# Package 'riskCommunicator'

July 23, 2025

**Title** G-Computation to Estimate Interpretable Epidemiological Effects

**Version** 1.0.1

**Depends** R (>= 3.5)

**Imports** boot, dplyr, ggplot2, ggpubr, magrittr, MASS, methods, purrr, rlang, stats, tidyr, tidyselect

**Description** Estimates flexible epidemiological effect measures including both differences and ratios using the parametric G-formula developed as an alternative to inverse probability weighting. It is useful for estimating the impact of interventions in the presence of treatment-confounder-feedback. G-computation was originally described by Robbins (1986) <doi:10.1016/0270-0255(86)90088-6> and has been described in detail by Ahern, Hubbard, and Galea (2009) <doi:10.1093/aje/kwp015>; Snowden, Rose, and Mortimer (2011) <doi:10.1093/aje/kwq472>; and Westreich et al. (2012) <doi:10.1002/sim.5316>.

**License** GPL-3

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.2.0

**Suggests** knitr, rmarkdown, testthat, tidyverse, printr, stringr, formatR, sandwich

**VignetteBuilder** knitr

**NeedsCompilation** no

**Author** Jessica Grembi [aut, cre, cph] (ORCID: <https://orcid.org/0000-0001-6142-4913>), Elizabeth Rogawski McQuade [ctb] (ORCID: <https://orcid.org/0000-0002-4942-3747>)

**Maintainer** Jessica Grembi <jess.grembi@gmail.com>

**Repository** CRAN

**Date/Publication** 2022-05-31 23:20:02 UTC

## Contents

---

`cvdd` *A subset of the* `framingham` *teaching data*

---

## Description

A subset of the `framingham` teaching dataset containing the following changes:

- removal of all observations where PERIOD == 2 or PERIOD == 3 (i.e. keep only PERIOD == 1)
- removal of all observations where PREVCHD == 1 (i.e. all patients with coronary heart disease at baseline)
- created a new variable, `cvd_dth` signifying an outcome of cardiovascular disease OR death (i.e. if the patient had either CVD or DEATH, this new variable is 1, otherwise 0)
- created a new variable, `timeout`, which calculates the number of days from the start of the study to cardiovascular disease, death, or loss to follow-up
- created a new variable, `logpdays`, which is the log of `timeout`
- created a new variable, `nhosp`, which is a simulated number of hospitalizations

## Usage

```
data(cvdd)
```

## Format

A data frame with 4240 rows and 31 variables:

**RANDID** Unique identification number for each participant.

**SEX** Participant sex. 0 = Male, 1 = Female.

**TOTCHOL** Serum Total Cholesterol (mg/dL).

**AGE** Age at exam (years).

**SYSBP** Systolic Blood Pressure (mean of last two of three measurements) (mmHg).

**DIABP** Diastolic Blood Pressure (mean of last two of three measurements) (mmHg).

**CURSMOKE** Current cigarette smoking at exam. 0 = Not current smoker, 1 = Current smoker.

**CIGPDAY** Number of cigarettes smoked each day. 0 = Not current smoker.

**BMI** Body Mass Index, weight in kilograms/height meters squared.

**DIABETES** Diabetic according to criteria of first exam treated or first exam with casual glucose of 200 mg/dL or more. 0 = Not a diabetic, 1 = Diabetic.

**BPMEDS** Use of Anti-hypertensive medication at exam. 0 = Not currently used, 1 = Current use.

**HEARTRTE** Heart rate (Ventricular rate) in beats/min.

**GLUCOSE** Casual serum glucose (mg/dL).

**educ** Level of completed education. 1 = 0-11 years, 2 = high school or GED, 3 = some college, 4 = college graduate or higher.

**PREVSTRK** Prevalent Stroke. 0 = Free of disease, 1 = Prevalent disease.

**PREVHYP** Prevalent Hypertensive. Subject was defined as hypertensive if treated or if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg. 0 = Free of disease, 1 = Prevalent disease.

**DEATH** Death from any cause. 0 = Did not occur during followup, 1 = Did occur during followup.

**ANGINA** Angina Pectoris. 0 = Did not occur during followup, 1 = Did occur during followup.

**HOSPMI** Hospitalized Myocardial Infarction. 0 = Did not occur during followup, 1 = Did occur during followup.

**MI_FCHD** Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease. 0 = Did not occur during followup, 1 = Did occur during followup.

**ANYCHD** Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease. 0 = Did not occur during followup, 1 = Did occur during followup.

**STROKE** Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease. 0 = Did not occur during followup, 1 = Did occur during followup.

**CVD** Myocardial infarction (Hospitalized and silent or unrecognized), Fatal Coronary Heart Disease, Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease. 0 = Did not occur during followup, 1 = Did occur during followup.

**HYPERTEN** Hypertensive. Defined as the first exam treated for high blood pressure or second exam in which either Systolic is 6 140 mmHg or Diastolic 6 90mmHg. 0 = Did not occur during followup, 1 = Did occur during followup.

