Package 'subdetect'

May 4, 2025

Type Package

Title Detect Subgroup with an Enhanced Treatment Effect

Version 1.3

Date 2025-05-04

Author Ailin Fan [aut],
Shannon T. Holloway [aut, cre]

Maintainer Shannon T. Holloway <shannon.t.holloway@gmail.com>

Description A test for the existence of a subgroup with enhanced treatment effect. And, a sample size calculation procedure for the subgroup detection test.

License GPL-3

Depends methods, stats

NeedsCompilation no

Repository CRAN

Date/Publication 2025-05-04 17:50:02 UTC

Contents

`sample_size` *Calculate Required Sample Size for the Test on Subgroup Existence.*

Description

Estimation of the required sample size for the test on subgroup existence. With a pre-specified significance level of the test α and a desired power $1 - \beta$ at a treatment effect τ, and other information about data, the required sample size that achieves power $1 - \beta$ can be estimated.

Usage

```
sample_size(outcome, theta0, sigma2, tau, N = 1000L, prob = 0.5,
            alpha = 0.05, power = 0.9, K = 1000L, M = 1000L, seed = NULL,
            precision = 0.01, ...)
```

Arguments

outcome	A formula object. The model of the indicator function. The model must include an intercept. Any lhs variable will be ignored.
theta0	A named numeric vector. The true parameters of the indicator model.
sigma2	A numeric object. The variance of the random error.
tau	A numeric object. The desired treatment effect.
N	An integer object. The number of random samples to draw. Default value is 1000.
prob	A numeric object. The probability of assigning individuals treatment 1. 0<=prob<=1.
alpha	A numeric object. The significance level of the test, 0<alpha<1.
power	A numeric object. The desired power of the test, 0<power<1
K	An integer object. The number of random sampled points on the unit ball surface $\{\theta : ||\theta||^2 = 1\}$. These randomly sampled points are used for approximating the Gaussian process in the null and local alternative distributions of the test statistic with multivariate normal distributions. It is recommended that K be set to 10^p, where p is the number of parameters in theta0. Note that it is recommended that the number of covariates be less than 10 for this implementation. Default value is 1000.
M	An integer object. The number of resamplings of the perturbed test statistic. This sample is used to calculate the critical value of the test. Default and minimum values are 1000.
seed	An integer object or NULL. If set, the seed for generating random values set at the onset of the calculation. If NULL, current seed in R environment is used.
precision	A numeric object. The precision tolerance for estimating the power from the calculated sample size. Specifically, the power of the sample size returned, P, will be (power - precision) < P < (power + precision).
...	For each covariate in outcome, user must provide a list object indicating the distribution function to be sampled and any parameters to be set when calling that function. Each list contains "FUN", the function name of the random generator of the distribution, and any formal arguments of that function. For example: x1 = list("FUN" = rnorm, sd = 2.0, mean = 10.0) x2 = list("FUN" = rbinom, size = 1, prob = 0.5) The number of points generated is determined by input N. Most distributions available through R's stats package can be used. Exceptions are: rhyper, rsignrank, rwilcox. Specifically, any random generator that passes the the number of observations to generate through formal argument 'n' can be used.

Details

The sample size calculation is based on the asymptotic null and local alternative distributions of the test statistic. More details can be found in the reference paper.

The difference between true baseline mean function and posited mean function, $\mu(X) - h(X, \beta_0)$, is set to zero when calculating the sample size.

Because the calculated sample size is based on simulated data following the null and local alternative distributions of the test statistic, the results can be different with different choices of `M` and `K`, as well as different seeds. When the signal `tau` to noise `sigma2` ratio is large, the calculated sample size can be more robust.

Value

A list consisting of

`n`	An integer object. The estimated sample size.
`power`	A numeric object. The estimated power.
`seed`	If seed was provided as input, the user specified integer seed. If seed was not provided, not present.

References

Ailin Fan, Rui Song, and Wenbin Lu, (2016). Change-plane analysis for subgroup detection and sample size calculation, Journal of the American Statistical Association, in press.

