Package: SPPcomb (via r-universe)

October 11, 2024

Type Package

Title Combining Different Spatial Datasets in Cancer Risk Estimation

Version 0.1

Maintainer Ming Wang <willnju@gmail.com>

Author Ming Wang, Yongtao Guan, Kun Xu

Description We propose a novel two-step procedure to combine epidemiological data obtained from diverse sources with the aim to quantify risk factors affecting the probability that an individual develops certain disease such as cancer. See Hui Huang, Xiaomei Ma, Rasmus Waagepetersen, Theodore R. Holford, Rong Wang, Harvey Risch, Lloyd Mueller & Yongtao Guan (2014) A New Estimation Approach for Combining Epidemiological Data From Multiple Sources, Journal of the American Statistical Association, 109:505, 11-23, <doi:10.1080/01621459.2013.870904>.

License GPL

LazyData TRUE

RoxygenNote 5.0.1

Depends R (>= 2.10)

Imports nleqslv, stats

NeedsCompilation no

Repository CRAN

Date/Publication 2016-12-20 22:32:35

Contents

DA_AllEE5 *Data Analysis for Combining (N1,M1) + (N1,M2) + (N2,M1) + (N2,M2)*

Description

The main function is to solve the estimating equations constructed by combining all pairs (N1,M1), (N1,M2), (N2,M1) and (N2,M2) with selection bias probility $\pi(s, \eta)$ included.

Usage

```
DA_AllEE5(realdata_covariates, realdata_alpha, beta0)
```

Arguments

realdata_covariates	a list contains the following data matrics: CASEZ_1, CASEZ_2, CASEZhat_1, CASEZhat_2, CASEZhat_22, CONTZ_1, CONTZ_2, CONTZhat_1, CONTZhat_2, CONTZhat_22. For details please see definition in the help of realdata_covariates. Please be noted that all the variables have to use the same name as listed above.
realdata_alpha	a list contains the following data matrics: prob_case_1, prob_case_11, prob_case_2, prob_case_22, prob_cont_1, prob_cont_2, pwt_cont_2. details please see definition in the help of realdata_alpha. Please be noted that all the variables have to use the same name as listed above.
beta0	We need an initial parameter for solver "nleqslv". Default value is beta0=c(-5.4163,0.7790,-0.1289,0.2773,-0.5510,0.1568,0.4353,-0.6895)

Details

The function solves GMM combined estimating equation with handling selection bias, see Huang(2014).

Value

A list of estimator and its standard deviation.

References

Huang, H., Ma, X., Waagepetersen, R., Holford, T.R. , Wang, R., Risch, H., Mueller, L. & Guan, Y. (2014). A New Estimation Approach for Combining Epidemiological Data From Multiple Sources, Journal of the American Statistical Association, 109:505, 11-23.

Examples

```
#you can use glm to get the estimate as the initial value of beta0
#beta0=c(-5.4163,0.7790,-0.1289,0.2773,-0.5510,0.1568,0.4353,-0.6895)
#DA_AllEE5(realdata_covariates,realdata_alpha,beta0=beta0)
```

DA_AllEE5_inside *Internal calculation of estimating equation for DA_AllEE5*

Description

This is the internal function to solve the estimating equation constructed by pair (N1,M1), (N1,M2), (N2,M1) and (N2,M2) with selection bias probility $\pi(s, \eta)$ included. Since it's a internal function for function DA_AllEE5, thus it's not a necessary or important function.

