ate_arf.R

Documentation for Running the R Program

October 2023

Peter Z. Schochet (pschochet@mathematica-mpr.com)

In clustered randomized controlled trials (RCTs), sample recruitment is often conducted after the randomization of clusters. This timing can lead to recruitment bias if access to the intervention affects the composition of cluster entrants in the study population and study consenters. The software, ***ate_arf.R***, is an R function that estimates average treatment effects (ATEs) in these settings using an inverse probability weighting (IPW) estimator developed in Schochet (Statistics in Medicine, 2024) for a causal estimand that pertains to the always-recruited in either research condition. The estimator uses data on *recruits* only and employs a generalized estimating equations approach to obtain clustered standard errors that adjusts for estimation error in the IPW weights from the propensity score logit models. The method allows for baseline covariates in the ATE regression models to obtain doubly robust ATE estimators.

INPUTS

Enter inputs using the c() function

Variable	Example input	Description
data_csv <-	c('serv_rec.csv')	Input **csv** data file. A complete case analysis is conducted where missing data are excluded for all input variables provided below.
y_var <-	c('y','y1')	Outcome variables
trt_var <-	c('t')	Treatment indicator (1 = treatment group, 0 = control group)
x_logit <-	c('x1','x2')	Covariates for the propensity score logit model modeling the probability of being recruited in a *control* cluster versus a treatment cluster (covariates are required)
x_wls <-	c('x1','x2','x3') or c(0)	Covariates for the weighted least squares models to obtain doubly, robust ATE estimators; c(0) = no covariates
clus_var <-	c('clus_id')	Name of clustering variable (cluster IDs)
wgt <-	c('ww1') or c(0)	Name of weight to adjust for the sample design or missing data c(0) = no weights
marg <-	c(1) or c(0)	c(1) = marginal logit model that conditions on the covariates only c(0) = random effects logit model that also conditions on random intercepts
se_ipw <-	c(1) or c(0)	c(1) = standard errors adjust for error in the IPW weights c(0) = no standard error adjustments (could run faster)
df_wls <-	c(1) or c(0)	c(1) = degrees of freedom adjust for regression model variables only c(0) = degrees of freedom also adjust for propensity model variables
out_regr <-	c('cace regr log.txt')	Name of output **txt** file summarizing data and regression results
out_est <-	c('cace est.txt')	Name of output **txt** file containing CACE estimation results

RUNNING THE PROGRAM

The program was developed using Version R 4.2.3 and tested using R Studio 2023.06.1.

Install the required libraries

Before running the program, you will need to install two R packages from the official R repository (CRAN): lme4 and estimatr. These packages can be installed, for example, using the install.packages("lme4") command and similarly for estimatr. If not installed, you may be asked if you want them installed the first time you run the program.

Steps for running the program

```
# Set the working directory
setwd("C:/MyDirectory")

# Enter the inputs from above

# Call the ate_arf.R script that was saved to the working directory
source("ate_arf.R")

# Call the function to conduct the analysis
ate_arf(data_csv,y_var,trt_var,x_logit,x_wls,clus_var,wgt,marg,se_ipw,df_wls,out_regr,out_est)

# You can view the output text files using Notepad

# You can call the function again with different inputs, where you only need to re-specify the inputs that
# you want to change. However, make sure to provide new inputs for the output txt files (out_regr and
# out_est) or the old output files will be overwritten.
```

TECHINCAL NOTES

1. The program assumes the data have been cleaned. The program does not check for invalid data values. It may not run properly with a csv data file with invalid data elements or missing data codes. However, the program removes observations with (i) missing values for any of the input variables, (ii) treatment value not equal to 0 or 1, and (iii) negative weights (if included).
2. If input weights are provided, the program normalizes them to sum to the treatment and control group sample sizes. They are then multiplied by the IPW weights to estimate treatment effects.
3. For the random effects logit model (marg <- c(0)), the program calculates propensity scores using the estimated parameters for the covariates only and not using the estimated random effects.
4. In calculating standard errors using the generalized estimating equation approach (which generates robust estimators), the program applies a small sample adjustment used by the HC1 option in the lm_robust function in R.
5. Adjusting standard errors for the estimation error in the IPW weights might increase program running time.