A dataframe containing variable names, codes, code descriptions, and other required columns for recoding variables
srvy	The survey instrument to be recoded (this is for filtering the codebook)

Details

This is a hidden file called only by exported package functions.

Index