**cvd_dth** Cardiovascular disease OR death. 0 = Did not occur during followup, 1 = Did occur during followup.

**timeout** Number of days from the start of the study to cardiovascular disease, death, or loss to follow-up.

**drop** Participant was lost to follow-up before 24 months complete followup. 0 = no, 1 = yes

**glucoseyear6** Casual serum glucose (mg/dL) after 6 years of follow-up

**logpdays** Natural log of `timeout`.

**bmicat** BMI category. 0 = Normal, 1 = Underweight, 2 = Overweight, 3 = Obese.

**nhosp** Simulated number of hospitalizations over 24 months, associated with age, sex, BMI, and diabetes (not collected in the Framingham study).

## Details

The National Heart, Lung, and Blood Institute of the National Institutes of Health developed a longitudinal, epidemiology-focused dataset using the Framingham Heart Study. The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1948 and 5,209 subjects were initially enrolled in the study. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. Clinic examination data has included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, Echocardiography, and medication use. Through regular surveillance of area hospitals, participant contact, and death certificates, the Framingham Heart Study reviews and adjudicates events for the occurrence of Angina Pectoris, Myocardial Infarction, Heart Failure, and Cerebrovascular disease. This dataset contains three clinic examinations and 20 year follow-up data on a large subset of the original Framingham cohort participants.

NOTE: This is a "teaching" dataset. Specific methods were employed to ensure an anonymous dataset that protects patient confidentiality; therefore, this dataset is inappropriate for publication purposes." The use of these data for the purposes of this package were approved on 11Mar2019 (request #7161) by NIH/NHLBI.

## Source

https://biolincc.nhlbi.nih.gov/teaching/

---

`framingham` *The* `framingham` *data set*

---

## Description

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1948 and 5,209 subjects were initially enrolled in the study. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. Clinic examination data has included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, Echocardiography, and medication use. Through regular surveillance of area hospitals, participant contact, and death certificates, the Framingham Heart Study reviews and adjudicates events for the occurrence of Angina Pectoris, Myocardial Infarction, Heart Failure, and Cerebrovascular disease. The enclosed dataset is a subset of the data collected as part of the Framingham study and includes laboratory, clinic, questionnaire, and adjudicated event data on 4,434 participants. Participant clinic data was collected during three examination periods, approximately 6 years apart, from roughly 1956 to 1968. Each participant was followed for a total of 24 years for the outcome of the following events: Angina Pectoris, Myocardial Infarction, Atherothrombotic Infarction or Cerebral Hemorrhage (Stroke) or death.

## Usage

```
data(framingham)
```

## Format

A data frame with 11627 rows and 39 variables:

**RANDID** Unique identification number for each participant. Values range from 2448-999312.

**SEX** Participant sex. 1 = Male (n = 5022), 2 = Female (n = 6605).

**TOTCHOL** Serum Total Cholesterol (mg/dL). Values range from 107-696.

**AGE** Age at exam (years). Values range from 32-81.

**SYSBP** Systolic Blood Pressure (mean of last two of three measurements) (mmHg). Values range from 83.5-295.

**DIABP** Diastolic Blood Pressure (mean of last two of three measurements) (mmHg). Values range from 30-150.

**CURSMOKE** Current cigarette smoking at exam. 0 = Not current smoker (n = 6598), 1 = Current smoker (n = 5029).

**CIGPDAY** Number of cigarettes smoked each day. 0 = Not current smoker. Values range from 0-90 cigarettes per day.

**BMI** Body Mass Index, weight in kilograms/height meters squared. Values range from 14.43-56.8.

**DIABETES** Diabetic according to criteria of first exam treated or first exam with casual glucose of 200 mg/dL or more. 0 = Not a diabetic (n = 11097), 1 = Diabetic (n = 530)

**BPMEDS** Use of Anti-hypertensive medication at exam. 0 = Not currently used (n = 10090), 1 = Current use (n = 944).

**HEARTRTE** Heart rate (Ventricular rate) in beats/min. Values range from 37-220.

**GLUCOSE** Casual serum glucose (mg/dL). Values range from 39-478.

**educ**

**PREVCHD** Prevalent Coronary Heart Disease defined as pre-existing Angina Pectoris, Myocardial Infarction (hospitalized, silent or unrecognized), or Coronary Insufficiency (unstable angina). 0 = Free of disease (n = 10785), 1 = Prevalent disease (n = 842).

**PREVAP** Prevalent Angina Pectoris at exam. 0 = Free of disease (n = 11000), 1 = Prevalent disease (n = 627).

**PREVMI** Prevalent Myocardial Infarction. 0 = Free of disease (n = 11253), 1 = Prevalent disease (n = 374).

**PREVSTRK** Prevalent Stroke. 0 = Free of disease (n = 11475), 1 = Prevalent disease (n = 152).

**PREVHYP** Prevalent Hypertensive. Subject was defined as hypertensive if treated or if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg. 0 = Free of disease (n = 6283), 1 = Prevalent disease (n = 5344).

