Package 'SelectionBias'

July 21, 2025

Title Calculates Bounds for the Selection Bias for Binary Treatment and Outcome Variables

Version 2.0.0

Description Computes bounds and sensitivity parameters as part of sensitivity analysis for selection bias. Different bounds are provided: the SV (Smith and VanderWeele), AF (assumption-free) bound, GAF (generalized AF), and CAF (counterfactual AF) bounds. The calculation of the sensitivity parameters for the SV and GAF bounds assume an additional dependence structure in form of a generalized M-structure. The bounds can be calculated for any structure as long as the necessary assumptions hold. See Smith and VanderWeele (2019) <doi:10.1097/EDE.0000000000001032>, Zetterstrom and Waernbaum (2022) <doi:10.1515/em-2022-0108> and Zetterstrom (2024) <doi:10.1515/em-2023-0033>.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.1

Imports arm, lifecycle, stats

Suggests knitr, table1, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

Depends R (>= 3.5.0)

LazyData true

VignetteBuilder knitr

URL https://github.com/stizet/SelectionBias

BugReports https://github.com/stizet/SelectionBias/issues

NeedsCompilation no

Author Stina Zetterstrom [aut, cre] (ORCID: <https://orcid.org/0000-0003-0730-9425>), Ingeborg Waernbaum [aut] (ORCID: <https://orcid.org/0000-0002-4457-5311>)

Maintainer Stina Zetterstrom <stina.zetterstrom@statistics.uu.se>

Repository CRAN

Date/Publication 2024-03-27 13:00:05 UTC

Contents

`AFbound` *Assumption-free bound*

Description

`AFbound()` returns a list with the AF upper and lower bounds.

Usage

```
AFbound(whichEst, outcome, treatment, selection = NULL)
```

Arguments

`whichEst`	Input string. Defining the causal estimand of interest. Available options are as follows. (1) Relative risk in the total population: "RR_tot", (2) Risk difference in the total population: "RD_tot", (3) Relative risk in the subpopulation: "RR_sub", (4) Risk difference in the subpopulation: "RD_sub".
`outcome`	Input vector. A binary outcome variable. Either the data vector (length>=3) or two conditional outcome probabilities with P(Y=1|T=1,I_s=1) and P(Y=1|T=0,I_s=1) as first and second element.
`treatment`	Input vector. A binary treatment variable. Either the data vector (length>=3) or two conditional treatment probabilities with P(T=1|I_s=1) and P(T=0|I_s=1) as first and second element.
`selection`	Input vector or input scalar. A binary selection variable or a selection probability. Can be omitted for subpopulation estimands.

Value

A list containing the upper and lower AF bounds.

References

Zetterstrom, Stina and Waernbaum, Ingeborg. "Selection bias and multiple inclusion criteria in observational studies" Epidemiologic Methods 11, no. 1 (2022): 20220108.

Zetterstrom, Stina. "Bounds for selection bias using outcome probabilities" Epidemiologic Methods 13, no. 1 (2024): 20230033

Examples

```
# Example with selection indicator variable.
y = c(0, 0, 0, 0, 1, 1, 1, 1)
tr = c(0, 0, 1, 1, 0, 0, 1, 1)
sel = c(0, 1, 0, 1, 0, 1, 0, 1)
AFbound(whichEst = "RR_tot", outcome = y, treatment = tr, selection = sel)

# Example with selection probability.
selprob = mean(sel)
AFbound(whichEst = "RR_tot", outcome = y[sel==1], treatment = tr[sel==1],
 selection = selprob)

# Example with simulated data.
n = 1000
tr = rbinom(n, 1, 0.5)
y = rbinom(n, 1, 0.2 + 0.05 * tr)
sel = rbinom(n, 1, 0.4 + 0.1 * tr + 0.3 * y)
AFbound(whichEst = "RD_tot", outcome = y, treatment = tr, selection = sel)
```

CAFbound *Counterfactual assumption-free bound*

Description

`CAFbound()` returns a list with the CAF upper and lower bounds. The sensitivity parameters are inserted directly.

