Package 'waou'

May 8, 2026

Title Weighting All of Us

Version 0.1.0

Description Utilities for using a probability sample to reweight prevalence estimates calculated from the All of Us research program. Weighted estimates will still not be representative of the general U.S. population. However, they will provide an early indication for how unweighted estimates may be biased by the sampling bias in the All of Us sample.

License AGPL (>= 3)

Encoding UTF-8

RoxygenNote 7.3.2

Suggests testthat (>= 3.0.0)

Imports glmnet, dplyr, stringr, stats, glue, mice, nonprobsvy, survey, ggplot2, purrr

Config/testthat/edition 3

Depends R (>= 3.5)

LazyData true

NeedsCompilation no

Author Daniel Brannock [aut, cre] (ORCID: <https://orcid.org/0000-0001-8095-547X>), Mahmoud Elkasabi [aut] (ORCID: <https://orcid.org/0000-0002-0720-4319>)

Maintainer Daniel Brannock <mbrannock@rti.org>

Repository CRAN

Date/Publication 2025-09-15 09:10:02 UTC

Contents

`adult2023` *NHIS Adult Data 2023*

Description

Raw survey results from adults for the 2023 National Health Interview Survey (NHIS). This is public use data. Documentation for the dataset can be found at the source link. NHIS is conducted by the National Center for Health Statistics within the Centers for Disease Control.

Usage

```
adult2023
```

Format

`adult2023`:

A data frame with 29,522 rows and 647 columns.

Source

https://www.cdc.gov/nchs/nhis/documentation/2023-nhis.html

`aou_synthetic` *Synthetic All of Us Data*

Description

Synthetic data intended to show how NHIS survey results can be used to generate weights from All of Us.

Usage

```
aou_synthetic
```

Format

Data frame with columns

SEX_A_R_I Sex: 0 (female), 1 (male)

AGEP_A_R_I Age in years: 1 (18-29), 2 (30-39), 3 (40-49), ..., 6 (70+)

HISPALLP_A_R_I Race/ethnicity: 1 (Hispanic), 2 (White), 3 (Black/African American), 4 (Other)

ORIENT_A_R_I Sexual orientation: 0 (Bisexual, Gay, or Lesbian), 1 (Straight)

HICOV_A_R_I Health insurance: 0 (Not insured), 1 (Insured)

EDUCP_A_R_I Education: 1 (Less than HS), 2 (HS or GED), 3 (Some college), 4 (College graduate), 5 (Advanced degree)

REGION_R_I Region: 1 (Northeast), 2 (Midwest), 3 (South), 4 (West)

EMPLASTWK_A_R_I Employment: 0 (Unemployed), 1 (Employed)

HOUTENURE_A_R_I Home ownership: 0 (Does not own home), 1 (Owns home)

MARITAL_A_R_I Marital status: 0 (Not married), 1 (Married)

DEPEV_A_R_I Depression: 0 (No diagnosis of depression), 1 (Has diagnosis of depression)

DEMENEV_A_R_I Depression: 0 (No diagnosis of dementia), 1 (Has diagnosis of dementia)

DIBTYPE_A_R_I Depression: 0 (No diagnosis of type 2 diabetes), 1 (Has diagnosis of type 2 diabetes)

Source

Generated from `data-raw/aou_synthetic.R`.

`calculate_weights` *Calculate Weights*

Description

Calculate weights using three methods: IPW, Calibration, and Calibration+IPW

Usage

```
calculate_weights(
  sample_a,
  sample_b,
  method,
  aux_variables,
  study_variables,
  weight,
  strata,
  psu
)
```

Arguments

sample_a	data.frame with representative sample
sample_b	data.frame with All of Us sample
method	string or string vector specifying weighting method to use: "ipw", "cal", and "ipw+cal"
aux_variables	character vector with names of calibration variables
study_variables	character vector with names of study variables
weight	character vector with name of the weight variable in sample_a
strata	character vector with name of the strata variable in sample_a
psu	character vector with name of the primary sampling units variable in sample_a

Details

Calculates weights intended to reduce the sampling bias present in All of Us. Three versions of weights are calculated from different reweighting strategies: IPW, Calibration, and Calibration+IPW.