**TIME** Number of days since baseline exam. Values range from 0-4854

**PERIOD** Examination Cycle. 1 = Period 1 (n = 4434), 2 = Period 2 (n = 3930), 3 = Period 3 (n = 3263)

**HDLC** High Density Lipoprotein Cholesterol (mg/dL). Available for Period 3 only. Values range from 10-189.

**LDLC** Low Density Lipoprotein Cholesterol (mg/dL). Available for Period 3 only. Values range from 20-565.

**DEATH** Death from any cause. 0 = Did not occur during followup, 1 = Did occur during followup.

**ANGINA** Angina Pectoris. 0 = Did not occur during followup, 1 = Did occur during followup.

**HOSPMI** Hospitalized Myocardial Infarction. 0 = Did not occur during followup, 1 = Did occur during followup.

**MI_FCHD** Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease. 0 = Did not occur during followup, 1 = Did occur during followup.

**ANYCHD** Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease. 0 = Did not occur during followup, 1 = Did occur during followup.

**STROKE** Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease. 0 = Did not occur during followup, 1 = Did occur during followup.

**CVD** Myocardial infarction (Hospitalized and silent or unrecognized), Fatal Coronary Heart Disease, Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease. 0 = Did not occur during followup, 1 = Did occur during followup.

**HYPERTEN** Hypertensive. Defined as the first exam treated for high blood pressure or second exam in which either Systolic is 6 140 mmHg or Diastolic 6 90mmHg. 0 = Did not occur during followup, 1 = Did occur during followup.

**TIMEAP** Number of days from Baseline exam to first Angina during the followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup.

**TIMEMI** Number of days from Baseline exam to first HOSPMI event during followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup.

**TIMEMIFC** Number of days from Baseline exam to first MI_FCHD event during followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup.

**TIMECHD** Number of days from Baseline exam to first ANYCHD event during followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup.

**TIMESTRK** Number of days from Baseline exam to first STROKE event during followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup.

**TIMECVD** Number of days from Baseline exam to first CVD event during followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup.

**TIMEDTH** Number of days from Baseline exam to death if occurring during followup or Number of days from Baseline to censor date. Censor date may be end of followup, or last known contact date if subject is lost to followup.

**TIMEHYP** Number of days from Baseline exam to first HYPERTEN event during followup or Number of days from Baseline to censor date. Censor date may be end of followup, death or last known contact date if subject is lost to followup.

## Details

This dataset is the teaching dataset from the Framingham Heart Study (No. N01-HC-25195), provided with permission from #' the National Heart, Lung, and Blood Institute (NHLBI). The Framingham Heart Study is conducted and supported by the NHLBI in collaboration with Boston University. This package was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

---

`gComp` *Estimate difference and ratio effects with 95% confidence intervals.*

---

## Description

Obtain a point estimate and 95% confidence interval for difference and ratio effects comparing exposed and unexposed (or treatment and non-treatment) groups using g-computation.

## Usage

```
gComp(
  data,
 outcome.type = c("binary", "count", "count_nb", "rate", "rate_nb", "continuous"),
  formula = NULL,
  Y = NULL,
  X = NULL,
  Z = NULL,
  subgroup = NULL,
  offset = NULL,
  rate.multiplier = 1,
  exposure.scalar = 1,
  R = 200,
  clusterID = NULL,
  parallel = "no",
  ncpus = getOption("boot.ncpus", 1L)
)
```

## Arguments

| | |
|---|---|
| `data` | (Required) A data.frame containing variables for `Y`, `X`, and `Z` or with variables matching the model variables specified in a user-supplied formula. Data set should also contain variables for the optional `subgroup` and `offset`, if they are specified. |

`outcome.type` (Required) Character argument to describe the outcome type. Acceptable responses, and the corresponding error distribution and link function used in the `glm`, include:

**binary** (Default) A binomial distribution with link = 'logit' is used.

**count** A Poisson distribution with link = 'log' is used.

**count_nb** A negative binomial model with link = 'log' is used, where the theta parameter is estimated internally; ideal for over-dispersed count data.

**rate** A Poisson distribution with link = 'log' is used; ideal for events/person-time outcomes.

**rate_nb** A negative binomial model with link = 'log' is used, where the theta parameter is estimated internally; ideal for over-dispersed events/person-time outcomes.

**continuous** A gaussian distribution with link = 'identity' is used.

`formula` (Optional) Default NULL. An object of class "formula" (or one that can be coerced to that class) which provides the the complete model formula, similar to the formula for the glm function in R (e.g. 'Y ~ X + Z1 + Z2 + Z3'). Can be supplied as a character or formula object. If no formula is provided, Y and X must be provided.

`Y` (Optional) Default NULL. Character argument which specifies the outcome variable. Can optionally provide a formula instead of `Y` and `X` variables.

