Package 'CBRM'

August 19, 2014

Title CBRM

Version 0.0.2

Date 2014-04-11

Author Minsun Song

Description
An R package for testing Calibration of Binary Risk Model (CBRM) using different goodness-of-fit statistics

Maintainer Minsun Song <minsun.song@nih.gov>

Depends MCMCpack

License GPL-2

R topics documented:

CBRM *CBRM*

Description

An R package for testing Calibration of Binary Risk Model (CBRM) using different goodness-of-fit statistics.

Details

The main function is `GOFR` which can fit alternative models to a given dataset and return corresponding goodness-of-fit test results. The function `getSummary` can be called for creating summary tables using the returned object from GOFR.

Author(s)

Minsun Song <minsun.song@nih.gov>

References

Song M, Kraft P, Joshi A D, Barrdahl M, and Chatterjee N (2014) Testing calibration of risk models at extremens of disease-risk, Biostatistics.

Hosmer D W and Lemeshow S (2000) Applied logistic regression. New York : John Wiley & Sons.

Windmeijer, F A G (1990) The asymptotic distribution of the sum of weighted squared residuals in binary choice models, Statistica Neerlandica 44(2), 69-78.

`getSummary` *getSummary*

Description

Returns summary information.

Usage

```
getSummary(fit)
```

Arguments

`fit`	The return object from GOFR.

Details

This function returns estimated parameters, standard errors, z values and p-values. It also returns goodness-of-fit test statistics and the corresponding p-values. In addition to that, it returns the goodness-of-fit test statistics and p-values from the Tail-Based Max test, Hosmer-Lemeshow test, and Windmeijer test. When Hosmer-Lemeshow test is performed, the grouping method is based on 10 percent quantiles of the fitted probabilities.

Value

A list with objects `summary_of_fit` and `summary_of_test`, where `summary_of_fit` is a matrix with column names "Estimate", "Std. Error", "z value", and "Pr(>|z|)". The rownames of the returned matrix will be the names of parameters. The object `summary_of_test` is a matrix with column names "Test Statistic" and "P-value". The rownames of this matrix will be "Tail-Based Max Test", "Hosmer and Lemeshow Test", and "Windmeijer Test".

References

Song M, Kraft P, Joshi A D, Barrdahl M, and Chatterjee N (2014) Testing calibration of risk models at extremens of disease-risk. Biostatistics.

Hosmer D W and Lemeshow S (2000) Applied logistic regression. New York : John Wiley & Sons.

Windmeijer, F A G (1990) The asymptotic distribution of the sum of weighted squared residuals in binary choice models, Statistica Neerlandica 44(2), 69-78.

See Also

GOFR

GOFR *GOFR*

Description

Goodness-Of-Fit tests for binary Risk models (GOFR).

Usage

```
GOFR(formula, data, link="logit", grid=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9),
     n_simulations, tol=10^-5, covariate=NULL)
```

Arguments

formula	A symbolic description of the model to be fitted. When there are covariates, one needs to include the covariates using +. For example, formula should be in the following form response~risk_factor_1+...+risk_factor_p+covariate_1+...+covariate_q when the model is as follows: the risk is additive with respect to risk factors as many as p and one needs to adjust covariates as many as q.
data	A data frame containing the variables in the model.
link	"logit" or "identity" to describe the link function to be used in the model. The default is "logit".
grid	Vector of risk percentiles to be used as grid points for evaluation of the test-statistics. Default is the vector of deciles (see Details).
n_simulations	The number of simulations to compute p-value of the Tail-Based Max-test-statistic (see Details).
tol	The convergence tolerance for fitting the model in the identity scale. The default is 10^-5.
covariate	The names of (categorical) variables which affect disease risk but are not used for constructing a risk prediction model. When the variables are specified as covariate, the odds ratio of risk factors is assumed to be homogeneous across the configurations of covariates. The default is NULL.