Value

list of data.frame with added (or replaced) weight columns and survey designs

Examples

```
# Prepare the NHIS data
calVars <- c(
  "SEX_A_R", "AGEP_A_R", "HISPALLP_A_R", "ORIENT_A_R", "HICOV_A_R", "EDUCP_A_R", "REGION_R",
  "EMPLASTWK_A_R", "HOUTENURE_A_R", "MARITAL_A_R"
)
stuVars <- "DIBTYPE_A_R"
vars_dummies <- c("AGEP_A_R","HISPALLP_A_R","EDUCP_A_R","REGION_R")
nhis_keep_vars <- c("PPSU","PSTRAT","WTFA_A")
nhis_imputed <- impute_data(nhis_processed, c(calVars, stuVars), nhis_keep_vars)
nhis_dummied <- dummies(nhis_imputed, vars=paste0(vars_dummies, '_I'))
factor_vars <- setdiff(names(nhis_dummied), nhis_keep_vars)
nhis_dummied[factor_vars] <- lapply(nhis_dummied[factor_vars], as.factor)

# Prepare the synthetic All of Us data
aou_imputed <- impute_data(aou_synthetic, c(calVars, stuVars))
aou_dummied <- dummies(aou_imputed, vars=paste0(vars_dummies, '_I'))
aou_dummied[] <- lapply(aou_dummied, as.factor)

# Calculate IPW weights using NHIS data and applied to All of Us
weights_df <- calculate_weights(
  nhis_dummied,
  aou_dummied,
  'ipw',
  paste0(calVars, '_I'),
  paste0(stuVars, '_I'),
```

```
  weight='WTFA_A',
  strata='PSTRAT',
  psu='PPSU'
)
```

---

dummies *Create Dummy Variables*

Description

Create dummy variables of factors and character vectors in a data frame

Usage

```
dummies(input, vars)
```

Arguments

input	data.frame with calibration variables
vars	character vector with names of variables requiring dummy encoding

Value

data.frame with the new dummy variables

Examples

```
calVars <- c(
  "SEX_A_R", "AGEP_A_R", "HISPALLP_A_R", "ORIENT_A_R", "HICOV_A_R", "EDUCP_A_R", "REGION_R",
  "EMPLASTWK_A_R", "HOUTENURE_A_R", "MARITAL_A_R"
)
stuVars <- "DIBTYPE_A_R"
nhis_keep_vars <- c("PPSU","PSTRAT","WTFA_A")

# First impute
nhis_imputed <- impute_data(nhis_processed, c(calVars, stuVars), nhis_keep_vars)

# Then create dummy variables
nhis_vars_dummies <- c("AGEP_A_R","HISPALLP_A_R","EDUCP_A_R","REGION_R")
nhis_dummied <- dummies(nhis_imputed, vars=paste0(nhis_vars_dummies, '_I'))
```

extract_totals *Extract population totals*

Description

Calculate weights using three methods: IPW, Calibration, and Calibration+IPW

Usage

```
extract_totals(sample, vars, weight)
```

Arguments

sample	data.frame with representative sample
vars	character vector with names of calibration variables
weight	character vector with name of the weight variable

Details

Calculates weights intended to reduce the sampling bias present in All of Us. Three versions of weights are calculated from different reweighting strategies: IPW, Calibration, and Calibration+IPW.