`X` (Optional) Default NULL. Character argument which specifies the exposure variable (or treatment group assignment), which can be binary, categorical, or continuous. This variable can be supplied as a factor variable (for binary or categorical exposures) or a continuous variable. For binary/categorical exposures, `X` should be supplied as a factor with the lowest level set to the desired referent. Numeric variables are accepted, but will be centered (see Note). Character variables are not accepted and will throw an error. Can optionally provide a formula instead of `Y` and `X` variables.

`Z` (Optional) Default NULL. List or single character vector which specifies the names of covariates or other variables to adjust for in the `glm` function. All variables should either be factors, continuous, or coded 0/1 (i.e. not character variables). Does not allow interaction terms.

`subgroup` (Optional) Default NULL. Character argument that indicates subgroups for stratified analysis. Effects will be reported for each category of the subgroup variable. Variable will be automatically converted to a factor if not already.

`offset` (Optional, only applicable for rate/count outcomes) Default NULL. Character argument which specifies the variable name to be used as the person-time denominator for rate outcomes to be included as an offset in the Poisson regression model. Numeric variable should be on the linear scale; function will take natural log before including in the model.

`rate.multiplier`

(Optional, only applicable for rate/count outcomes). Default 1. Numeric variable signifying the person-time value to use in predictions; the offset variable will be set to this when predicting under the counterfactual conditions. This value should be set to the person-time denominator desired for the rate difference measure and must be inputted in the units of the original offset variable

(e.g. if the offset variable is in days and the desired rate difference is the rate per 100 person-years, rate.multiplier should be inputted as 365.25*100).

`exposure.scalar`
: (Optional, only applicable for continuous exposure) Default 1. Numeric value to scale effects with a continuous exposure. This option facilitates reporting effects for an interpretable contrast (i.e. magnitude of difference) within the continuous exposure. For example, if the continuous exposure is age in years, a multiplier of 10 would result in estimates per 10-year increase in age rather than per a 1-year increase in age.

`R`
: (Optional) Default 200. The number of data resamples to be conducted to produce the bootstrap confidence interval of the estimate.

`clusterID`
: (Optional) Default NULL. Character argument which specifies the variable name for the unique identifier for clusters. This option specifies that clustering should be accounted for in the calculation of confidence intervals. The `clusterID` will be used as the level for resampling in the bootstrap procedure.

`parallel`
: (Optional) Default "no." The type of parallel operation to be used. Available options (besides the default of no parallel processing) include "multicore" (not available for Windows) or "snow." This argument is passed directly to `boot`. See note below about setting seeds and parallel computing.

`ncpus`
: (Optional, only used if parallel is set to "multicore" or "snow") Default 1. Integer argument for the number of CPUs available for parallel processing/ number of parallel operations to be used. This argument is passed directly to `boot`

## Details

The `gComp` function executes the following steps:

1. Calls the `pointEstimate` function on the data to obtain the appropriate effect estimates (difference, ratio, etc.).
2. Generates `R` bootstrap resamples of the data, with replacement. If the resampling is to be done at the cluster level (set using the `clusterID` argument), the number of clusters will remain constant but the total number of observations in each resampled data set might be different if clusters are not balanced.
3. Calls the `pointEstimate` function on each of the resampled data sets.
4. Calculates the 95% confidence interval of the difference and ratio estimates using the results obtained from the `R` resampled parameter estimates.

As bootstrap resamples are generated with random sampling, users should set a seed (`set.seed` for reproducible confidence intervals.

While offsets are used to account for differences in follow-up time between individuals in the `glm` model, rate differences are calculated assuming equivalent follow-up of all individuals (i.e. predictions for each exposure are based on all observations having the same offset value). The default is 1 (specifying 1 unit of the original offset variable) or the user can specify an offset to be used in the predictions with the rate.multiplier argument.

## Value

An object of class gComp which is a named list with components:

| | |
|---|---|
| $summary | Summary providing parameter estimates and 95% confidence limits of the outcome difference and ratio (in a print-pretty format) |
| $results.df | Data.frame with parameter estimates, 2.5% confidence limit, and 97.5% confidence limit each as a column (which can be used for easy incorporation into tables for publication) |
| $n | Number of unique observations in the original dataset |
| $R | Number of bootstrap iterations |
| $boot.result | Data.frame containing the results of the R bootstrap iterations of the g-computation |
| $contrast | Contrast levels compared |
| $family | Error distribution used in the model |
| $formula | Model formula used to fit the glm |
| $predicted.outcome | A data.frame with the marginal mean predicted outcomes (with 95% confidence limits) for each exposure level (i.e. under both exposed and unexposed counterfactual predictions) |
| $glm.result | The glm class object returned from the fitted regression of the outcome on the exposure and relevant covariates. |

## Note

Note that for a protective exposure (risk difference less than 0), the 'Number needed to treat/harm' is interpreted as the number needed to treat, and for a harmful exposure (risk difference greater than 0), it is interpreted as the number needed to harm. Note also that confidence intervals are not reported for the number needed to treat/harm. If the confidence interval (CI) for the risk difference crosses the null, the construction of the CI for the number needed to treat/harm is not well defined. Challenges and options for reporting the number needed to treat/harm CI are reviewed extensively in Altman 1998, Hutton 2000, and Stang 2010, with a consensus that an appropriate interval would have two segments, one bounded at negative infinity and the other at positive infinity. Because the number needed to treat/harm is most useful as a communication tool and is directly derived from the risk difference, which has a CI that provides a more interpretable measure of precision, we do not report the CI for the number needed to treat/harm. If the CI of the risk difference does not cross the null, the number needed to treat/harm CI can be calculated straightforwardly by taking the inverse of each confidence bound of the risk difference.