Usage

```
CAFbound(whichEst, M, m, outcome, treatment, selection = NULL)
```

Arguments

`whichEst`	Input string. Defining the causal estimand of interest. Available options are as follows. (1) Relative risk in the total population: "RR_tot", (2) Risk difference in the total population: "RD_tot", (3) Relative risk in the subpopulation: "RR_sub", (4) Risk difference in the subpopulation: "RD_sub".
`M`	Input value. Sensitivity parameter. Must be between 0 and 1 and larger than m.
`m`	Input value. Sensitivity parameter. Must be between 0 and 1 and smaller than M.
`outcome`	Input vector. A binary outcome variable. Either the data vector (length>=3) or two conditional outcome probabilities with P(Y=1|T=1,I_s=1) and P(Y=1|T=0,I_s=1) as first and second element.
`treatment`	Input vector. A binary treatment variable. Either the data vector (length>=3) or two conditional treatment probabilities with P(T=1|I_s=1) and P(T=0|I_s=1) as first and second element.
`selection`	Input vector or input scalar. A binary selection variable or a selection probability. Can be omitted for subpopulation estimands.

Value

A list containing the upper and lower CAF bounds.

References

Zetterstrom, Stina. "Bounds for selection bias using outcome probabilities" Epidemiologic Methods 13, no. 1 (2024): 20230033

Examples

```
# Example with selection indicator variable.
y = c(0, 0, 0, 0, 1, 1, 1, 1)
tr = c(0, 0, 1, 1, 0, 0, 1, 1)
sel = c(0, 1, 0, 1, 0, 1, 0, 1)
Mt = 0.8
mt = 0.2
CAFbound(whichEst = "RR_tot", M = Mt, m = mt, outcome = y, treatment = tr,
 selection = sel)

# Example with selection probability.
selprob = mean(sel)
CAFbound(whichEst = "RR_tot", M = Mt, m = mt, outcome = y[sel==1],
 treatment = tr[sel==1], selection = selprob)

# Example with subpopulation and no selection variable or probability.
Ms = 0.7
ms = 0.1
CAFbound(whichEst = "RR_sub", M = Ms, m = ms, outcome = y, treatment = tr)

# Example with simulated data.
n = 1000
tr = rbinom(n, 1, 0.5)
y = rbinom(n, 1, 0.2 + 0.05 * tr)
sel = rbinom(n, 1, 0.4 + 0.1 * tr + 0.3 * y)
Mt = 0.5
mt = 0.05
CAFbound(whichEst = "RD_tot", M = Mt, m = mt, outcome = y, treatment = tr,
 selection = sel)
```

GAFbound *Generalized assumption-free bound*

Description

`GAFbound()` returns a list with the GAF upper and lower bounds. The sensitivity parameters can be inserted directly or as output from `sensitivityparametersM()`.

Usage

```
GAFbound(whichEst, M, m, outcome, treatment, selection = NULL)
```

Arguments

whichEst	Input string. Defining the causal estimand of interest. Available options are as follows. (1) Relative risk in the total population: "RR_tot", (2) Risk difference in the total population: "RD_tot", (3) Relative risk in the subpopulation: "RR_sub", (4) Risk difference in the subpopulation: "RD_sub".
M	Input value. Sensitivity parameter. Must be between 0 and 1, larger than m and larger than max_t P(Y=1|T=t,I_s=1).
m	Input value. Sensitivity parameter. Must be between 0 and 1, smaller than M and smaller than min_t P(Y=1|T=t,I_s=1).
outcome	Input vector. A binary outcome variable. Either the data vector (length>=3) or two conditional outcome probabilities with P(Y=1|T=1,I_s=1) and P(Y=1|T=0,I_s=1) as first and second element.
treatment	Input vector. A binary treatment variable. Either the data vector (length>=3) or two conditional treatment probabilities with P(T=1|I_s=1) and P(T=0|I_s=1) as first and second element.
selection	Input vector or input scalar. A binary selection variable or a selection probability. Can be omitted for subpopulation estimands.

Value

A list containing the upper and lower GAF bounds.