Usage

```
DA_AllEE5_inside(beta, CASEZ_1, CASEZ_2, CASEZhat_1, CASEZhat_2, CASEZhat_22,
  CONTZ_1, CONTZ_2, CONTZhat_1, CONTZhat_2, CONTZhat_22, prob_case_1,
  prob_case_11, prob_case_2, prob_case_22, prob_cont_1, prob_cont_2, p,
  pi_case_1, pi_case_1_t, pi_case_2, pi_case_2_t, pi_cont_1, pi_cont_1_t,
  pi_cont_2, Z_case_pi_1, Z_case_pi_1_t, Z_case_pi_2, Z_case_pi_2_t,
  Z_cont_pi_1, Z_cont_pi_1_t, Z_cont_pi_2, J_step3, V_step3, pwt_cont_2,
  subset_2, subset_3, subset_4)
```

Arguments

beta	Parameter β.
CASEZ_1, CASEZhat_1	case data(N1) from case-control study, details please see definition in the help of realdata_covariates.
CASEZ_2, CASEZhat_2, CASEZhat_22	CTR data(N2), details please see definition in the help of realdata_covariates.
CONTZ_1, CONTZhat_1	control data(M1) from case-control study, details please see definition in the help of realdata_covariates.
CONTZ_2, CONTZhat_2, CONTZhat_22	BRFSS data(M2), details please see definition in the help of realdata_covariates.
prob_cont_1, prob_cont_2, prob_case_1, prob_case_11, prob_case_2, prob_case_22, pwt_cont_2	please see definition in the help of realdata_alpha.
p	Number of parameters, a constant value of 8.
pi_case_1, pi_case_1_t, pi_case_2, pi_case_2_t, pi_cont_1, pi_cont_1_t, pi_cont_2	selection bias
Z_case_pi_1, Z_case_pi_1_t, Z_case_pi_2, Z_case_pi_2_t, Z_cont_pi_1, Z_cont_pi_1_t, Z_cont_pi_2	part of variables from covariates, used for the estiamtion of variance.
J_step3	Derivative of the estimating equation.
V_step3	Variance of the estimating equation.
subset_2	A vector of 1:(p-2).

subset_3	A vector of 1:p.
subset_4	A vector of 1:(p-2).

Details

The function solves estimating equation based on GMM combined estimating equations with handling selection bias. It also accounts for the uncertainty due to the estimated value of eta. The function will output the estimating equation at current input value beta. Hence it can be used in "nleqslv" to solve for β. Because the function also outputs J and V, the asymptotic variance of β can be calculated in a straightforward way. $\hat{Z}_l$ may be highly correlated with Z_d, so it is removed in the estimation. And it has to be careful in constructing f, J and V.

Value

A list of (f,J,V)

1. f The final form of the estimating equation after adjusting eta.
2. J_step3 The derivative of the estimating equation.
3. V_step3 The variance of the estimating equation.

DA_FDN2M1M2 *Data Analysis for Combining (N2,M1) + (N2,M2)*

Description

The main function to solve the estimating equations constructed by combining pair (N2,M1) and (N2,M2). Since there is just one case data, no selection bias needed.

Usage

```
DA_FDN2M1M2(realdata_covariates, realdata_alpha, subset_2, subset_4, p, beta0)
```

Arguments

realdata_covariates	a list contains the following data matrics: CASEZ_2, CASEZhat_2, CASEZhat_22, CONTZ_1, CONTZhat_1, CONTZhat_2, CONTZhat_22
realdata_alpha	a list contains the following data matrics: prob_case_22, prob_cont_1, prob_cont_2, pwt_cont_2
subset_2	A vector of 1:(p-2), which is the subset of $\hat{Z}_{21}$, i.e. $\hat{Z}^{\star}_{21}$ in equation (10) of Huang(2014). $\hat{Z}_l$ may be highly correlated with Z_d, so it is removed in the estimation. For the view of including more information, you can use the whole dataset.
subset_4	A vector of 1:(p-2), which is the subset of $\hat{Z}_{22}$.
p	number of parameters.
beta0	an initial parameter for solver "nleqslv".

Details

The function solves estimating equation based on (N2,M1) and (N2,M2), see Huang(2014).

Value

A list of estimator and its standard deviation.

References

Huang, H., Ma, X., Waagepetersen, R., Holford, T.R. , Wang, R., Risch, H., Mueller, L. & Guan, Y. (2014). A New Estimation Approach for Combining Epidemiological Data From Multiple Sources, Journal of the American Statistical Association, 109:505, 11-23.

