Package 'svycdiff'

May 9, 2026

Type Package

Title Controlled Difference Estimation for Complex Surveys

Version 0.2.0

Maintainer Stephen Salerno <ssalerno@fredhutch.org>

Description Estimates the population average controlled difference for a given outcome between levels of a binary treatment, exposure, or other group membership variable of interest for clustered, stratified survey samples where sample selection depends on the comparison group. Provides three methods for estimation, namely outcome modeling and two factorizations of inverse probability weighting. Under stronger assumptions, these methods estimate the causal population average treatment effect. Salerno et al., (2024) <doi:10.48550/arXiv.2406.19597>.

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

VignetteBuilder knitr

Imports MASS, betareg, numDeriv, stats, survey

Depends R (>= 4.1.0)

Suggests knitr, rmarkdown, markdown, spelling

URL https://github.com/salernos/svycdiff, https://salernos.github.io/svycdiff/

BugReports https://github.com/salernos/svycdiff/issues

Language en-US

NeedsCompilation no

Author Stephen Salerno [aut, cre, cph] (ORCID: <https://orcid.org/0000-0003-2763-0494>), Emily K Roberts [aut], Tyler H McCormick [aut], Fan Li [aut], Bhramar Mukherjee [aut], Xu Shi [aut]

Repository CRAN

Date/Publication 2025-07-22 20:40:12 UTC

Contents

NHANES *Race, SES, and Telomere Length Data*

Description

National Health and Nutrition Examination Survey (NHANES) data on race, socioeconomic status, and leukocyte telomere length from the 1999-2000 and 2001-2002 survey waves.

Usage

```
data(NHANES)
```

Format

A dataset with 5,298 observations (rows) of 29 variables (columns):

SEQN Numeric: Respondent Sequence Number

iWTMEC4YR Numeric: 1/WTMEC4YR (Full Sample 4 Year Probability of Selection)

WTMEC4YR Numeric: Full Sample 4 Year Interview Weight

SDMVPSU Numeric: Masked Variance Pseudo-PSU

SDMVSTRA Numeric: Masked Variance Pseudo-Stratum

TELOMEAN Numeric: Mean T/S Ratio (See Details)

lTELOMEAN Numeric: log(TELOMEAN)

RACE_2CAT Numeric: 0 = Non-Hispanic White, 1 = Non-Hispanic Black (0/1 Coded for Current Functionality)

AGE Numeric: Age at Screening (Years)

SEX Factor w/ 2 Levels: Self-Reported Sex - Male, Female

EDUC_3CAT Factor w/ 3 Levels: Education - High School or GED, Some College, College Graduate

MARTL_3CAT Factor w/ 3 Levels: Marital Status - Never Married, Widowed/Divorced/Separated, Married/Living with Partner

HHSIZE_3CAT Factor w/ 5 Levels: Household Size - 1 Person, 2 People, 3 People, 4 People, 5+ People

HHINC_5CAT Factor w/ 5 Levels: Annual Household Income - $0 - $20,000, $20,000 - $35,000, $35,000 - $55,000, $55,000 - $75,000, $75,000+

PIR Factor w/ 3 Levels: Family Poverty-Income Ratio Category - < 1.3, 1.3 <= PIR < 3.5, >= 3.5

EMPSTAT_4CAT Factor w/ 4 Levels: Employment Status - Full-Time, Part-Time, Retired, Not Working

OCC_5CAT Factor w/ 5 Levels: Occupation Category - No Work, Low Blue Collar, High Blue Collar, Low White Collar, High White Collar

WIC_2CAT Factor w/ 2 Levels: WIC Utilization - No WIC, Received WIC

FDSEC_3CAT Factor w/ 3 Levels: Food Security Status - Food Secure, Marginally Food Secure, Food Insecure

HOD_4CAT Factor w/ 4 Levels: Home Type - Family Home Detached, Family Home Attached, Apartment, Other