References

Zetterstrom, Stina. "Bounds for selection bias using outcome probabilities" Epidemiologic Methods 13, no. 1 (2024): 20230033

Examples

```
# Example with selection indicator variable.
y = c(0, 0, 0, 0, 1, 1, 1, 1)
tr = c(0, 0, 1, 1, 0, 0, 1, 1)
sel = c(0, 1, 0, 1, 0, 1, 0, 1)
Mt = 0.8
mt = 0.2
GAFbound(whichEst = "RR_tot", M = Mt, m = mt, outcome = y, treatment = tr,
 selection = sel)

# Example with selection probability.
selprob = mean(sel)
GAFbound(whichEst = "RR_tot", M = Mt, m = mt, outcome = y[sel==1],
 treatment = tr[sel==1], selection = selprob)

# Example with subpopulation and no selection variable or probability.
Ms = 0.7
```

```
ms = 0.1
GAFbound(whichEst = "RR_sub", M = Ms, m = ms, outcome = y, treatment = tr)

# Example with simulated data.
n = 1000
tr = rbinom(n, 1, 0.5)
y = rbinom(n, 1, 0.2 + 0.05 * tr)
sel = rbinom(n, 1, 0.4 + 0.1 * tr + 0.3 * y)
Mt = 0.5
mt = 0.05
GAFbound(whichEst = "RD_tot", M = Mt, m = mt, outcome = y, treatment = tr,
 selection = sel)
```

`sensitivityparametersM` *Sensitivity parameters for the Smith and VanderWeele bound and the GAF bound*

Description

`sensitivityparametersM()` returns a list with the sensitivity parameters and an indicator if bias is negative and the treatment coding is reversed for an assumed model.

Usage

```
sensitivityparametersM(
  whichEst,
  whichBound,
  Vval,
  Uval,
  Tcoef,
  Ycoef,
  Scoef,
  Mmodel,
  pY1_T1_S1,
  pY1_T0_S1
)
```

Arguments

`whichEst`	Input string. Defining the causal estimand of interest. Available options are as follows. (1) Relative risk in the total population: "RR_tot", (2) Risk difference in the total population: "RD_tot", (3) Relative risk in the subpopulation: "RR_sub", (4) Risk difference in the subpopulation: "RD_sub".
`whichBound`	Input string. Defining the bound of interest. Available options are as follows. (1) SV bound: "SV", (2) GAF bound: "GAF".

Vval	Input matrix. The first column is the values of the categories of V. The second column is the probabilities of the categories of V. If V is continuous, use a fine grid of values and probabilities.
Uval	Input matrix. The first column is the values of the categories of U. The second column is the probabilities of the categories of U. If U is continuous, use a fine grid of values and probabilities.
Tcoef	Input vector. Two numerical elements. The first element is the intercept in the model for the treatment. The second element is the slope in the model for the treatment.
Ycoef	Input vector. Three numerical elements. The first element is the intercept in the model for the outcome. The second element is the slope for T in the model for the outcome. The third element is the slope for U in the model for the outcome.
Scoef	Input matrix. Numerical matrix of size K by 4, where K is the number of selection variables. Each row is the coefficients for one selection variable. The first column is the intercepts in the models for the selection variables. The second column is the slopes for V in the models for the selection variables. The third column is the slopes for U in the models for the selection variables. The fourth column is the slopes for T in the models for the selection variables.
Mmodel	Input string. Defining the models for the variables in the M structure. If "P", the probit model is used. If "L", the logit model is
pY1_T1_S1	Input scalar. The observed probability P(Y=1|T=1,I_S=1).
pY1_T0_S1	Input scalar. The observed probability P(Y=1|T=0,I_S=1). used.

Value

A list containing the sensitivity parameters and, for the SV bound, an indicator if the treatment has been reversed.

References

Smith, Louisa H., and Tyler J. VanderWeele. "Bounding bias due to selection." Epidemiology (Cambridge, Mass.) 30.4 (2019): 509.

Zetterstrom, Stina and Waernbaum, Ingeborg. "Selection bias and multiple inclusion criteria in observational studies" Epidemiologic Methods 11, no. 1 (2022): 20220108.

