SelectionBias

```
library(SelectionBias)
```

Introduction

Selecting a study population from a larger source population, based on the research question, is a common procedure, for example in an observational study with data from a population register. Subjects who fulfill all the selection criteria are included in the study population, and subjects who do not fulfill at least one selection criterion are excluded from the study population. These selections might introduce a systematic error when estimating a causal effect, commonly referred to as selection bias. Selection bias can also arise if the selections are involuntary, for example, if there are dropouts or other missing values for some individuals in the study. In an applied study, it is often of interest to assess the magnitude of potential biases using a sensitivity analysis, such as bounding the bias. Two bounds, the SV and AF, for selection bias can be calculated in the R package `SelectionBias`. The content in `SelectionBias` is:

- `zika_learner`: a simulated dataset of zika virus and microcephaly inspired both by data and a previous example (Araújo et al. 2018; Smith and VanderWeele 2019).
- `SVboundparametersM()`: a function that calculates the sensitivity parameters for the SV bound for the an assumed model following the M-structure in Figure 1.
- `SVbound()`: a function that calculates the SV bound for the relative risk or risk difference in either the total or subpopulation for sensitivity parameters given by the user, or calculated from `SVboundparametersM()`.
- `AFbound()`: a function that calculates the AF bound, for the relative risk or risk difference in either the total or subpopulation, for a dataset that includes observations on an outcome and treatment variable and either a selection variable or a selection probability.
- `SVboundsharp()`: a function that evaluates if the SV bound for the subpopulation is sharp, inconclusive or not sharp.

For the formulas of the bounds as well as the theory behind them, we refer to the original papers (Smith and VanderWeele 2019; Zetterstrom and Waernbaum 2022) **ARXIV också**.

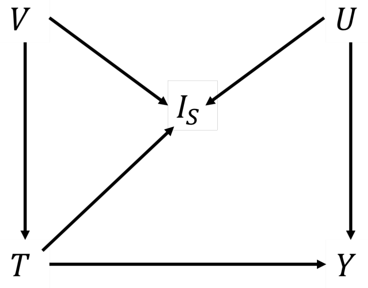

Figure 1: Figure 1. The generalized M-structure.

R package

Simulated zika dataset

T illustrate the bounds, a simulated, `zika_learner`, is constructed. It is inspired both by a numerical zika example used in Smith and VanderWeele (2019) together with a case-control study that investigates the effect of zika virus on microcephaly (Araújo et al. 2018). The variables included are:

- *Living area* (V).
- *Socioeconomic status, SES* (U).
- *Zika* (T).
- *Microcephaly* (Y).
- *Birth* (S_1).
- *Public hospital* (S_2).

The variables are related as seen in Figure 2 and Table 1. The prevalences of the variables, and strengths of dependencies between them, are chosen to mimic real data and the assumed values for the sensitivity parameters in Smith and VanderWeele (2019). The simulated data mimics a cohort with 5000 observations, even though the original study is a case-control study. For more details of the variables and the models, see **REF TILL OSS PÅ ARXIV**.

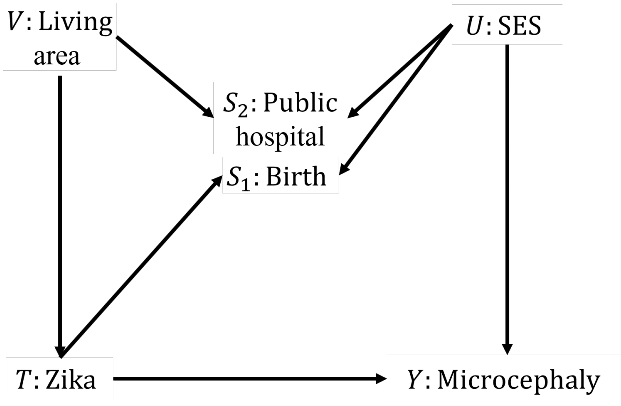

Figure 2: Figure 2. Causal model for the `zika_learner` dataset.