Value

list of data.frame with added (or replaced) weight columns and survey designs

impute_data *Impute Data*

Description

Add imputed data columns to existing data.frame

Usage

```
impute_data(
  input,
  vars,
  keep_vars = c(),
  return_mice = FALSE,
  impute_constant = NULL
)
```

Arguments

input	data.frame with calibration variables
vars	character vector with names of variables to be imputed
keep_vars	character vector with names of additional variables that should be retained
return_mice	boolean for whether to return mice object (for looking at logged events)
impute_constant	numeric if not NULL will impute with provided constant

Details

For each of the specified variables, use all variables to predict missing values. Populate actual (when available) and imputed values into new columns appended with names appended with _I.

If you choose to return the mice object with return_mice, the function output will be a list that includes the final data.frame and the mice output.

Value

data.frame with imputed versions of variables

Examples

```
calVars <- c(
  "SEX_A_R", "AGEP_A_R", "HISPALLP_A_R", "ORIENT_A_R", "HICOV_A_R", "EDUCP_A_R", "REGION_R",
  "EMPLASTWK_A_R", "HOUTENURE_A_R", "MARITAL_A_R"
)
stuVars <- "DIBTYPE_A_R"
nhis_keep_vars <- c("PPSU","PSTRAT","WTFA_A")

nhis_imputed <- impute_data(nhis_processed, c(calVars, stuVars), nhis_keep_vars)
```

---

nhis_processed *Processed NHIS Data*

Description

Survey data from NHIS that has been sampled down, recoded, and subsetted.

Usage

```
nhis_processed
```

Format

Data frame with columns

SEX_A_R_I Sex: 0 (female), 1 (male)

AGEP_A_R_I Age in years: 1 (18-29), 2 (30-39), 3 (40-49), ..., 6 (70+)

HISPALLP_A_R_I Race/ethnicity: 1 (Hispanic), 2 (White), 3 (Black/African American), 4 (Other)

ORIENT_A_R_I Sexual orientation: 0 (Bisexual, Gay, or Lesbian), 1 (Straight)

HICOV_A_R_I Health insurance: 0 (Not insured), 1 (Insured)

EDUCP_A_R_I Education: 1 (Less than HS), 2 (HS or GED), 3 (Some college), 4 (College graduate), 5 (Advanced degree)

REGION_R_I Region: 1 (Northeast), 2 (Midwest), 3 (South), 4 (West)

EMPLASTWK_A_R_I Employment: 0 (Unemployed), 1 (Employed)

HOUTENURE_A_R_I Home ownership: 0 (Does not own home), 1 (Owns home)

MARITAL_A_R_I Marital status: 0 (Not married), 1 (Married)

DEPEV_A_R_I Depression: 0 (No self-reported depression), 1 (Has self-reported depression)

DEMENEV_A_R_I Depression: 0 (No self-reported dementia), 1 (Has self-reported dementia)

DIBTYPE_A_R_I Depression: 0 (No self-reported type 2 diabetes), 1 (Has self-reported type 2 diabetes)

PPSU Person-level ID

PSTRAT Stratification to be used as part of the survey design

WTFA_A Weights used to assure representativeness of U.S. population (may not be valid for sampled data)

Source

Generated from `data-raw/nhis_processed`.

`plot_prevalence` *Visualize Prevalence Estimates*

Description

Visualize prevalence estimates for calibration or outcome variables using different weighting methods.

Usage

```
plot_prevalence(df, mean, mean_se, method, cal_vars, cal_levels)
```

Arguments

df	data.frame with representative sample
mean	character name of mean prevalence estimate variable
mean_se	character name of mean prevalence estimate variable
method	character name of the weighting method variable
cal_vars	character name of the variable with calibration variable names
cal_levels	character name of the variable with calibration variable levels

Details

Specify columns and weighting methodologies of interest to visualize.