Zetterstrom, Stina. "Bounds for selection bias using outcome probabilities" Epidemiologic Methods 13, no. 1 (2024): 20230033

Examples

```
# Examples with no selection bias.
V = matrix(c(1, 0, 0.1, 0.9), ncol = 2)
U = matrix(c(1, 0, 0.1, 0.9), ncol = 2)
Tr = c(0, 1)
Y = c(0, 0, 1)
S = matrix(c(1, 0, 0, 0, 1, 0, 0, 0), nrow = 2, byrow = TRUE)
probT1 = 0.534
probT0 = 0.534
```

```
sensitivityparametersM(whichEst = "RR_tot", whichBound = "SV", Vval = V,
 Uval = U, Tcoef = Tr, Ycoef = Y, Scoef = S, Mmodel = "P",
 pY1_T1_S1 = probT1, pY1_T0_S1 = probT0)

sensitivityparametersM(whichEst = "RR_tot", whichBound = "GAF", Vval = V,
 Uval = U, Tcoef = Tr, Ycoef = Y, Scoef = S, Mmodel = "P",
 pY1_T1_S1 = probT1, pY1_T0_S1 = probT0)


# Examples with selection bias. DGP from the zika example.
V = matrix(c(1, 0, 0.85, 0.15), ncol = 2)
U = matrix(c(1, 0, 0.5, 0.5), ncol = 2)
Tr = c(-6.2, 1.75)
Y = c(-5.2, 5.0, -1.0)
S = matrix(c(1.2, 2.2, 0.0, 0.5, 2.0, -2.75, -4.0, 0.0), ncol = 4)
probT1 = 0.286
probT0 = 0.004
sensitivityparametersM(whichEst = "RR_sub", whichBound = "SV", Vval = V,
 Uval = U, Tcoef = Tr, Ycoef = Y, Scoef = S, Mmodel = "L",
 pY1_T1_S1 = probT1, pY1_T0_S1 = probT0)

sensitivityparametersM(whichEst = "RR_sub", whichBound = "GAF", Vval = V,
 Uval = U, Tcoef = Tr, Ycoef = Y, Scoef = S, Mmodel = "L",
 pY1_T1_S1 = probT1, pY1_T0_S1 = probT0)
```

SVbound *Smith and VanderWeele bound*

Description

`SVbound()` returns a list with the SV bound. All sensitivity parameters for the population of interest must be set to numbers, and the rest can be left as `NULL`. The sensitivity parameters can be inserted directly or as output from `sensitivityparametersM()`. If the causal estimand is expected to be larger than the observational estimand, the recoding of the treatment has to be done manually.

Usage

```
SVbound(
  whichEst,
  pY1_T1_S1,
  pY1_T0_S1,
  RR_UY_T1 = NULL,
  RR_UY_T0 = NULL,
  RR_SU_T1 = NULL,
  RR_SU_T0 = NULL,
  RR_UY_S1 = NULL,
  RR_TU_S1 = NULL
)
```

Arguments

whichEst	Input string. Defining the causal estimand of interest. Available options are as follows. (1) Relative risk in the total population: "RR_tot", (2) Risk difference in the total population: "RD_tot", (3) Relative risk in the subpopulation: "RR_sub", (4) Risk difference in the subpopulation: "RD_sub".
pY1_T1_S1	Input value. The probability P(Y=1|T=1,I_S=1). Must be between 0 and 1.
pY1_T0_S1	Input value. The probability P(Y=1|T=0,I_S=1). Must be between 0 and 1.
RR_UY_T1	Input value. The sensitivity parameter RR_UY|T=1. Must be greater than or equal to 1. Used in the bounds for the total population.
RR_UY_T0	Input value. The sensitivity parameter RR_UY|T=0. Must be greater than or equal to 1. Used in the bounds for the total population.
RR_SU_T1	Input value. The sensitivity parameter RR_SU|T=1. Must be greater than or equal to 1. Used in the bounds for the total population.
RR_SU_T0	Input value. The sensitivity parameter RR_SU|T=0. Must be greater than or equal to 1. Used in the bounds for the total population.
RR_UY_S1	Input value. The sensitivity parameter RR_UY|S=1. Must be greater than or equal to 1. Used in the bounds for the subpopulation.
RR_TU_S1	Input value. The sensitivity parameter RR_TU|S=1. Must be greater than or equal to 1. Used in the bounds for the subpopulation.