The causal dependencies are generated by the logistic models described in Table 3.

Table 1: Table 1. Data generating process for the dataset `zika_learner`. Models generating causal dependencies are logistic, $g(X'\theta)$, for predictor variable X and model parameter θ.

Model	Coefficients (θ)/Proportions	Function argument
$P(V = 1)$	0.85	`Vval`
$P(U = 1)$	0.50	`Uval`
$P(T = 1\|V) = g(V'\theta_T)$	$(-6.20, 1.75)$	`Tcoef`
$P(Y = 1\|T, U) = g[(T, U)'\theta_Y]$	$(-5.20, 5.00, -1.00)$	`Ycoef`
$P(S_1 = 1\|T, U) = g[(V, U, T)'\theta_{S1}]$	$(1.20, 0.00, 2.00, -4.00)$	`Scoef`
$P(S_2 = 1\|T, U) = g[(V, U, T)'\theta_{S1}]$	$(2.20, 0.50, -2.75, 0.00)$	`Scoef`

The data was generated on R, version 4.2.0, using the package arm, version 1.13-1, with the following code:

```r
# Seed.
set.seed(158118)
```

Table 2: Table 2. Proportions for the simulated dataset, by treatment status and overall.

	Not zika infected	Zika infected	Overall
	(N=4939)	(N=61)	(N=5000)
Microcephaly			
Mean	0.003	0.361	0.008
Living area			
Mean	0.849	0.951	0.850
SES			
Mean	0.499	0.426	0.498

```r
# Number of observations.
nObs = 5000

# The unmeasured variable, living area (V).
urban = rbinom(nObs, 1, 0.85)

# The treatment variable, zika.
zika_prob = arm::invlogit(-6.2 + 1.75 * urban)
zika = rbinom(nObs, 1, zika_prob)

# The unmeasured variable, SES (U).
SES = rbinom(nObs, 1, 0.5)

# The outcome variable, microcephaly.
mic_ceph_prob = arm::invlogit(-5.2 + 5 * zika - 1 * SES)
mic_ceph = rbinom(nObs, 1, mic_ceph_prob)

# The first selection variable, birth.
birth_prob = arm::invlogit(1.2 - 4 * zika + 2 * SES)
birth = rbinom(nObs, 1, birth_prob)

# The second selection variable, hospital.
hospital_prob = arm::invlogit(2.2 + 0.5 * urban - 2.75 * SES)
hospital = rbinom(nObs, 1, hospital_prob)

# The selection indicator.
sel_ind = birth * hospital
```

The resulting proportions of the `zika_learner` data, for the total dataset, the subset with $S_1 = 1$ and the subset with $S_1 = S_2 = 1$ are seen in Tables 2-4.

The dataset and data generating process (DGP) can be used to test the functions in `SelectionBias`.

`SVboundparametersM()`

The sensitivity parameters for the SV bound are calculated for the M-structure, illustrated in Figure 1. The sensitivity parameters are only calculated for an assumed model structure, since they depend on the unobserved variables, U. However, the observed probabilities of the outcome, $P(Y = 1|T = t, I_S = 1)$, $t = 0, 1$ are inputs in since they are used to check if the causal estimand for the assumed DGP is greater or smaller than the observed estimand. The code and the output are:

Table 3: Table 3. Proportions for the simulated dataset, by treatment status and overall, after the first selection.

	Not zika infected	Zika infected	Overall
	(N=4268)	(N=11)	(N=4279)
Microcephaly			
Mean	0.003	0.273	0.004
Living area			
Mean	0.845	1.000	0.846
SES			
Mean	0.556	0.818	0.557

Table 4: Table 4. Proportions for the simulated dataset, by treatment status and overall, after both selections.