For continuous exposure variables, the default effects are provided for a one unit difference in the exposure at the mean value of the exposure variable. Because the underlying parametric model for a binary outcome is logistic regression, the risks for a continuous exposure will be estimated to be linear on the log-odds (logit) scale, such that the odds ratio for any one unit increase in the continuous variable is constant. However, the risks will not be linear on the linear (risk difference) or log (risk ratio) scales, such that these parameters will not be constant across the range of the continuous exposure. Users should be aware that the risk difference, risk ratio, number needed to treat/harm (for a binary outcome) and the incidence rate difference (for a rate/count outcome) reported with a continuous exposure apply specifically at the mean of the continuous exposure. The

effects do not necessarily apply across the entire range of the variable. However, variations in the effect are likely small, especially near the mean.

Interaction terms are not allowed in the model formula. The `subgroup` argument affords interaction between the exposure variable and a single covariate (that is forced to categorical if supplied as numeric) to estimate effects of the exposure within subgroups defined by the interacting covariate. To include additional interaction terms with variables other than the exposure, we recommend that users create the interaction term as a cross-product of the two interaction variables in a data cleaning step prior to running the model.

The documentation for `boot` includes details about reproducible seeds when using parallel computing.

## References

Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. Am. J. Epidemiol. 2009;169(9):1140–1147. doi:10.1093/aje/kwp015

Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common. BMJ. 1998;317(7168):1318. doi:10.1136/bmj.317.7168.1318

Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC. Book link

Hutton JL. Number needed to treat: properties and problems. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2000;163(3):381–402. doi:10.1111/1467985X.00175

Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling. 1986;7(9):1393–1512. doi:10.1016/02700255(86)900886

Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am. J. Epidemiol. 2011;173(7):731–738. doi:10.1093/aje/kwq472

Stang A, Poole C, Bender R. Common problems related to the use of number needed to treat. Journal of Clinical Epidemiology. 2010;63(8):820–825. doi:10.1016/j.jclinepi.2009.08.006

Westreich D, Cole SR, Young JG, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. Stat Med. 2012;31(18):2000–2009. doi:10.1002/sim.5316

## See Also

`pointEstimate` `boot`

## Examples

```
## Obtain the risk difference and risk ratio for cardiovascular disease or death between
## patients with and without diabetes.
data(cvdd)
set.seed(538)
diabetes <- gComp(cvdd, formula = "cvd_dth ~ DIABETES + AGE + SEX + BMI + CURSMOKE + PREVHYP",
outcome.type = "binary", R = 20)
```

get_results_dataframe *Take predicted dataframe and calculate the outcome (risk difference/ratio, incidence rate difference/ratio, mean difference, and/or number needed to treat)*

## Description

Take predicted dataframe and calculate the outcome (risk difference/ratio, incidence rate difference/ratio, mean difference, and/or number needed to treat)

## Usage

```
get_results_dataframe(predict.df, outcome.type)
```

## Arguments

| | |
|---|---|
| predict.df | (Required) A data.frame output from the make_predict_df function with predicted outcome for each observation at each level of treatment/exposure. |
| outcome.type | (Required) Character argument to describe the outcome type. Acceptable responses, and the corresponding error distribution and link function used in the glm, include: |

**binary** (Default) A binomial distribution with link = 'logit' is used.

**count** A Poisson distribution with link = 'log' is used.

**count_nb** A negative binomial model with link = 'log' is used, where the theta parameter is estimated internally; ideal for over-dispersed count data.

**rate** A Poisson distribution with link = 'log' is used; ideal for events/person-time outcomes.

**rate_nb** A negative binomial model with link = 'log' is used, where the theta parameter is estimated internally; ideal for over-dispersed events/person-time outcomes.

