# Package 'svycdiff'

August 20, 2024

**Type** Package

**Title** Controlled Difference Estimation for Complex Surveys

**Version** 0.1.1

**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** betareg, numDeriv, stats, survey

**Depends** R (>= 3.5.0)

**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] (<https://orcid.org/0000-0003-2763-0494>), Emily K Roberts [aut], Tyler H McCormick [aut], Bhramar Mukherjee [aut], Xu Shi [aut]

**Repository** CRAN

**Date/Publication** 2024-08-20 14:40:03 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, `1TELOMEAN`, 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 (continuous) outcome, variable of interest, confounder, and selection mechanism.

## Usage

```
simdat(
  N,
  X_dist = "continuous",
  S_known = FALSE,
  tau_0 = 0,
  tau_X = 1,
  beta_0 = 0,
  beta_A = 1,
  beta_X = 1,
  hetero = TRUE,
  alpha_0 = 0,
  alpha_X = 1,
  alpha_A = 1,
  alpha_AX = 0.1
)
```

## Arguments

| | |
|---|---|
| `N` | int - Number of observations to be generated |
| `X_dist` | string - Distribution of the confounding variable, X. Defaults to "continuous" for a N(1, 1) variable, or "binary" for a Bernoulli(0.5) variable |

| | |
|---|---|
| S_known | boolean - Logical for whether the selection mechanism should be treated as known (deterministic) or needs to be estimated (simulated with Gaussian error; defaults to FALSE) |
| tau_0 | double - Intercept for propensity model (defaults to 0) |
| tau_X | double - Coefficient for X in propensity model (defaults to 1) |
| 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 - Coefficient for X in selection model (defaults to 1) |
| hetero | boolean - Logical for heterogeneous treatment effect in the outcome model (defaults to TRUE) |
| alpha_0 | double - Intercept for outcome model (defaults to 0) |
| alpha_X | double - Coefficient for X in outcome model (defaults to 1) |
| alpha_A | double - Coefficient for A in outcome model (defaults to 1) |
| alpha_AX | double - Coefficient for interaction between A and X in outcome model (only used if hetero == TRUE; defaults to 0.1) |

## Details

The data are generated as follows. For a user-given number, N, observations in our so-called super population, we first generate a confounding variable, X, which relates to our outcome, Y, our variable of interest, A, and our selection indicator, S. We generate population-level data with X ~ N(1,1) or X ~ Bern(0.5) depending on whether distribution of X is chosen to be X_dist = "continous" or X_dist = "binary", respectively.

We then generate the remaining data from three models:

**1. Propensity Model**

**2. Selection Model**

**3. Outcome Model**

## Value

A data.frame with N observations of 7 variables:

**Y** Observed outcome (continuous)

**A** Comparison group variable of interest (binary)

**X** Confounding variable (continuous or binary)

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

**P_S_cond_AX** True probability of selection (S = 1) conditional on A and X (continuous)

**P_S_cond_A1X** True probability of selection (S = 1) conditional on A = 1 and X (continuous)

**P_S_cond_A0X** True probability of selection (S = 1) conditional on A = 0 and X (continuous)

**CDIFF** True controlled difference in outcomes by comparison group (double)

## 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", or "IPW2". See 'Details' for more 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", else 'y_form = y ~ 1'. |
| y_fam | a family function. Only used if 'id_form' = "OM", else 'y_fam = NULL'. |
| 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", or "IPW2", corresponding to the three 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). "OM", "IPW1", or "IPW2" are useful in different settings, which warrants some brief discussion here. "OM" requires you to specify an outcome regression model, whereas "IPW1" and "IPW2" do not require estimation, nor do they assume additivity or interactivity. However, while OM, IPW1, and IPW2 are consistent, OM is most efficient if correctly specified.

For IPW1/IPW2, `y_form` should be of the form `Y ~ 1`.

For known S, `s_form` should be of the form `S ~ 1`, where `S` is the variable corresponding to the probability of selection. 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$P_S_cond_AX)

samp <- dat[S == 1,]

y_mod <- Y ~ A * X

a_mod <- A ~ X
```

```
s_mod <- P_S_cond_AX ~ A + X

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

fit

summary(fit)
```

# Index