	Not zika infected	Zika infected	Overall
	(N=2869)	(N=7)	(N=2876)
Microcephaly			
Mean	0.004	0.286	0.005
Living area			
Mean	0.858	1.000	0.858
SES			
Mean	0.382	0.714	0.382

```r
SVboundparametersM(whichEst = "RR_sub",
                   Vval = matrix(c(1, 0, 0.85, 0.15), ncol = 2),
                   Uval = matrix(c(1, 0, 0.5, 0.5), ncol = 2 ),
                   Tcoef = c(-6.2, 1.75),
                   Ycoef = c(-5.2, 5.0, -1.0),
                   Scoef = matrix(c(1.2, 2.2, 0.0, 0.5,
                                    2.0, -2.75, -4.0, 0.0),
                                  ncol = 4),
                   Mmodel = "L",
                   prob = c(0.286, 0.004))
#>      [,1]                [,2]
#> [1,] "BF_U"              1.5625
#> [2,] "RR_UY|S=1"         2.7089
#> [3,] "RR_TU|S=1"         2.3293
#> [4,] "Reverse treatment" TRUE
```

The first argument is `whichEst`, where the user inputs the causal estimand of interest. It must be one of the four `RR_tot`, `RD_tot`, `RR_sub` or `RD_sub`. Second, the argument `Vval` takes the matrix for V as input. The first column contains the values that V can take, and the second column contains the corresponding probabilities. In this example, V is binary, so the first two elements in the matrix are 1 and 0. However, any discrete V can be used. An approximation of a continuous V can be used, if it is discretized. The third argument is `Uval`, which takes the matrix for U as input. The matrix U has a similar structure as V. The fourth argument is `Tcoef`, containing the coefficients used in the model for T. The first entry in `Tcoef` is the intercept of the model, and the second the slope for V. The fifth argument is `Ycoef`, containing the coefficient vector for the outcome model, where the first entry is the intercept, the second the slope coefficient for T and third is the slope coefficient for U. The sixth argument is `Scoef`. `Scoef` is the coefficient matrix for the selection variables. The number of rows is equal to the number of selection variables, and the number of columns is equal to four. The columns represent the intercept, and slope coefficients for V, U and T, respectively. A summary of the code notation is seen in the last column of Table 3. The seventh argument is `Mmodel`, which indicates whether the models in the M-structure are probit (`Mmodel = "P"`) or logit (`Mmodel = "L"`). The eigth and last argument is `prob`. It is a vector of length two, where the first entry is $P(Y = 1|T = 1, I_S = 1)$ and the second entry is $P(Y = 1|T = 0, I_S = 1)$. The output is the sensitivity parameters for SV bound and an indicator stating if the bias is negative and the coding for the treatment has been reversed.

In the zika example, the estimand of interest is the relative risk in the subpopulation, `whichEst = RR_sub`, the DGP is found in Table 1, logistic models are used in the DGP and the probabilities are found in Table 2.The output is $RR_{TU|S=1} = 2.33$ and $RR_{UY|S=1} = 2.71$, which gives $BF_U = 1.56$, and the treatment coding is reversed.

`SVbound()`

The SV bound can be calculated using the function `SVbound()`. The first argument is `whichEst`, indicating the causal estimand of interest (`RR_tot`, `RD_tot`, `RR_sub` or `RD_sub`). The subsequent arguments are the sensitivity parameters provided by the user. The default value for all sensitivity parameters are `NULL`, and the user must then specify numeric values on the sensitivity parameters that are necessary for the bound for the chosen estimand. The sensitivity parameter can either be calculated using `SVboundparametersM()`, or found elsewhere. For sensitivity parameters found elsewhere, `SVbound()` is not restricted to the M-structure. However, the necessary assumptions for the SV bound must still be fulfilled (Smith and VanderWeele 2019). The output is the SV bound. The code and output are:

```
SVbound(whichEst = "RR_sub",
        RR_UY_S1 = 2.71,
        RR_TU_S1 = 2.33)
#>      [,1]       [,2]
#> [1,] "SV bound" 1.56
```

As before in the zika example, the causal estimand is the relative risk in the subpopulation, `whichEst = RR_sub`. The sensitivity parameters are $RR_{UY|S=1} = 2.71$ and $RR_{TU|S=1} = 2.33$, calculated above in `SVboundparametersM()`, which gives an SV bound equal to 1.56. If the causal estimand is underestimated, the recoding of the treatment must be done manually.