Examples

```
model <- ~ x1 + x2
theta0 <- c("(Intercept)" = 0.0, "x1" = 1.0, "x2" = 0.25)
sample_size(outcome = ~ x1 + x2,
            theta0 = theta0,
            N = 1000,
            sigma2 = 0.25,
            tau = 0.25,
            K = 100,
            M = 1000,
            x1 = list(FUN=runif, min = -1, max = 1),
            x2 = list(FUN=rnorm, mean = 0.0, sd = 0.75))
```

`subgroup_detect` *Test for and Identify a Subgroup with an Enhanced Treatment Effect.*

Description

Tests for the existence of a subgroup with an enhanced treatment effect. The subgroup of interest is represented by $\{\theta : \theta^T X \geq 0\}$. The test returns a p-value for $H_0 : \tau = 0$, where τ is the treatment effect in this subgroup. If H_0 is rejected, estimates for θ can be used to obtain the estimated subgroup.

Usage

```
subgroup_detect(outcome, propen, data,
                K = 1000L, M = 1000L, seed = NULL)
```

Arguments

outcome	A formula object. The linear model for the outcome regression. The left-hand-side variable must be the response. R function lm will be used to estimate model parameters. The response must be continuous.
propen	A formula object. The model for the propensity score. The left-hand-side variable must be the treatment variable. R function glm will be used with input option family = binomial(link="logit") to estimate model parameters. The treatment must be binary.
data	A data.frame object. All covariates, treatment, and response variables. Note that the treatment must be binary and that the response must be continuous.
K	An integer object. The number of random sampled points on the unit ball surface $\{\theta : ||\theta||^2 = 1\}$. These randomly sampled points are used for approximating the Gaussian process in the null and local alternative distributions of the test statistic with multivariate normal distributions. It is recommended that K be set to 10^p, where p is the number of parameters in the outcome model. Note that it is recommended that the number of covariates be less than 10 for this implementation. Default value is 1000.
M	An integer object. The number of resamplings of the perturbed test statistic. This sample is used to calculate the critical value of the test. Default and minimum values are 1000.
seed	An integer object or NULL. If integer, the seed for random number generation, set at the onset of the calculation. If NULL, current seed in R environment is used.

Details

In this function, a linear model with least squares estimate is used for fitting the baseline model $\mu(X)$, and a logistic model with maximum likelihood estimate is used for fitting the propensity score model $P(a = 1|X)$. These settings cannot be changed by the user.

Value

A list consisting of

outcome	An lm object. The object returned by the lm fit of the outcome.
propen	A glm object. The object returned by the glm fit of the propensity.
p_value	A numeric object. The p-value of the test.
theta	A named numeric vector. The change-plane parameter estimates for subgroup.
prop	A numeric object. The proportion of sampled points on θ unit ball surface that are used for calculating test statistic. For some values of $theta$, the subgroup contains no samples or all samples. These are discarded.

seed — If seed was provided as input, the user specified integer seed. If seed was not provided, not present.

References

Ailin Fan, Rui Song, and Wenbin Lu, (2016). Change-plane analysis for subgroup detection and sample size calculation, Journal of the American Statistical Association, in press.

Examples

```
#set parameters
tau <- 0.5
theta_t <- c(-0.15,0.3,sqrt(1-(-0.15)^2-(0.3)^2))
beta <- c(1,1,1)
sigma <- 0.5
n <- 50
p <- 2

#generate data
x1 <- rbinom(n,size=1,prob=0.5)
x2 <- runif(n,min=-1,max=1)
X <- cbind(1,x1,x2)
a <- rbinom(n,1,prob=0.5)
y <- drop(X%*%beta) + tau*a*(drop(X%*%theta_t)>=0) + rnorm(n,0,sigma)

data <- data.frame(X[,2:3], a, y)

subgroup_detect(outcome = y~x1+x2,
                propen = a~x1+x2,
                data = data)
```

Index