OWNHOME_2CAT Factor w/ 2 Levels: Home Ownership - Does Not Own Home, Owns Home

HIQ_2CAT Factor w/ 2 Levels: Insurance Status - Not Insured, Insured

LBXWBCSI Numeric: White Blood Cell Count (SI)

LBXLYPCT Numeric: Lymphocyte Percent (%)

LBXMOPCT Numeric: Monocyte Percent (%)

LBXNEPCT Numeric: Segmented Neutrophils Percent (%)

LBXEOPCT Numeric: Eosinophils Percent (%)

LBXBAPCT Numeric: Basophils Percent (%)

LBXBPB_LOD Numeric: Blood Lead Concentration (ug/dL; LOD = 0.3 ug/dL; Imputed by LOD / sqrt(2))

Details

Our initial sample consisted of 7,839 participants in the 1999-2002 NHANES waves with laboratory measures recorded, including telomere length, `LTELOMEAN`, which was assayed via quantitative polymerase chain reaction (PCR; see Cawthorn, 2002). Our primary endpoint is the log-transformed mean ratio of an individual's telomere length to a standard reference DNA sample across all leukocyte cell types (mean T/S), `TELOMEAN`. We focus on the 1999-2002 NHANES waves, as they featured 4-year adjusted survey weights, `WTMEC4YR`, designed for aggregating data across cohorts. Among the initial 7,839 participants, 5,308 (67.7%) self-identified as either non-Hispanic White or non-Hispanic Black. Excluding those participants without our outcome of interest, our final analytic sample contained 5,298 Non-Hispanic White or Non-Hispanic Black identifying participants with measured telomere length. Race, `RACE_2CAT`, is our variable of interest. We further included study participant age, sex, and blood cell composition to account for known differences in these factors, as well as twelve indicators of SES. Ten of these, namely marital status, education level, household income, insurance status, Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) usage, household size, home ownership, home type, food security status, and an individual's poverty income ratio (PIR), were extracted directly from the NHANES demographic and occupation questionnaires. Occupation category was constructed by mapping occupation group codes in the NHANES 1999-2002 occupation questionnaire to the national statistics socioeconomic job classifications, and employment status was derived from three occupational measures: type of work done last week, hours worked last week at all jobs, and main reason for not working last week (see Rehkopf et al., 2008, Rose et al., 2005).

Source

<https://www.cdc.gov/nchs/nhanes/index.htm>

References

Richard M Cawthon. Telomere measurement by quantitative pcr. Nucleic acids research, 30(10):e47–e47, 2002.

David H Rehkopf, Lisa F Berkman, Brent Coull, and Nancy Krieger. The non-linear risk of mortality by income level in a healthy population: Us national health and nutrition examination survey mortality follow-up cohort, 1988–2001. BMC Public Health, 8(1):1–11, 2008.

David Rose, David J Pevalin, and Karen O'Reilly. The National Statistics Socio-economic Classification: origins, development and use. Palgrave Macmillan, 2005.

Examples

```
data(NHANES)
```

`simdat` *Simulate data with varying degrees of selection and confounding bias*

Description

Function to simulate data based on specified relationships between the generated outcome, group variable, confounder(s), and selection mechanism.

Usage

```
simdat(
  N = 1e+06,
  p = 1,
  q = 0,
  n_strat = 1,
  n_clust = 1,
  sigma_strat = 1,
  sigma_clust = 1,
  X_fam = c("gaussian", "binary"),
  tau_0 = 0,
  tau_A = 1,
  tau_X = rep(1, p),
  tau_X12 = 0,
  beta_0 = 0,
  beta_A = 1,
  beta_X = rep(1, p),
  beta_U = rep(1, q),
  Y_fam = c("gaussian", "binary", "poisson"),
  alpha_0 = 0,
```

```
  alpha_A = 1,
  alpha_X = rep(1, p),
  alpha_AX = 0
)
```

