Package 'samplesizelogisticcasecontrol'

February 4, 2017

Title Sample Size Calculations for Case-Control Studies

Version 0.0.6

Date 2017-01-31

Author Mitchell H. Gail

Description To determine sample size for case-control studies to be analyzed using logistic regression.

Maintainer William Wheeler <WheelerB@imsweb.com>

Depends mvtnorm

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2017-02-04 06:49:33

R topics documented:

`file.list` *List to describe the covariate and exposure data*

Description

The list to describe the covariate and exposure data for the `data` option.

Format

The format is: List of 7

file Data file containing the confounders and exposure variables. No default.

exposure Name or column number in `file` for the exposure variable. This can also be a vector giving the columns to form an interaction variable (see details). No default.

covars Character vector of variables names or numeric vector of column numbers in `file` that will be confounders. These variables must be numeric. The length and order of the `logOR` argument must match the length and order of c(`covars`, `exposure`). The default is NULL.

header 0 or 1 if `file` has the first row as variable names. The default is determined from the first line of the `file`.

delimiter The delimiter in `file`. The default is determined from the first two lines of the `file`.

in.miss Vector of character strings to define the missing values. This option corresponds to the option `na.strings` in `read.table()`. The default is "NA".

subsetData List of sublists to subset the data. Each sublist should contain the names "var", "operator" and "value" corresponding to a variable name, operator and values of the variable. Multiple sublists are logically connected by the AND operator. For example,
subsetData=list(list(var="GENDER", operator="==", value="MALE"))
will only include subjects with the string "MALE" for the GENDER variable.
subsetData=list(list(var="AGE", operator=">", value=50),
list(var="STUDY", operator="%in%", value=c("A", "B", "C")))
will include subjects with AGE > 50 AND in STUDY A, B or C. The default is NULL.

Details

In this list, `file` and `exposure` must be specified. If `exposure` is a vector of column names or column numbers, then an exposure variable will be created by multipling the columns defined in the vector to form the interaction variable. Thus, the columns must be numeric variables. In this case, the length and order of `logOR` must match the length and order of c(`covars`, <new interaction variable>).

`samplesizelogisticcasecontrol`

Sample size calculations for case-control studies

Description

This package can be used to calculate the required sample size needed for case-control studies to have sufficient power. To calculate the sample size, one needs to specify the significance level α, power γ, and the hypothesized non-null θ. Here θ is a log odds ratio for an exposure main effect or θ is an interaction effect on the logistic scale. Choosing θ requires subject matter knowledge to understand how strong the association needs to be to have practical importance. Sample size varies inversely with θ^2 and is thus highly dependent on θ.

Details

The main functions in the package are for different types of exposure variables, where the exposure variable is the variable of interest in a hypothesis test. The function `sampleSize_binary` can be used for a binary exposure variable (X = 0 or 1), while the function `sampleSize_ordinal` is a more general function that can be used for an ordinal exposure variable (X takes the values 0, 1, ..., k). `sampleSize_continuous` is useful for a continuous exposure variable and `sampleSize_data` can be used when pilot data is available that defines the distribution of the exposure and other confounding variables. Each function will return the sample sizes for a Wald-type test and a score test. When there are no adjustments for confounders, the user can specify a general distribution for the exposure variable. With confounders, either pilot data or a function to generate random samples from the multivariate distribution of the confounders and exposure variable must be given.

If the parameter of interest, θ, is one dimensional, then the test statistic is often asymptotically equivalent to a test of the form $T > Z_{1-\alpha}\sigma_0 n^{-\frac{1}{2}}$ or $T > Z_{1-\alpha}\sigma_\theta n^{-\frac{1}{2}}$, where $Z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution, n is the total sample size (cases plus controls), and $n^{\frac{1}{2}}T$ is normally distributed with mean 0 and null variance σ_0^2. Depending on which critical value $Z_{1-\alpha}\sigma_0 n^{-\frac{1}{2}}$ or $Z_{1-\alpha}\sigma_\theta n^{-\frac{1}{2}}$ of the test was used, the formulas for sample size are obtained by inverting the equations for power:

$n_1 = (Z_\gamma\sigma_\theta + Z_{1-\alpha}\sigma_0)^2/\theta^2$ or $n_2 = (Z_\gamma + Z_{1-\alpha})^2\sigma_\theta^2/\theta^2$.