Value

ggplot object

Examples

```
library(dplyr)
library(stringr)

# Prepare the NHIS data
calVars <- c(
  "SEX_A_R", "AGEP_A_R", "HISPALLP_A_R", "ORIENT_A_R", "HICOV_A_R", "EDUCP_A_R", "REGION_R",
  "EMPLASTWK_A_R", "HOUTENURE_A_R", "MARITAL_A_R"
)
stuVars <- "DIBTYPE_A_R"
vars_dummies <- c("AGEP_A_R","HISPALLP_A_R","EDUCP_A_R","REGION_R")
nhis_keep_vars <- c("PPSU","PSTRAT","WTFA_A")
nhis_imputed <- impute_data(nhis_processed, c(calVars, stuVars), nhis_keep_vars)
nhis_dummied <- dummies(nhis_imputed, vars=paste0(vars_dummies, '_I'))
factor_vars <- setdiff(names(nhis_dummied), nhis_keep_vars)
nhis_dummied[factor_vars] <- lapply(nhis_dummied[factor_vars], as.factor)

# Prepare the synthetic All of Us data
aou_imputed <- impute_data(aou_synthetic, c(calVars, stuVars))
aou_dummied <- dummies(aou_imputed, vars=paste0(vars_dummies, '_I'))
aou_dummied[] <- lapply(aou_dummied, as.factor)

# Calculate IPW weights using NHIS data and applied to All of Us
weights_df <- calculate_weights(
  nhis_dummied,
  aou_dummied,
  'ipw',
  paste0(calVars, '_I'),
  paste0(stuVars, '_I'),
  weight='WTFA_A',
  strata='PSTRAT',
  psu='PPSU'
)
```

```
# Get IPW results by group
ipw_outcome_df <- summarize_results_by_group(
  weights_df,
  paste0(stuVars, '_I'),
  paste0(calVars, '_I'),
  weight_col='ipw_weight',
  label='AoU: IPW'
)

# Process data prior to plotting to make labels more readable
plot_df <- ipw_outcome_df %>%
  mutate(
    Name = case_when(
      group_var == 'SEX_A_R_I' & level_var == 1 ~ 'Sex: Male',
      group_var == 'SEX_A_R_I' & level_var == 0 ~ 'Sex: Female',
      group_var == 'AGEP_A_R_I1' & level_var == 1 ~ 'Age: 18-29',
      group_var == 'AGEP_A_R_I2' & level_var == 1 ~ 'Age: 30-39',
      group_var == 'AGEP_A_R_I3' & level_var == 1 ~ 'Age: 40-49',
      group_var == 'AGEP_A_R_I4' & level_var == 1 ~ 'Age: 50-59',
      group_var == 'AGEP_A_R_I5' & level_var == 1 ~ 'Age: 60-69',
      group_var == 'AGEP_A_R_I6' & level_var == 1 ~ 'Age: 70+',
      group_var == 'HISPALLP_A_R_I1' & level_var == 1 ~ 'Race/Eth: Hispanic',
      group_var == 'HISPALLP_A_R_I2' & level_var == 1 ~ 'Race/Eth: White',
      group_var == 'HISPALLP_A_R_I3' & level_var == 1 ~ 'Race/Eth: Black',
      group_var == 'HISPALLP_A_R_I4' & level_var == 1 ~ 'Race/Eth: Other',
      TRUE ~ group_var
    )
  ) %>%
  filter(str_detect(group_var, "SEX|AGEP|HISPALLP")) %>%
  filter(!str_detect(Name, "_")) %>%
  mutate(
    condition = case_when(
      outcome_var == 'DIBTYPE_A_R_I' ~ "Diabetes"
    ),
    VAR = case_when(
      str_detect(group_var, "SEX") ~ "Sex",
      str_detect(group_var, "AGE") ~ "Age",
      str_detect(group_var, "HISPALL") ~ "Race",
      str_detect(group_var, "EDUC") ~ "Educ"
    )
  )

# Plot
plot_prevalence(
  plot_df,
  'WMEAN',
  'SEMEAN',
  'Method',
  'VAR',
  'Name'
)
```

select_variables *Select Variables*

Description

Select variables relevant to propensity for inclusion in All of Us

Usage

```
select_variables(sample_a, sample_b, aux_variables)
```

Arguments

sample_a	data.frame of the reference probability sample (i.e., NHIS)
sample_b	data.frame of the All of Us sample
aux_variables	character vector with names of auxiliary variables

Details

Chooses which variables are meaningful in modeling propensity for inclusion in All of Us (sample_b) as compared to the general US population as represented by a reference probability sample (sample_a). This function assumes that variable names in both sample_a and sample_b are harmonized (i.e., definitions and names are the same across the two sources).