Value

A list containing the Smith and VanderWeele bound.

References

Smith, Louisa H., and Tyler J. VanderWeele. "Bounding bias due to selection." Epidemiology (Cambridge, Mass.) 30.4 (2019): 509.

Zetterstrom, Stina and Waernbaum, Ingeborg. "Selection bias and multiple inclusion criteria in observational studies" Epidemiologic Methods 11, no. 1 (2022): 20220108.

Examples

```
# Example for relative risk in the total population.
SVbound(whichEst = "RR_tot", pY1_T1_S1 = 0.05, pY1_T0_S1 = 0.01,
 RR_UY_T1 = 2, RR_UY_T0 = 2, RR_SU_T1 = 1.7, RR_SU_T0 = 1.5)

# Example for risk difference in the total population.
SVbound(whichEst = "RD_tot", pY1_T1_S1 = 0.05, pY1_T0_S1 = 0.01,
 RR_UY_T1 = 2, RR_UY_T0 = 2, RR_SU_T1 = 1.7, RR_SU_T0 = 1.5)

# Example for relative risk in the subpopulation.
SVbound(whichEst = "RR_sub", pY1_T1_S1 = 0.05, pY1_T0_S1 = 0.01,
 RR_UY_S1 = 2.71, RR_TU_S1 = 2.33)

# Example for risk difference in the subpopulation.
SVbound(whichEst = "RD_sub", pY1_T1_S1 = 0.05, pY1_T0_S1 = 0.01,
```

```
RR_UY_S1 = 2.71, RR_TU_S1 = 2.33)
```

SVboundparametersM *Sensitivity parameters for the Smith and VanderWeele bound*

Description

[Deprecated] `SVboundparametersM()` has been deprecated and is replaced by `sensitivityparametersM()`.

`SVboundparametersM()` returns a list with the sensitivity parameters and an indicator if the bias is negative and the treatment coding is reversed for an assumed model.

Usage

```
SVboundparametersM(
  whichEst,
  Vval,
  Uval,
  Tcoef,
  Ycoef,
  Scoef,
  Mmodel,
  pY1_T1_S1,
  pY1_T0_S1
)
```

Arguments

`whichEst`	Input string. Defining the causal estimand of interest. Available options are as follows. (1) Relative risk in the total population: "RR_tot", (2) Risk difference in the total population: "RD_tot", (3) Relative risk in the subpopulation: "RR_sub", (4) Risk difference in the subpopulation: "RD_sub".
`Vval`	Input matrix. The first column is the values of the categories of V. The second column is the probabilities of the categories of V. If V is continuous, use a fine grid of values and probabilities.
`Uval`	Input matrix. The first column is the values of the categories of U. The second column is the probabilities of the categories of U. If U is continuous, use a fine grid of values and probabilities.
`Tcoef`	Input vector. Two numerical elements. The first element is the intercept in the model for the treatment. The second element is the slope in the model for the treatment.
`Ycoef`	Input vector. Three numerical elements. The first element is the intercept in the model for the outcome. The second element is the slope for T in the model for the outcome. The third element is the slope for U in the model for the outcome.

Scoef	Input matrix. Numerical matrix of size K by 4, where K is the number of selection variables. Each row is the coefficients for one selection variable. The first column is the intercepts in the models for the selection variables. The second column is the slopes for V in the models for the selection variables. The third column is the slopes for U in the models for the selection variables. The fourth column is the slopes for T in the models for the selection variables.
Mmodel	Input string. Defining the models for the variables in the M structure. If "P", the probit model is used. If "L", the logit model is
pY1_T1_S1	Input scalar. The observed probability P(Y=1|T=1,I_S=1).
pY1_T0_S1	Input scalar. The observed probability P(Y=1|T=0,I_S=1). used.

Value

A list containing the sensitivity parameters and an indicator if the treatment has been reversed.

References

Smith, Louisa H., and Tyler J. VanderWeele. "Bounding bias due to selection." Epidemiology (Cambridge, Mass.) 30.4 (2019): 509.