Examples

```
#p <- 8
#subset_2 <- 1:p
#subset_4 <- 1:p
#beta0=c(-5.4163,0.7790,-0.1289,0.2773,-0.5510,0.1568,0.4353,-0.6895)
#DA_FDN2M1M2(realdata_covariates,realdata_alpha,subset_2,subset_4,p=p,beta0=beta0)
```

DA_FDN2M1M2_inside *Internal calculation of estimating equation for DA_FDN2M1M2*

Description

The internal function to solve the estimating equations constructed by combining pair (N2,M1) and (N2,M2). Since there is just one case data, no selection bias needed. Since it's a internal function for function DA_FDN2M1M2, thus it's not a necessary or important function.

Usage

```
DA_FDN2M1M2_inside(beta, CASEZ_2, CASEZhat_2, CASEZhat_22, CONTZ_1, CONTZhat_1,
  CONTZhat_2, CONTZhat_22, prob_case_2, prob_case_22, prob_cont_1, prob_cont_2,
  p, J, V, subset_2, subset_4, pwt_cont_2)
```

Arguments

beta	Parameter β.
CASEZ_2, CASEZhat_2, CASEZhat_22	CTR data(N2), details please see definition in the help of realdata_covariates.
CONTZ_1, CONTZhat_1	control data(M1) from case-control study, details please see definition in the help of realdata_covariates.
CONTZhat_2, CONTZhat_22	BRFSS data(M2), details please see definition in the help of realdata_covariates.

`prob_cont_1`, `prob_cont_2`, `prob_case_2`, `prob_case_22`, `pwt_cont_2`
: please see definition in the help of realdata_alpha.

`p`
: Number of parameters, a constant value of 8.

`J`
: The derivative of the estimating equation.

`V`
: The variance of the estimating equation.

`subset_2`
: A vector of 1:(p-2).

`subset_4`
: A vector of 1:(p-2).

Details

The function solves estimating equation based on (N2,M1) and (N2, M2) with handling selection bias. It also accounts for the uncertainty due to the estimated value of eta. The function will output the estimating equation at current input value beta. Hence it can be used in "nleqslv" to solve for β. Because the function also outputs J and V, the asymptotic variance of β can be calculated in a straightforward way. $\hat{Z}_l$ may be highly correlated with Z_d, so it is removed in the estimation.

Value

A list of (f,J,V)

1. f The final form of the estimating equation after adjusting eta.
2. J The derivative of the estimating equation.
3. V The variance of the estimating equation.

`DA_FDN2M2` *Data Analysis of (N2,M2)*

Description

The main function to solve the estimating equations constructed by (N2,M2). Since there is just one case data, no selection bias needed.

Usage

```
DA_FDN2M2(realdata_covariates, realdata_alpha, p, beta0)
```

Arguments

`realdata_covariates`
: a list contains the following data matrics: CASEZhat_2, CASEZhat_22, CONTZhat_2, CONTZhat_22

`realdata_alpha`
: a list contains the following data matrics: prob_case_22,prob_cont_2, pwt_cont_2

`p`
: number of parameters.

`beta0`
: an initial parameter for solver "nleqslv".

Details

The function solves estimating equation based on (N2,M2), see Huang(2014).

Value

A list of estimator and its standard deviation.

References

Huang, H., Ma, X., Waagepetersen, R., Holford, T.R. , Wang, R., Risch, H., Mueller, L. & Guan, Y. (2014). A New Estimation Approach for Combining Epidemiological Data From Multiple Sources, Journal of the American Statistical Association, 109:505, 11-23.

Examples

```
#p <- 8
#beta0=c(-5.4163,0.7790,-0.1289,0.2773,-0.5510,0.1568,0.4353,-0.6895)
#DA_FDN2M2(realdata_covariates,realdata_alpha,p=p,beta0=beta0)
```

---

realdata_alpha *A list of matrices containing value of alpha at each location.*

Description

A list of matrices containing value of alpha at each location.