Value

character vector with selected variable names

Examples

```
# Prepare the NHIS data
calVars <- c(
  "SEX_A_R", "AGEP_A_R", "HISPALLP_A_R", "ORIENT_A_R", "HICOV_A_R", "EDUCP_A_R", "REGION_R",
  "EMPLASTWK_A_R", "HOUTENURE_A_R", "MARITAL_A_R"
)
stuVars <- "DIBTYPE_A_R"
vars_dummies <- c("AGEP_A_R","HISPALLP_A_R","EDUCP_A_R","REGION_R")
nhis_keep_vars <- c("PPSU","PSTRAT","WTFA_A")
nhis_imputed <- impute_data(nhis_processed, c(calVars, stuVars), nhis_keep_vars)
nhis_dummied <- dummies(nhis_imputed, vars=paste0(vars_dummies, '_I'))
factor_vars <- setdiff(names(nhis_dummied), nhis_keep_vars)
nhis_dummied[factor_vars] <- lapply(nhis_dummied[factor_vars], as.factor)

# Prepare the synthetic All of Us data
aou_imputed <- impute_data(aou_synthetic, c(calVars, stuVars))
aou_dummied <- dummies(aou_imputed, vars=paste0(vars_dummies, '_I'))
aou_dummied[] <- lapply(aou_dummied, as.factor)

# Define base variable names of auxiliary variables
```

```
aux_variables <- c(
  "SEX_A_R_I","AGEP_A_R_I", "HISPALLP_A_R_I","EDUCP_A_R_I",
  "REGION_R_I","ORIENT_A_R_I","HICOV_A_R_I",
  "EMPLASTWK_A_R_I","HOUTENURE_A_R_I","MARITAL_A_R_I"
)

# Provide All of Us and NHIS data to select variables
selected_base_vars <- select_variables(nhis_dummied, aou_dummied, aux_variables)
```

`summarize_results` *Summarize Results*

Description

Get adjusted totals and prevalence for provided variables.

Usage

```
summarize_results(
  df,
  vars,
  weight_col = NULL,
  id_col = 1,
  strata_col = NULL,
  label = NULL
)
```

Arguments

`df`	data.frame with sample and weights (if using a survey design)
`vars`	string vector of variables to calculate prevalences for
`weight_col`	string specifying the column with weights or NULL for unweighted
`id_col`	string specifying the column with IDs for cluster-aware standard error (SE) calculations
`strata_col`	string specifying the column with strata for cluster-aware SE calculations
`label`	string label for weighting method

Value

data.frame with totals, means, and standard errors (if using a survey design)