**continuous** A gaussian distribution with link = 'identity' is used.

## Value

A list containing the calculated results for the applicable measures (based on the outcome.type): Risk Difference, Risk Ratio, Odds Ratio, Incidence Risk Difference, Incidence Risk Ratio, Mean Difference, Number Needed to Treat, Average Tx (average predicted outcome of all observations with treatment/exposure), and Average noTx (average predicted outcome of all observations without treatment/exposure)

make_predict_df *Using* glm *results, predict outcomes for each individual at each level of treatment/exposure*

## Description

Using glm results, predict outcomes for each individual at each level of treatment/exposure

## Usage

```
make_predict_df(
  glm.res,
  df,
  X,
  subgroup = NULL,
  offset = NULL,
  rate.multiplier = 1
)
```

## Arguments

| | |
|---|---|
| glm.res | (Required) A fitted object of class inheriting from "glm" that will be used with new dataset for prediciton. |
| df | (Required) A new data frame in which to look for variables with which to predict. This is equivalent to the newdata argument in predict.glm. |
| X | (Required) Character argument which provides variable identifying exposure/treatment group assignment. |
| subgroup | (Optional) Default NULL. Character argument of the variable name to use for subgroup analyses. Variable automatically transformed to a factor within the function if not supplied as such. |
| offset | (Optional, only applicable for rate/count outcomes) Default NULL. Character argument which specifies the variable name to be used as the person-time denominator for rate outcomes to be included as an offset in the Poisson regression model. Numeric variable should be on the linear scale; function will take natural log before including in the model. |
| rate.multiplier | (Optional, only applicable for rate/count outcomes). Default 1 Numeric variable signifying the person-time value to use in predictions; the offset variable will be set to this when predicting under the counterfactual conditions. This value should be set to the person-time denominator desired for the rate difference measure and must be inputted in the units of the original offset variable (e.g. if the offset variable is in days and the desired rate difference is the rate per 100 person-years, rate.multiplier should be inputted as 365.25*100). |

## Value

A data.frame of predicted outcomes for each level of treatment/exposure. Additional columns are provided for each subgroup *x*treatment, if specified.

---

`plot.gComp` *Plot estimates of difference and ratio effects obtained in the bootstrap computations of the g-computation*

---

## Description

Plot histograms and Q-Q plots for each the difference and ratio estimates

## Usage

```
## S3 method for class 'gComp'
plot(x, ...)
```

## Arguments

| | |
|---|---|
| `x` | (Required) An object of class `gComp`. |
| `...` | (Optional) additional arguments to be supplied to the ‘`geom_histogram` call (e.g. to adjust binwidth for histogram, assign colors, etc.). |

## Value

a plot containing histograms and Q-Q plots of the difference and ratio estimates returned from R bootstrap iterations

## Examples

```
## Obtain the risk difference and risk ratio for cardiovascular disease or death
## between patients with and without diabetes, while controlling for
## age,
## sex,
## BMI,
## whether the individual is currently a smoker, and
## if they have a history of hypertension.
data(cvdd)
set.seed(58)
diabetes.result <- gComp(data = cvdd, Y = "cvd_dth", X = "DIABETES",
Z = c("AGE", "SEX", "BMI", "CURSMOKE", "PREVHYP"), outcome.type = "binary", R = 60)
plot(diabetes.result)
```

`pointEstimate` *Perform g-computation to estimate difference and ratio effects of an exposure*

## Description

Generate a point estimate of the outcome difference and ratio using G-computation

## Usage

```
pointEstimate(
  data,
 outcome.type = c("binary", "count", "count_nb", "rate", "rate_nb", "continuous"),
  formula = NULL,
  Y = NULL,
  X = NULL,
  Z = NULL,
  subgroup = NULL,
  offset = NULL,
  rate.multiplier = 1,
  exposure.scalar = 1,
  exposure.center = TRUE
)
```

## Arguments

`data` (Required) A data.frame containing variables for `Y`, `X`, and `Z` or with variables matching the model variables specified in a user-supplied formula. Data set should also contain variables for the optional `subgroup` and `offset`, if they are specified.

`outcome.type` (Required) Character argument to describe the outcome type. Acceptable responses, and the corresponding error distribution and link function used in the `glm`, include:

**binary** (Default) A binomial distribution with link = 'logit' is used.

**count** A Poisson distribution with link = 'log' is used.

**count_nb** A negative binomial model with link = 'log' is used, where the theta parameter is estimated internally; ideal for over-dispersed count data.

**rate** A Poisson distribution with link = 'log' is used; ideal for events/person-time outcomes.

**rate_nb** A negative binomial model with link = 'log' is used, where the theta parameter is estimated internally; ideal for over-dispersed events/person-time outcomes.

**continuous** A gaussian distribution with link = 'identity' is used.

`formula` (Optional) Default NULL. An object of class "formula" (or one that can be coerced to that class) which provides the the complete model formula, similar to

the formula for the glm function in R (e.g. 'Y ~ X + Z1 + Z2 + Z3'). Can be supplied as a character or formula object. If no formula is provided, Y and X must be provided.

`Y` (Optional) Default NULL. Character argument which specifies the outcome variable. Can optionally provide a formula instead of `Y` and `X` variables.

`X` (Optional) Default NULL. Character argument which specifies the exposure variable (or treatment group assignment), which can be binary, categorical, or continuous. This variable can be supplied as a factor variable (for binary or categorical exposures) or a continuous variable. For binary/categorical exposures, `X` should be supplied as a factor with the lowest level set to the desired referent. Numeric variables are accepted, but will be centered (see Note). Character variables are not accepted and will throw an error. Can optionally provide a formula instead of `Y` and `X` variables.