Usage

```
realdata_alpha
```

Format

An object of class `list` of length 8.

Value

A list of 8 matrices of calculated value of alpha for case and control points.

counts_agebysex_state Age-by-sex stratification in Connecticut, which is a matrix with 18 rows and 2 variables: Male, Female. In this dataset the age-by-sex distribution based on the Census for the following ten age groups: 35-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, 71-75, 76-80, and 81-83.

prob_case_1 Value of alpha for cases in case-control study, this $\alpha_1(s)$ is matched to controls' age-by-sex proportion

prob_case_11 Value of alpha for cases in case-control study, this $\alpha_2(s)$ is matched to BRFSS age-by-sex proportion

prob_case_2 Value of alpha for cases in CTR, which is a matrix with 1929 rows and 1 variables: $\alpha_1(s)$. This $\alpha_1(s)$ in CTR is matched to controls' age-by-sex proportion in case-control study

prob_case_22 Value of alpha for cases in CTR, which is a matrix with 1929 rows and 1 variables: $\alpha_2(s)$. This $\alpha_2(s)$ in CTR is matched to BRFSS age-by-sex proportion

prob_cont_1 Value of alpha for controls in case-control study, which is a matrix with 690 rows and 1 variables: $\alpha_1(s)$. A dataset of controls' $\alpha_1(s)$ of its own

prob_cont_2 Another Value of alpha for controls in BRFSS data, which is a matrix with 4459 rows and 1 variables: $\alpha_2(s)$. A dataset of controls' $\alpha_2(s)$ in BRFSS data of its own

pwt_cont_2 Value of weights(sampling probability) for controls in BRFSS data, which is $\alpha_2^\star$ in equation(14) of Huang(2014), a matrix with 4459 rows and 1 variables

Examples

```
# For example of each matrix, type the command in R: attributes(realdata_alpha)
# to obtain names of 8 matrices in the list:
#"counts_agebysex_state", "prob_case_1", "prob_case_11", "prob_case_2",
#"prob_case_22", "prob_cont_1", "prob_cont_2", "pwt_cont_2".
```

`realdata_covariates` *A data list of matrices containing covariates of cases and controls.*

Description

The list includes 10 matrices of covariates of cases and controls from different sources. Some of them need to impute the missing data, some of them need to estimate the variables even not missing to make sure the consistent format of input.

Usage

```
realdata_covariates
```

Format

A list of 10 matrices

Value

in the first case data which has complete cases:

1. CASEZ_1=$[1, Z_d, Z_t, Z_l]$
2. CASEZhat_1=$[1, Z_d, \hat{Z}_t, Z_l]$

in the second case data(CTR) which has missing lifestyle covariates:

1. CASEZ_2=$[1, Z_d, Z_t]$
2. CASEZhat_2=$[1, Z_d, Z_t, \hat{Z}_l]$

3. CASEZhat_22=$[1, Z_d, \hat{Z}_t, \hat{Z}_l]$

in the 1st control data which has complete controls:

1. CONTZ_1=$[1, Z_d, Z_t, Z_l]$
2. CONTZhat_1=$[1, Z_d, Z_t, \hat{Z}_l]$

in the 2nd control data which has missing traffic covariates(BRFSS):

1. CONTZ_2=$[1, Z_d, Z_l]$
2. CONTZhat_2=$[1, Z_d, \hat{Z}_t, Z_l]$
3. CONTZhat_22=$[1, Z_d, \hat{Z}_t, \hat{Z}_l]$

Examples

```
# For example of each matrix, type the command in R: attributes(realdata_covariates)
# to obtain names of 10 bulit-in matrices in the list:
# "CASEZ_1", "CASEZhat_1", "CASEZ_2", "CASEZhat_2", "CASEZhat_22", "CONTZ_1",
# "CONTZhat_1", "CONTZ_2", "CONTZhat_2", "CONTZhat_22".
```

Index