Examples

```
# Prepare the NHIS data
calVars <- c(
  "SEX_A_R", "AGEP_A_R", "HISPALLP_A_R", "ORIENT_A_R", "HICOV_A_R", "EDUCP_A_R", "REGION_R",
  "EMPLASTWK_A_R", "HOUTENURE_A_R", "MARITAL_A_R"
)
stuVars <- "DIBTYPE_A_R"
vars_dummies <- c("AGEP_A_R","HISPALLP_A_R","EDUCP_A_R","REGION_R")
nhis_keep_vars <- c("PPSU","PSTRAT","WTFA_A")
nhis_imputed <- impute_data(nhis_processed, c(calVars, stuVars), nhis_keep_vars)
nhis_dummied <- dummies(nhis_imputed, vars=paste0(vars_dummies, '_I'))
factor_vars <- setdiff(names(nhis_dummied), nhis_keep_vars)
nhis_dummied[factor_vars] <- lapply(nhis_dummied[factor_vars], as.factor)

# Prepare the synthetic All of Us data
aou_imputed <- impute_data(aou_synthetic, c(calVars, stuVars))
aou_dummied <- dummies(aou_imputed, vars=paste0(vars_dummies, '_I'))
aou_dummied[] <- lapply(aou_dummied, as.factor)

# Calculate IPW weights using NHIS data and applied to All of Us
weights_df <- calculate_weights(
  nhis_dummied,
  nhis_dummied,
  'ipw',
  paste0(calVars, '_I'),
  paste0(stuVars, '_I'),
  weight='WTFA_A',
  strata='PSTRAT',
  psu='PPSU'
)

results_ipw <- summarize_results(
  weights_df,
  c(paste0(calVars, '_I'), paste0(stuVars, '_I')),
  weight_col='ipw_weight',
  label='AoU: IPW'
)
```

---

summarize_results_by_group *Summarize Results by Group*

Description

Get adjusted totals and prevalences for provided variables, grouped by specified variables.

Usage

```
summarize_results_by_group(
  df,
  vars,
  group_vars,
  weight_col = NULL,
  id_col = NULL,
  strata_col = NULL,
  label = NULL
)
```

Arguments

df	data.frame with sample and weights (if using a survey design)
vars	string vector of variables to calculate prevalences for
group_vars	string vector of variables to group by
weight_col	string specifying the column with weights, "nhis" or nhis survey design, or NULL for unweighted
id_col	string specifying the column with IDs for cluster-aware standard error (SE) calculations
strata_col	string specifying the column with strata for cluster-aware SE calculations
label	string label for weighting method

Details

TODO: Merge into regular summarize_results function

Value

data.frame with totals, means, and standard errors (if using a survey design)

Examples

```
# Prepare the NHIS data
calVars <- c(
  "SEX_A_R", "AGEP_A_R", "HISPALLP_A_R", "ORIENT_A_R", "HICOV_A_R", "EDUCP_A_R", "REGION_R",
  "EMPLASTWK_A_R", "HOUTENURE_A_R", "MARITAL_A_R"
)
stuVars <- "DIBTYPE_A_R"
vars_dummies <- c("AGEP_A_R","HISPALLP_A_R","EDUCP_A_R","REGION_R")
nhis_keep_vars <- c("PPSU","PSTRAT","WTFA_A")
nhis_imputed <- impute_data(nhis_processed, c(calVars, stuVars), nhis_keep_vars)
nhis_dummied <- dummies(nhis_imputed, vars=paste0(vars_dummies, '_I'))
factor_vars <- setdiff(names(nhis_dummied), nhis_keep_vars)
nhis_dummied[factor_vars] <- lapply(nhis_dummied[factor_vars], as.factor)

# Prepare the synthetic All of Us data
aou_imputed <- impute_data(aou_synthetic, c(calVars, stuVars))
```

```
aou_dummied <- dummies(aou_imputed, vars=paste0(vars_dummies, '_I'))
aou_dummied[] <- lapply(aou_dummied, as.factor)

# Calculate IPW weights using NHIS data and applied to All of Us
weights_df <- calculate_weights(
  nhis_dummied,
  aou_dummied,
  'ipw',
  paste0(calVars, '_I'),
  paste0(stuVars, '_I'),
  weight='WTFA_A',
  strata='PSTRAT',
  psu='PPSU'
)

# Get IPW results by group
ipw_outcome_df <- summarize_results_by_group(
  weights_df,
  paste0(stuVars, '_I'),
  paste0(calVars, '_I'),
  weight_col='ipw_weight',
  label='AoU: IPW'
)
```

Index