`Z` (Optional) Default NULL. List or single character vector which specifies the names of covariates or other variables to adjust for in the `glm` function. All variables should either be factors, continuous, or coded 0/1 (i.e. not character variables). Does not allow interaction terms.

`subgroup` (Optional) Default NULL. Character argument that indicates subgroups for stratified analysis. Effects will be reported for each category of the subgroup variable. Variable will be automatically converted to a factor if not already.

`offset` (Optional, only applicable for rate/count outcomes) Default NULL. Character argument which specifies the variable name to be used as the person-time denominator for rate outcomes to be included as an offset in the Poisson regression model. Numeric variable should be on the linear scale; function will take natural log before including in the model.

`rate.multiplier`
(Optional, only applicable for rate/count outcomes). Default 1. Numeric variable signifying the person-time value to use in predictions; the offset variable will be set to this when predicting under the counterfactual conditions. This value should be set to the person-time denominator desired for the rate difference measure and must be inputted in the units of the original offset variable (e.g. if the offset variable is in days and the desired rate difference is the rate per 100 person-years, rate.multiplier should be inputted as 365.25*100).

`exposure.scalar`
(Optional, only applicable for continuous exposure) Default 1. Numeric value to scale effects with a continuous exposure. This option facilitates reporting effects for an interpretable contrast (i.e. magnitude of difference) within the continuous exposure. For example, if the continuous exposure is age in years, a multiplier of 10 would result in estimates per 10-year increase in age rather than per a 1-year increase in age.

`exposure.center`
(Optional, only applicable for continuous exposure) Default TRUE. Logical or numeric value to center a continuous exposure. This option facilitates reporting effects at the mean value of the exposure variable, and allows for a mean value to be provided directly to the function in cases where bootstrap resampling is being conducted and a standardized centering value should be used across all

bootstraps. See note below on continuous exposure variables for additional details.

## Details

The `pointEstimate` function executes the following steps on the data:

1. Fit a regression of the outcome on the exposure and relevant covariates, using the provided data set.
2. Using the model fit in step 1, predict counterfactuals (e.g. calculate predicted outcomes for each observation in the data set under each level of the treatment/exposure).
3. Estimate the marginal difference/ratio of treatment effect by taking the difference or ratio of the average of all observations under the treatment/no treatment regimes.

As counterfactual predictions are generated with random sampling of the distribution, users should set a seed (`set.seed`) prior to calling the function for reproducible confidence intervals.

## Value

A named list containing the following:

`$parameter.estimates`
: Point estimates for the risk difference, risk ratio, odds ratio, incidence rate difference, incidence rate ratio, mean difference and/or number needed to treat/harm, depending on the outcome.type

`$formula`
: Model formula used to fit the `glm`

`$contrast`
: Contrast levels compared

`$Y`
: The response variable

`$covariates`
: Covariates used in the model

`$n`
: Number of observations provided to the model

`$family`
: Error distribution used in the model

`$predicted.data`
: A data.frame with the predicted values for the exposed and unexposed counterfactual predictions for each observation in the original dataset (on the log scale)

`$predicted.outcome`
: A data.frame with the marginal mean predicted outcomes for each exposure level

`$glm.result`
: The `glm` class object returned from the fitted regression of the outcome on the exposure and relevant covariates.

formula = formula,

## Note

While offsets are used to account for differences in follow-up time between individuals in the `glm` model, rate differences are calculated assuming equivalent follow-up of all individuals (i.e. predictions for each exposure are based on all observations having the same offset value). The default is 1 (specifying 1 unit of the original offset variable) or the user can specify an offset to be used in the predictions with the rate.multiplier argument.

Note that for a protective exposure (risk difference less than 0), the 'Number needed to treat/harm' is interpreted as the number needed to treat, and for a harmful exposure (risk difference greater than 0), it is interpreted as the number needed to harm.

For continuous exposure variables, the default effects are provided for a one unit difference in the exposure at the mean value of the exposure variable. Because the underlying parametric model for a binary outcome is logistic regression, the risks for a continuous exposure will be estimated to be linear on the log-odds (logit) scale, such that the odds ratio for any one unit increase in the continuous variable is constant. However, the risks will not be linear on the linear (risk difference) or log (risk ratio) scales, such that these parameters will not be constant across the range of the continuous exposure. Users should be aware that the risk difference, risk ratio, number needed to treat/harm (for a binary outcome) and the incidence rate difference (for a rate/count outcome) reported with a continuous exposure apply specifically at the mean of the continuous exposure. The effects do not necessarily apply across the entire range of the variable. However, variations in the effect are likely small, especially near the mean.

@note Interaction terms are not allowed in the model formula. The `subgroup` argument affords interaction between the exposure variable and a single covariate (that is forced to categorical if supplied as numeric) to estimate effects of the exposure within subgroups defined by the interacting covariate. To include additional interaction terms with variables other than the exposure, we recommend that users create the interaction term as a cross-product of the two interaction variables in a data cleaning step prior to running the model.