Arguments

N	int - Number of observations to be generated. Defaults to 1000000.
p	int - Number of covariates to be generated. Defaults to 1.
q	int - Number of additional covariates that affect selection to be generated. Defaults to 0.
n_strat	int - Number of strata in the population to be generated. Defaults to 1.
n_clust	int - Number of clusters within each stratum in the population to be generated. Defaults to 1.
sigma_strat	double - Standard deviation of covariate means across strata. Defaults to 1.
sigma_clust	double - Standard deviation of covariate means across clusters. Defaults to 1.
X_fam	string - Distribution of the covariates, X. Defaults to a multivariate normal distribution with mean equal to the sum of the cluster and stratum means, and an identity covariance matrix. If "binary", continuous covariates are discretized at their median values.
tau_0	double - Intercept for propensity model. Defaults to 0.
tau_A	double - Scaling factor for group assignment. Defaults to 1.
tau_X	double - Coefficients for X in propensity model. Defaults to a 1 vector of length p.
tau_X12	double - Interaction term coefficient for X1*X2 if p > 1. Defaults to 0.
beta_0	double - Intercept for selection model. Defaults to 0.
beta_A	double - Coefficient for A in selection model. Defaults to 1.
beta_X	double - Coefficients for X in selection model. Defaults to a 1 vector of length p.
beta_U	double - Coefficients for U (additional covariates affection only selection) in selection model. Defaults to a 1 vector of length q.
Y_fam	string - Distribution of the outcome variable, Y. Defaults to "gaussian" for a normally distributed outcome. Other options include "binary" for a Bernoulli-distributed outcome and "poisson" for a Poisson-distributed outcome.
alpha_0	double - Intercept for outcome model. Defaults to 0.
alpha_A	double - Coefficient for A in outcome model. Defaults to 1.
alpha_X	double - Coefficients for X in outcome model. Defaults to a 1 vector of length p.
alpha_AX	double - Coefficient for interaction between A and X in outcome model. Defaults to 0.

Details

The function generates data in a hierarchical structure with stratified clusters. The data generation process follows these steps:

1. **Stratum and Cluster Means:** For each of the `n_strat` strata, a matrix of stratum-level means for `p` covariates is generated from a normal distribution with standard deviation `sigma_strat`. Similarly, for each of the `n_clust` clusters within each stratum, cluster-level means are generated from a normal distribution with standard deviation `sigma_clust`.

2. **Covariate Generation:** Within each cluster, covariates, `X`, for `N / (n_strat * n_clust)` individuals are generated from a multivariate normal distribution with mean equal to the sum of the cluster and stratum means, and an identity covariance matrix.

3. **Covariate Transformation:** If `X_fam` is `"binary"`, each covariate is discretized at its median, otherwise it remains continuous.

4. **Propensity Model:** The group variable, `A`, is generated using a logistic regression model with intercept `tau_0`, covariate effects `tau_X`, and an interaction effect between the first two covariates with coefficient `tau_X12`. The group membership probability, `pA`, is defined by the logistic model.

5. **Selection Model:** The probability of selection, `pS`, is generated using a logistic regression model with intercept `beta_0`, group effect `beta_A`, and covariate effects `beta_X`. Gaussian noise is added to the linear predictor.

6. **Outcome Model:** The outcome, `Y`, is generated based on a chosen outcome distribution, `Y_fam`. The linear predictor includes an intercept, `alpha_0`, group effect, `alpha_A`, covariate effects, `alpha_X`, and an optional interaction effect, `alpha_AX`, between the group variable and covariates.

7. **Controlled Difference:** The true controlled difference in the outcome between groups is calculated as `CDIFF`.

The output is a data frame containing the generated outcome, group variable, covariates, and selection probabilities.