Zetterstrom, Stina and Waernbaum, Ingeborg. "Selection bias and multiple inclusion criteria in observational studies" Epidemiologic Methods 11, no. 1 (2022): 20220108.

SVboundsharp *Check if the Smith and VanderWeele bound in the subpopulation is sharp*

Description

`SVboundsharp()` returns a string that indicates if the SV bound is sharp or if it's inconclusive. If the bias is negative, the recoding of the treatment has to be done manually.

Usage

```
SVboundsharp(BF_U, pY1_T0_S1)
```

Arguments

BF_U	Input scalar. The bounding factor for the SV bounds in the subpopulation. Must be equal to or above 1. Can be inserted directly or as output from `sensitivityparametersM()`.
pY1_T0_S1	Input scalar. The probability P(Y=1|T=0,I_S=1).

Value

A string stating if the SV bound is sharp or inconclusive.

References

Smith, Louisa H., and Tyler J. VanderWeele. "Bounding bias due to selection." Epidemiology (Cambridge, Mass.) 30.4 (2019): 509.

Zetterstrom, Stina, and Ingeborg Waernbaum. "SelectionBias: An R Package for Bounding Selection Bias." arXiv preprint arXiv:2302.06518 (2023).

Examples

```
# Example where the SV bound is sharp.
SVboundsharp(BF_U = 1.56, pY1_T0_S1 = 0.33)

# Example where the SV bound is inconclusive.
SVboundsharp(BF_U = 2, pY1_T0_S1 = 0.8)
```

---

zika_learner *Simulated data set emulating a zika outbreak in Brazil*

Description

The data set is simulated to mimic real data. For the data generating process, see the vignette.

Usage

```
data(zika_learner)
```

Format

A data frame with 5,000 observations on the following 7 binary variables:

mic_ceph Indication if the baby has microcephaly (1=microcephaly, 0=not microcephaly)

zika Indication if the mother is infected by zika (1=infected, 0=not infected)

urban Indication of the living area of the subject (1=urban, 0=rural)

SES Indication of the socioeconomic status of the subject (1=high, 0=low)

birth First selection variable. Indication if the baby is born (1=birth, 0=terminated birth)

hospital Second selection variable. Indication if the delivery is in a public hospital (1=public, 0=private)

sel_ind Selection indicator variable. Indication if the subject is included in the study (1=included, 0=not included)

Details

The data set is created to use in examples of selection bias. A similar example has previously been used in articles that construct bounds for selection bias (Smith and VanderWeele, 2019; Zetterstrom and Waernbaum, 2022).

References

de Araújo, Thalia Velho Barreto, et al. "Association between microcephaly, Zika virus infection, and other risk factors in Brazil: final report of a case-control study." The Lancet infectious diseases 18.3 (2018): 328-336.

de Oliveira, Wanderson Kleber, et al. "Infection-related microcephaly after the 2015 and 2016 Zika virus outbreaks in Brazil: a surveillance-based analysis." The Lancet 390.10097 (2017): 861-870.

Ali, Sofia, et al. "Environmental and social change drive the explosive emergence of Zika virus in the Americas." PLoS neglected tropical diseases 11.2 (2017): e0005135.

Lebov, Jill F., et al. "International prospective observational cohort study of Zika in infants and pregnancy (ZIP study): study protocol." BMC Pregnancy and Childbirth 19.1 (2019): 1-10.

Malta, Monica, et al. "Abortion in Brazil: the case for women's rights, lives, and choices." The Lancet Public Health 4.11 (2019): e552.

Smith, Louisa H., and Tyler J. VanderWeele. "Bounding bias due to selection." Epidemiology (Cambridge, Mass.) 30.4 (2019): 509.

Zetterstrom, Stina and Waernbaum, Ingeborg. "Selection bias and multiple inclusion criteria in observational studies" Epidemiologic Methods 11, no. 1 (2022): 20220108.

https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?locations=BR

https://agenciabrasil.ebc.com.br/en/geral/noticia/2020-12/number-births-registered-brazil-down-2019

https://www.angloinfo.com/how-to/brazil/healthcare/health-system

Index