@note For negative binomial models, `MASS::glm.nb` is used instead of the standard `stats::glm` function used for all other models.

## References

Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. Am. J. Epidemiol. 2009;169(9):1140–1147. doi:10.1093/aje/kwp015

Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common. BMJ. 1998;317(7168):1318. doi:10.1136/bmj.317.7168.1318

Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC. Book link

Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling. 1986;7(9):1393–1512. doi:10.1016/02700255(86)900886

Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am. J. Epidemiol. 2011;173(7):731–738. doi:10.1093/aje/kwq472

Westreich D, Cole SR, Young JG, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. Stat Med. 2012;31(18):2000–2009. doi:10.1002/sim.5316

## See Also

gComp

## Examples

```
## Obtain the risk difference and risk ratio for cardiovascular disease or death
## between patients with and without diabetes, while controlling for
## age,
## sex,
## BMI,
## whether the individual is currently a smoker, and
## if they have a history of hypertension.
data(cvdd)
ptEstimate <- pointEstimate(data = cvdd, Y = "cvd_dth", X = "DIABETES",
Z = c("AGE", "SEX", "BMI", "CURSMOKE", "PREVHYP"), outcome.type = "binary")
```

---

# print.gComp *Print estimates of difference and ratio effects obtained in the bootstrap computations of the g-computation*

---

## Description

Print results from bootstrap computations of the g-computation

## Usage

```
## S3 method for class 'gComp'
print(x, ...)
```

## Arguments

| | |
|---|---|
| x | (Required) An object of class gComp as produced by gComp(). |
| ... | (Optional) Further arguments passed to or from other methods. |

## Value

Returns the formula and resulting point estimate and 95% confidence intervals of the difference and ratio.

## Examples

```
## Obtain the risk difference and risk ratio for cardiovascular disease or
## death between patients with and without diabetes, while controlling for
## age, sex, BMI, whether the individual is currently a smoker, and
## if they have a history of hypertension.
data(cvdd)
set.seed(4832)
diabetes.result <- gComp(cvdd,
   formula = "cvd_dth ~ DIABETES + AGE + SEX + BMI + CURSMOKE + PREVHYP",
   outcome.type = "binary", R = 20)
print(diabetes.result)
```

---

`riskCommunicator` *riskCommunicator: Obtaining interpretable epidemiological effect estimates*

---

## Description

`riskCommunicator` is a package for estimating flexible epidemiological effect measures including both differences and ratios. The package is based on the parametric G-formula (g-computation with parametric models) developed by Robbins et. al. in 1986 as an alternative to inverse probability weighting. It is useful for estimating the impact of interventions in the presence of treatment-confounder-feedback and is a powerful tool for causal inference, but has seen limited success due to lack of software for the computationally intensive components. This package provides three main functions. The first, `pointEstimate`, obtains a point estimate of the difference and ratio effect estimates. This function is typically called within the `gComp` function, but is available for use in special cases for example when the user requires more explicit control over bootstrap resampling (e.g. nested clusters). The second function, `gComp`, is the workhorse function that obtains point estimates for difference and ratio effects along with their 95/ to visualize the bootstrap results. We provide the `framingham` dataset, which is the teaching dataset from the Framingham Heart Study, as well as a subset of that data, `cvdd` for users.

## References

Robins, James. 1986. "A New Approach To Causal Inference in Mortality Studies with a Sustained Exposure Period - Application To Control of the Healthy Worker Survivor Effect." Mathematical Modelling 7: 1393–1512. doi:10.1016/0270-0255(86)90088-6.

## See Also

`gComp`

`pointEstimate`

`plot.gComp`

summary.gComp *Print a summary of a gComp class object.*

## Description

Takes a gComp object produced by gComp() and produces various useful summaries from it.

## Usage

```
## S3 method for class 'gComp'
summary(object, ...)

## S3 method for class 'summary.gComp'
print(x, ...)
```

## Arguments

| | |
|---|---|
| object | (Required) An object of class gComp as produced by gComp(). |
| ... | Further arguments passed to or from other methods. |
| x | (Required) An object of class summary.gComp as produced by summary.gComp(). |

## Value

Returns the formula, family (with link function), contrast evaluated, resulting point estimate and 95% confidence intervals of the parameters estimated, and the underlying glm used for model predictions.

## Examples

```
## Obtain the risk difference and risk ratio for cardiovascular disease or
## death between patients with and without diabetes, while controlling for
## age, sex, BMI, whether the individual is currently a smoker, and
## if they have a history of hypertension.
data(cvdd)
set.seed(4832)
diabetes.result <- gComp(cvdd,
   formula = "cvd_dth ~ DIABETES + AGE + SEX + BMI + CURSMOKE + PREVHYP",
   outcome.type = "binary", R = 20)
summary(diabetes.result)
```

# Index