Author(s)

Mitchell H. Gail <gailm@mail.nih.gov>

References

Gail, M.H. and Haneuse, S. Power and sample size for case-control studies. In Handbook of Statistical Methods for Case-Control Studies. Editors: Ornulf Borgan, Norman Breslow, Nilanjan Chatterjee, Mitchell Gail, Alastair Scott, Christopher Wild. Chapman and Hall/CRC, Taylor and Francis Group, New York, in press.

`sampleSize_binary` *Sample size for a binary exposure*

Description

Calculates the required sample size of as case-control study with a binary exposure variable

Usage

```
sampleSize_binary(prev, logOR, probXeq1=NULL, distF=NULL, data=NULL,
     size.2sided=0.05, power=0.9, cc.ratio=0.5, interval=c(-100, 100), tol=0.0001,
      n.samples=10000)
```

Arguments

prev	Number between 0 and 1 giving the prevalence of disease. No default.
logOR	Vector of ordered log-odds ratios for the confounders and exposure. The last log-odds ratio in the vector is for the exposure. If the option data (below) is specified, then the order must match the order of data. No default.
probXeq1	NULL or a number between 0 and 1 giving the probability that the exposure variable is 1. If set to NULL, the the data option must be specified so that probXeq1 can be estimated. The default is NULL.
distF	NULL, a function or a character string giving the function to generate random vectors from the distribution of the confounders and exposure. The order of the returned vector must match the order of logOR. User defined functions are also allowed, provided the user-defined function has only one integer valued argument that inputs the number of random vectors to generate. For instance the header of a user-defined function called "userF" would be userF <- function(n). The default depends on other options (see details).
data	NULL, matrix, data frame or a list of type file.list that gives a sample from the distribution of the confounders and exposure. If a matrix or data frame, then the last column consists of random values for the exposure, while the other columns are for the confounders. The order of the columns must match the order of the vector logOR. The default is NULL.
size.2sided	Number between 0 and 1 giving the size of the 2-sided hypothesis test. The default is 0.05.
power	Number between 0 and 1 for the desired power of the test. The default is 0.9.
cc.ratio	Number between 0 and 1 for the proportion of cases in the case-control sample. The default is 0.5.
interval	Two element vector giving the interval to search for the estimated intercept parameter. The default is c(-100, 100).
tol	Positive value giving the stopping tolerance for the root finding method to estimate the intercept parameter. The default is 0.0001.
n.samples	Integer giving the number of random vectors to generate when the option distF is specified. The default is 10000.

Details

If there are no confounders (length(logOR) = 1), then either probXeq1 or data must be specified, where probXeq1 takes precedance. If there are confounders (length(logOR) > 1), then either data or distF must be specified, where data takes precedance.

Value

A list containing four sample sizes, where two of them are for a Wald test and two for a score test. The two sample sizes for each test correspond to the equations for n_1 and n_2.

See Also

sampleSize_continuous, sampleSize_ordinal, sampleSize_data

Examples

```
prev  <- 0.01
logOR <- 0.3

# No confounders, Prob(X=1)=0.2
sampleSize_binary(prev, logOR, probXeq1=0.2)

# Generate data for a N(0,1) confounder and binary exposure
data <- cbind(rnorm(1000), rbinom(1000, 1, 0.4))
beta <- c(0.1, 0.2)
sampleSize_binary(prev, beta, data=data)

# Define a function to generate random vectors for two confounders and the binary exposure
f <- function(n) {cbind(rnorm(n), rbinom(n, 3, 0.5), rbinom(n, 1, 0.3))}
logOR <- c(0.2, 0.3, 0.25)
sampleSize_binary(prev, logOR, distF=f)
```

`sampleSize_continuous` *Sample size for a continuous exposure*

Description

Calculates the required sample size of as case-control study with a continuous exposure variable

Usage

```
sampleSize_continuous(prev, logOR, distF=NULL, distF.support=c(-Inf, Inf),
     data=NULL, size.2sided=0.05, power=0.9, cc.ratio=0.5, interval=c(-100, 100),
      tol=0.0001, n.samples=10000, distF.var=NULL)
```

