Analysis scripts for “Bias due to self selection and loss to follow up in prospective cohort studies”

Guido Biele

7 mai 2018

### Inverse probability weighted beta binomial regression in a 2 step approach

#### Collect MoBa and MBRN data

The script 010\_prep\_data\_v9.R opens MoBa and MBRN files and extracts variables, i.e. exposures, outcomes and additional variables deemed to be useful for multiple imputation.

#### Impute data

Imputation is performed with the package mi. The script 020\_mi.R prepares data for the imputation and starts multiple chained imputation. 020\_mi.R depends on 021\_mi\_utils.R and 022\_make\_ipdata\_utils.R.

After the multiple chained imputation process has converged (checked with visual diagnostics from the mi package and Rhat values) the multiply imputed data sets are generated with the script 030\_make\_imputed\_data\_v9.R, which depends on 031\_make\_imputed\_data\_utils.R.

#### Compare population and Moba sample

The script 040\_comp\_SSBMOBAR\_v9.R loads statistics we received from Statistics Norway (SSB) and the multiple imputed MoBa data set and compares population and study sample with respect to mothers age, education and parity.

Some imputation is also needed for the SSB statistic, because Statistics NNorway does not provide counts for subgroups with fewer than 3 members[[1]](#footnote-25). Therefore, the counts for the smallest groups were imputed by using a poisson regression where age and parity determine counts of educational levels .

The script also prepares tables and figures for the manuscript.

#### Calculate inverse probability of participation weights

The script 050\_calc\_IPW\_2SB\_v9.R prepares data for the selection model, which is then used to calculate inverse probability of participation weights. The selection model and calculation of smoothed weights is performed on a cluster with the script 051\_calc\_IPW\_2SB\_v9\_sl.R. Lastly, the script 052\_plot\_check\_IPWs.R compares observed and predicted participation rates and plots those and the distribution of weights.

#### Specify regression models for the three analyses

Next we set up the regression models, i.e. we define outcome, predictors and adjustment variables for the analysis of effects of birth variables, mental health, and legal drug use on ADHD symptoms. This is done with the script 060\_analyses\_2SB\_v9.R, which uses 061\_load\_data\_2SB\_v9.R.

#### Save data for analysis with stan model

The script 07\_rdump\_data\_2SB\_v9.R saves multiply imputed predictors, outcomes and weights for analysis in Stan (20 data sets per analysis). Two implement a 2 stage Bayesian appraoch to inverse probability weighting that propagates uncertainty over weights into the regression model, the data included multiple weights calculate from the posterior samples of the selection model.

#### The Stan model

The Stan model Stan/bebi\_logit\_deltaAR\_2SB.stan simultaneously estimates the IPW and AR models as well as the difference in regression weights. This model needs an auxiliary file Stan/get\_iter.hpp with code to identify in which iteration the model currently is, because uncertainty in weights is propagated by letting weights vary between iterations of the sampling algorithm.

The stan model also calculates average marginal effects for the AR and IPW models, and as well as the difference of these .

The stan model is compiled with Stan/compile\_stanmodel.R

For each of the 3 analyses we ran 20 imputations, each with 3 chains using the script 080\_run\_bebi.R. This analysis was performed on a cluster.

#### Checking regression model results

The script 090\_show\_rhats.R checks Rhats (potential scale reduction factor) to insure convergence and saves combined posterior samples per analysis.

#### Analysis results

The script 100\_results\_bebi\_logit\_deltaAR\_2SB.R extracts posterior samples of average marginal effects and calculates statistics from these. The script also prepares tables and figures for the manuscript.

### Calculating SNP based genetic correlations

#### Prepare downloaded GWAS summary data

The script LDSR/make\_LDSR\_data.R prepares downloaded GWAS summary data for use in LD score regressions.

#### Run LD score regressions

The scripts LDSR/munge.bash implements further pre-processing, and the scripts LDSR/h2.bash and LDSR/rg.bash calculate SNP heritability estimates and all pairwise genetic correlations, respectively. Information about sample size is saved in the file var\_N.txt.

#### Extract LD score regression results

The script LDSR/show\_rgh.R collects results of LD score regressions and makes plot and tables for LD score regression results.

1. For example, in 2001 122 women aged 25-29 who already had 4 or more children gave birth. Of these, 12 had either a bachelor or master’s degree. Because one of these two groups had fewer than 3 members, Statistic Norway only reported the total count for all mothers irrepsective of education, and the counts for the educational groups that were not master or bachelor. [↑](#footnote-ref-25)