`AFbound()`

The AF bound is calcualted using the function `AFbound()`. The first argument is the causal estimand of interest (`RR_tot`, `RD_tot`, `RR_sub` or `RD_sub`). The second argument is `outcome`, where the user inputs the obesrved numeric vector with the outcome variable. The third argument is `treatment`, where the user inputs the observed treatment vector. The fourth argument is `selection`, where the user can either input the observed selection vector, or the selection probability. The output is the AF bound. The code and output are:

```
attach(zika_learner)

AFbound(whichEst = "RR_sub",
        outcome = mic_ceph,
        treatment = 1 - zika,
        selection = sel_ind)
#>      [,1]       [,2]
#> [1,] "AF bound" 3.5
```

Similar to before, `whichEst = "RR_sub"`. Furthermore, the outcome, treatment, and selection are the variables microcephaly, zika and the selection indicator. The coding of the treatment is manually reversed if needed. This has to be done manually. The output is the AF bound, which is 3.50 in this example.

In the above example, all observations were included in the vectors, even those with $I_S = 0$. However, if the data is not available for those subjects with $I_S = 0$, as could be the case with missing data, one can input the selection probability instead of the vector with the selection indicator variable. In this example, the selection probability is calculated as

```
mean(sel_ind)
#> [1] 0.5752
```

The code and output are:

```
AFbound(whichEst = "RR_sub",
        outcome = mic_ceph[sel_ind == 1],
        treatment = 1 - zika[sel_ind == 1],
        selection = mean(sel_ind))
#>      [,1]       [,2]
#> [1,] "AF bound" 3.5
```

When using the selection probability instead of the selection indicator variable, the other two vectors must be restricted to only include subjects with $I_S = 1$. The result is the same for both functions, since, in this example, the selection probability is calculated from the complete dataset.

`SVboundsharp()`

The sharpness of an SV bound can be evaluated using `SVboundsharp()` [**ARXIV**]. The first argument, `BF_U`, is the value of BF_U which can be calculated using `SVboundparametersM`. The second argument, `prob`, is the probability $P(Y = 1|T = 0, I_S = 1)$. Next, there are two optional arguments, `SVbound` and `AFbound`. These are not necessary to check if the SV bound is sharp, or if it is inconclusive, but they are necessary if the user wants to check if the bound is *not* sharp.The output is a string stating if the SV bound is sharp, inconclusive or not sharp. The code and output are:

```
SVboundsharp(BF_U = 1.56,
             prob = 0.27,
             SVbound = 1.56,
             AFbound = 3.5)
#> [1] "SV bound is sharp."
```

In the zika example, $BF_U = 1.56$, $P(Y = 1|T = 0, I_S = 1) = 0.27$ (calculated from the `zika_learner`), and the SV and AF bounds are 1.56 and 3.5. Note that if the causal estimand is underestimated, the recoding of the treatment has to be done manually. In this setting, the SV bound is sharp. As before, the bias is negative, and we have reversed the coding of the treatment.

References

Araújo, Thalia Velho Barreto de, Ricardo Arraes de Alencar Ximenes, Demócrito de Barros Miranda-Filho, Wayner Vieira Souza, Ulisses Ramos Montarroyos, Ana Paula Lopes de Melo, Sandra Valongueiro, et al. 2018. "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): 328–36.

Smith, Louisa H, and Tyler J VanderWeele. 2019. "Bounding Bias Due to Selection." *Epidemiology* 30 (4): 509–16.

Zetterstrom, Stina, and Ingeborg Waernbaum. 2022. "Selection Bias and Multiple Inclusion Criteria in Observational Studies." *Epidemiologic Methods* 11 (1): 1–21.