Value

A `data.frame` with `N` observations and the following variables:

Strata Stratum index (integer)

Cluster Cluster index (integer)

X1, X2, ..., Xp Confounding covariates (continuous or binary, depending on `X_fam`)

pA True probability of A = 1 conditional on X (continuous)

A Group assignment (binary)

pS True probability of selection conditional on A and X (continuous)

Y0 Potential outcome under A = 0 (continuous, binary, or count depending on `Y_fam`)

Y1 Potential outcome under A = 1 (continuous, binary, or count depending on `Y_fam`)

Y Observed outcome, based on treatment assignment (continuous, binary, or count depending on `Y_fam`)

CDIFF True controlled difference in outcomes by comparison group (double, computed as mean(Y1 - Y0))

Examples

```
N <- 100000

dat <- simdat(N)

head(dat)
```

---

svycdiff *Controlled Difference Estimation for Complex Surveys*

Description

This is the main function to estimate population average controlled difference (ACD), or under stronger assumptions, the population average treatment effect (PATE), for a given outcome between levels of a binary treatment, exposure, or other group membership variable of interest for clustered, stratified survey samples where sample selection depends on the comparison group.

Usage

```
svycdiff(
  df,
  id_form,
  a_form,
  s_form,
  y_form,
  y_fam = NULL,
  strata = NULL,
  cluster = NULL
)
```

Arguments

df	a 'data.frame' or 'tibble' containing the variables in the models.
id_form	a 'string' indicating which identification formula to be used. Options include "OM", "IPW1", "IPW2", or "DR". See 'Details' for information.
a_form	an object of class 'formula' which describes the propensity score model to be fit.
s_form	an object of class 'formula' which describes the selection model to be fit.
y_form	an object of class 'formula' which describes the outcome model to be fit. Only used if id_form = "OM" or id_form = "DR", else y_form = y ~ 1.
y_fam	a 'family' function. Only used if id_form = "OM" or id_form = "DR", else y_fam = NULL. Current options include gaussian, binomial, or poisson.
strata	a 'string' indicating strata, else strata = NULL for no strata.
cluster	a 'string' indicating cluster IDs, else cluster = NULL for no clusters.

Details

The argument `id_form` takes possible values `"OM"`, `"IPW1"`, `"IPW2"`, or `"DR"`, corresponding to the four formulas presented in Salerno et al. `"OM"` refers to the method that uses outcome modeling and direct standardization to estimate the controlled difference, while `"IPW1"` and `"IPW2"` are inverse probability weighted methods. `"IPW1"` and `"IPW2"` differ with respect to how the joint propensity and selection mechanisms are factored (see Salerno et al. for additional details). `"DR"` refers to the doubly robust form of estimator, which essentially combines `"OM"` and `"IPW2"`.

For `id_form = "IPW1"` or `id_form = "IPW2"`, `y_form` should be of the form `Y ~ 1`.

For known selection mechanisms, `s_form` should be of the form `pS ~ 1`, where `pS` is the variable corresponding to the probability of selection (e.g., inverse of the selection weight), and there should be two additional variables in the dataset: `P_S_cond_A1X` and `P_S_cond_A0X`, corresponding to the known probability of selection conditional on $A = 1$ or 0 and $X = x$, respectively. If these quantities are not known, `s_form` should contain the variables which affect sample selection on the right hand side of the equation, including the comparison group variable of interest.

Value

'svycdiff' returns an object of class "svycdiff" which contains:

id_form A string denoting Which method was selected for estimation

cdiff A named vector containing the point estimate (est), standard error (err), lower confidence limit (lcl), upper confidence limit (ucl), and p-value (pval) for the estimated controlled difference

fit_y An object of class inheriting from "glm" corresponding to the outcome model fit, or NULL for IPW1 and IPW2

fit_a An object of class inheriting from "glm" corresponding to the propensity model fit

wtd_fit_a An object of class inheriting from "glm" corresponding to the weighted propensity model fit

fit_s An object of class "betareg" corresponding to the selection model fit, or NULL if the selection mechanism is known

Examples

```
N <- 1000

dat <- simdat(N)

S <- rbinom(N, 1, dat$pS)

samp <- dat[S == 1,]

y_mod <- Y ~ A * X1

a_mod <- A ~ X1

s_mod <- pS ~ A + X1

fit <- svycdiff(samp, "DR", a_mod, s_mod, y_mod, "gaussian")
```

```
fit

summary(fit)
```

Index