Arguments

`prev`	Number between 0 and 1 giving the prevalence of disease. No default.
`logOR`	Vector of ordered log-odds ratios for the confounders and exposure. The last log-odds ratio in the vector is for the exposure. If the option `data` (below) is specified, then the order must match the order of `data`. No default.
`distF`	NULL, a function or a character string giving the pdf of the exposure variable for the case of no confounders, or giving the function to generate random vectors from the distribution formed by the confounders and exposure. For the case of no confounders, examples are `dnorm`, "dnorm(x, mean=0.5, sd=2.1)", "dbeta(?, shape1=0.3, shape2=3)", "dchisq(whatever, df=1)". Notice that when `distF` is a character string, the first argument can be anything but must be given to serve as a place holder. For the case of two confounders, an example might be a random vector generator from a multivariate normal distribution "rmvnorm(x, c(0,0,0))". User defined functions are also allowed, provided the user-defined function has only one input argument. The input argument would be a vector

of quantiles if the user-defined function is a pdf, or the input argument would be an integer specifiying the number of random vectors to generate if the user-defined function is a function to generate random vectors from the distribution of the confounders and exposure. An example pdf is the function H, where H <- function(x) { dunif(x, min=2, max=7) }. The default depends on other options (see details).

`distF.support` Two element vector giving the domain of `distF`. This option is only used when `distF` is a pdf. The default is c(-Inf, Inf).

`data` NULL, matrix, data frame or a list of type `file.list` that gives a sample from the distribution of the confounders and exposure. If a matrix or data frame, then the last column consists of random values for the exposure, while the other columns are for the confounders. The order of the columns must match the order of the vector `logOR`. The default is NULL.

`size.2sided` Number between 0 and 1 giving the size of the 2-sided hypothesis test. The default is 0.05.

`power` Number between 0 and 1 for the desired power of the test. The default is 0.9.

`cc.ratio` Number between 0 and 1 for the proportion of cases in the case-control sample. The default is 0.5.

`interval` Two element vector giving the interval to search for the estimated intercept parameter. The default is c(-100, 100).

`tol` Positive value giving the stopping tolerance for the root finding method to estimate the intercept parameter. The default is 0.0001.

`n.samples` Integer giving the number of random vectors to generate when the option `distF` is specified and is a random vector generation function. The default is 10000.

`distF.var` The variance of the exposure variable for the case of no confounders. This option is for efficiency purposes. If not specified, the variance will be estimated by either the empirical variance of a random sample from the distribution of the exposure or by numerical integration. The default is NULL.

Details

The `data` option takes precedance over the other options. If `data` is not specified, then the distribution of the exposure will be N(0,1) or MVN(0, 1) depending on whether there are confounders.

Value

A list containing four sample sizes, where two of them are for a Wald test and two for a score test. The two sample sizes for each test correspond to the equations for n_1 and n_2.

See Also

`sampleSize_binary`, `sampleSize_ordinal`, `sampleSize_data`

Examples

```
prev  <- 0.01
logOR <- 0.3

# No confounders, exposure assumed to be N(0,1)
sampleSize_continuous(prev, logOR)

# Two confounders and exposure assumed to be MVN(0,1)
beta <- c(0.1, 0.1, logOR)
sampleSize_continuous(prev, beta)

# No confounders, exposure is beta(0.3, 3)
sampleSize_continuous(prev, logOR, distF="dbeta(m, shape1=0.3, shape2=3)",
                      distF.support=c(0, 1))
```

`sampleSize_data` *Sample size using pilot data*

Description

Calculates the required sample size of a case-control study with pilot data

Usage

```
sampleSize_data(prev, logOR, data, size.2sided=0.05, power=0.9, cc.ratio=0.5,
        interval=c(-100, 100), tol=0.0001)
```

Arguments

`prev`	Number between 0 and 1 giving the prevalence of disease. No default.
`logOR`	Vector of ordered log-odds ratios for the confounders and exposure. The last log-odds ratio in the vector is for the exposure. If the option `data` (below) is specified, then the order must match the order of `data`. No default.
`data`	Matrix, data frame or a list of type `file.list` that gives a sample from the distribution of the confounders and exposure. If a matrix or data frame, then the last column consists of random values for the exposure, while the other columns are for the confounders. The order of the columns must match the order of the vector `logOR`. The default is NULL.
`size.2sided`	Number between 0 and 1 giving the size of the 2-sided hypothesis test. The default is 0.05.
`power`	Number between 0 and 1 for the desired power of the test. The default is 0.9.
`cc.ratio`	Number between 0 and 1 for the proportion of cases in the case-control sample. The default is 0.5.
`interval`	Two element vector giving the interval to search for the estimated intercept parameter. The default is c(-100, 100).
`tol`	Positive value giving the stopping tolerance for the root finding method to estimate the intercept parameter. The default is 0.0001.