Details

The procedure implements three alternative goodness-of-fit tests. The **Hosmer-Lemeshow (HL)** test-statistic, which is based on observed vs expected event counts at group level, is implemented by categorizing subjects based on deciles of risk-distribution. The **Windmeijer** test-statistic is defined by sum of individual level model residuals over all subjects. The third procedure implements a **Tail-Based Max**-test-statistic **(TBM)**, where different tail-based (TB) test-statistics are first formed by summing model residuals over only those subjects which reach certain upper or lower risk-thresholds and then maximizing those statistics over different risk-thresholds. For example, if grid=(0.20,0.40,0.60,0.80), the **TBM** procedure uses two regions at the upper tail of the risk distribution (ie > 80th percentile and >60th percentile) and two regions at the lower tail of the risk distribution (ie < 20th percentile and <40th percentile). A TB statistic is evaluated for each of the four one-sided tail regions and also for the four combinations of the upper- and lower-tail risk-regions. In addition, a TB statistic is always evaluated using all subjects as it corresponds to union of the upper and lower tail regions defined by the median value of the risk distribution. The **TBM** statistic is then obtained by maximizing over nine different TB-statistics. If grid=(0.25,0.50,0.75), the **TBM** procedure uses one region at the upper tail of the risk distribution (ie >75th percentile) and one region at the lower tail of the risk distribution (ie < 25th percentile). A TB statistics is evaluated for each of the two one-sided tail regions and also for the one combination of the upper- and lower-tail risk-regions. In addition, the TB statistic is also evaluated using all subjects. Thus, in this case, the **TBM** statistic is obtained by maximizing over four different TB-statistics. To compute the p-value, we generate a multivariate Gaussian random variable from the covariance of the normalized TB statistics and mean of zero and obtain the maximum of the absolute values of the multivariate realization generated from the normal distribution. We then repeat the process as many as `n_simulations` and the p-value would be the ratio of the cases where the maxima from the simulations are larger than or equal to the maximum of the normalized TB statistics. The fitting of model under the identity link is done under a logistic representation of the identity-link function that assumes the disease is rare in the underlying population (see the reference by Song et al. for details).

Value

A list with the following objects:

- `coefficients` A named vector of coefficients which are the maximum likelihood estimates of the regression parameters of the specified model
- `fitted.values` A vector of fitted probabilities for the subjects in the data
- `cov` Estimates of variance-covariance matrix of the coefficients
- `TBM_ts` The test-statistic for the TBM method
- `TBM_pv` P-value for the TMB method
- `Windmeijer_ts` The test statistic for the Windmeijer method
- `Windmeijer_pv` P-value for the Windmeijer method
- `HL_ts` The test statistic for Hosmer-Lemeshow method
- `HL_pv` P-value for Hosmer-Lemeshow method

References

Song M, Kraft P, Joshi A D, Barrdahl M, and Chatterjee N (2014) Testing calibration of risk models at extremens of disease-risk, Biostatistics.

Hosmer D W and Lemeshow S (2000) Applied logistic regression. New York : John Wiley &

Sons.

Windmeijer, F A G (1990) The asymptotic distribution of the sum of weighted squared residuals in binary choice models, Statistica Neerlandica 44(2), 69-78.

See Also

getSummary

Examples

```
# Example 1  : Use simulated dataset which was generated under an "alternative" model that is in
# between additive and multiplicative model (see Song et al., 2013)

data(Xdata1,package="CBRM")

## Fit under the multiplicative model
fit_multiplicative<-GOFR(case.control~SNP1+SNP2+SNP3+SNP4+SNP5+SNP6+SNP7+SNP8+SNP9+SNP10,
      data=Xdata1, link="logit", grid=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9), n_simulations=10000)

## Compute a summary table for the analysis under the multiplicative model
getSummary(fit_multiplicative)

## Fit under the additive model
fit_additive<-GOFR(case.control~SNP1+SNP2+SNP3+SNP4+SNP5+SNP6+SNP7+SNP8+SNP9+SNP10,data=Xdata1,
    link="identity", grid=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9),n_simulations=10000, tol=10^-5)

## Compute a summary table for the analysis under the additive model
getSummary(fit_additive)

## Example 2 : Data are generated assuming a logistic model for 10 SNPs where  4 of the SNPs have
##  pair-wise interactions. The main effect is log(1.05) and the interaction effect is log(1.1).
data(Xdata2,package="CBRM")

## Fit under the model without interaction
fit_without_interactions<-GOFR(case.control~SNP1+SNP2+SNP3+SNP4+SNP5+SNP6+SNP7+SNP8+SNP9+SNP10,
   data=Xdata2, link="logit", grid=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9), n_simulations=10000)

## Compute a summary table for the analysis under the model without  interaction
getSummary(fit_without_interactions)


## Now fit  the model with correct interaction terms included ##
fit_with_interactions<-GOFR(case.control~SNP1+SNP2+SNP3+SNP4+SNP5+SNP6+SNP7+SNP8+SNP9+SNP10+
                            SNP1:SNP2+SNP1:SNP3+SNP1:SNP4+SNP2:SNP3+SNP2:SNP4+SNP3:SNP4,
                data=Xdata2, link="logit", grid=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9),
                n_simulations=10000)

## Compute a summary table for the analysis under the model with interaction among 4 SNPs
getSummary(fit_with_interactions)
```

Index