Value

A list containing four sample sizes, where two of them are for a Wald test and two for a score test. The two sample sizes for each test correspond to the equations for n_1 and n_2.

See Also

sampleSize_binary, sampleSize_ordinal, sampleSize_continuous

Examples

```
prev  <- 0.01
logOR <- 0.3
data  <- matrix(rnorm(100, mean=1.5), ncol=1)

# Assuming exposuure is N(1.5, 1)
sampleSize_data(prev, logOR, data)
```

---

sampleSize_ordinal *Sample size for an ordinal exposure*

Description

Calculates the required sample size of as case-control study with an ordinal exposure variable

Usage

```
sampleSize_ordinal(prev, logOR, probX=NULL, distF=NULL, data=NULL,
     size.2sided=0.05, power=0.9, cc.ratio=0.5, interval=c(-100, 100), tol=0.0001,
      n.samples=10000)
```

Arguments

prev	Number between 0 and 1 giving the prevalence of disease. No default.
logOR	Vector of ordered log-odds ratios per category increase for the confounders and exposure. The last log-odds ratio in the vector is for the exposure. If the exposures are coded 0, 1, ..., k (k+1 categories), then the logOR corresponds to a one category increase. If the option data (below) is specified, then the order must match the order of data. No default.
probX	NULL or a vector that sums to 1 giving the probability that the exposure variable is equal to i, i = 0, 1, ..., k. If set to NULL, the the data option must be specified so that probX can be estimated. The default is NULL.
distF	NULL, a function or a character string giving the function to generate random vectors from the distribution of the confounders and exposure. The order of the returned vector must match the order of logOR. The default depends on other options (see details).

data	NULL, matrix, data frame or a list of type file.list that gives a sample from the distribution of the confounders and exposure. If a matrix or data frame, then the last column consists of random values for the exposure, while the other columns are for the confounders. The order of the columns must match the order of the vector logOR. The default is NULL.
size.2sided	Number between 0 and 1 giving the size of the 2-sided hypothesis test. The default is 0.05.
power	Number between 0 and 1 for the desired power of the test. The default is 0.9.
cc.ratio	Number between 0 and 1 for the proportion of cases in the case-control sample. The default is 0.5.
interval	Two element vector giving the interval to search for the estimated intercept parameter. The default is c(-100, 100).
tol	Positive value giving the stopping tolerance for the root finding method to estimate the intercept parameter. The default is 0.0001.
n.samples	Integer giving the number of random vectors to generate when the option distF is specified and is a random vector generation function. The default is 10000.

Details

If there are no confounders (length(logOR) = 1), then either probX or data must be specified, where probX takes precedance. If there are confounders (length(logOR) > 1), then either data or distF must be specified, where data takes precedance.

Value

A list containing four sample sizes, where two of them are for a Wald test and two for a score test. The two sample sizes for each test correspond to the equations for n_1 and n_2.

See Also

sampleSize_continuous, sampleSize_binary, sampleSize_data

Examples

```
prev  <- 0.01
logOR <- 0.3

# No confounders, Prob(X=1)=0.2
sampleSize_ordinal(prev, logOR, probX=c(0.8, 0.2))

# Generate data for a N(0,1) confounder and ordinal exposure with 3 levels
data <- cbind(rnorm(1000), rbinom(1000, 2, 0.5))
beta <- c(0.1, 0.2)
sampleSize_ordinal(prev, beta, data=data)

# Define a function to generate random vectors for two confounders and an ordinal
#   exposure with 5 levels
f <- function(n) {cbind(rnorm(n), rbinom(n, 1, 0.5), rbinom(n, 4, 0.5))}
beta <- c(0.2, 0.3, 0.25)
```

```
sampleSize_ordinal(prev, beta, distF=f)
```

Index