HulC on VD

Contents

This file to reproduce the data analysis on the First Steps program of V&D (2021) and to compare the confidence interval by their approach and HulC.

I could not find the data online, so I just generate a random data.

```
# create pseudo data
set.seed(1)
N <- 1000
data <- data.frame(maternal_age = sample(20:50, N, replace = T),
                   sex = sample(c("M", "F"), N, replace = T),
                   race = sample(c("a", "b", "h", "w"), N, replace = T),
                   smoking = rbinom(N, 1, 1/10),
                   martial = rbinom(N, 1, 7/10))


# first_step: logit (martial == 0)*5 + (race == "w")*1 + (race == "b")*2
tmp <- (data$martial==0)*5 + (data$race == "w")*1 + (data$race == "b")*2
prob_first_step <- exp(tmp)/(1+exp(tmp))
data$first_step <- rbinom(N, 1, prob_first_step)

# weight:
data$weight <- data$maternal_age^2*2 - data$smoking * 100
data$low_weight <- (data$weight < 2500)
```

1 EDA

Variable	Description
Infant birth weight (in grams)	Response variable
first_step	First steps program participation
maternal age	
child's sex	
race	asian, black, hispanic, white or other
smoking status	
marital status	

2 Linear model

The authors consider a linear model with and without interaction between the participation on the first step program and maternal age.

2.1 No interaction

```
# Y: infant birth weight (in grams)
# X: participation onthe First Steps program, maternal age, child's sex, mother's age, race (asian, bla
formula_lm <- weight ~ first_step + maternal_age + sex + race + smoking + martial
```

```r
# lm
fit_lm <- lm(formula_lm, data)
confint(fit_lm)

# HulC
HulC(data, FUN = lm, formula = formula_lm)
```

2.2 With interaction

```r
formula_int <- weight ~ first_step*maternal_age + sex + race + smoking + martial

# lm with interaction
fit_lm_int <- lm(formula_int, data)
confint(fit_lm)

# HulC
HulC(data, FUN = lm, formula = fit_lm_int)
```

3 GLM

Similarly, the author repeated the analysis after dichotomising the outcome (an infant was considered to have low birth weight if they weighed $< 2{,}500$g).

```r
formula_glm <- low_weight ~ first_step + maternal_age + sex + race + smoking + martial
# glm
fit_glm <- glm(formula_glm, data, family = "binomial")
summary(fit_glm)

# HulC
HulC(data, FUN = glm, formula = formula_glm, family = "binomial")
```

4 Nonparametric

Assume

$$g\{E(Y|A,L)\} = \beta A + \omega(L)$$

, the authors tried the "partialling out" estimator of Robinson (1988) and their proposal. All the nuisance are estimated using the `grf` package. They did not specify how they choose the tuning parameters so I assume they used the default.

```r
formula <- weight ~ maternal_age + sex + race + smoking + martial - 1
```

4.1 "partialling out" estimator of Robinson (1988)

$$\frac{\sum_{i=1}^{n}\{A_i - \hat{E}(A_i|L_i)\}\{Y_i - \hat{E}(Y_i|L_i)\}}{\sum_{i=1}^{n}\{A_i - \hat{E}(A_i|L_i)\}^2}$$

where the nuisances are estimated using `grf`.

```
partialling_out_estimator(data, formula = formula,
                          Y_var = "weight", A_var = "first_step")

HulC(data, FUN = partialling_out_estimator,
     formula = formula, Y_var = "weight", A_var = "first_step")
```

4.2 VD proposal

Copying from Page 15

1. Obtain the estimates $\hat{E}(A|L)$ and $\hat{E}(Y|A,L)$, e.g. using machine learning.
2. If A is binary, estimate $E[g\{E(Y|A,L)\}|L]$ as

$$\hat{E}[g\{\hat{E}(Y|A,L)\}|L] = g\{\hat{E}(Y|A=1,L)\}\hat{E}(A|L) + g\{\hat{E}(Y|A=0,L)\}\{1-\hat{E}(A|L)\}$$

   otherwise, use an additional machine learning fit (with $g\{\hat{E}(Y|A,L)\}$ as outcome).
3. Obtain an estimate of $\mu(Y,A,L)$:

$$\hat{\mu}(Y,A,L) = g^{-1}\{\hat{E}(Y|A,L)\}\{Y-\hat{E}(Y|A,L)\} + g\{\hat{E}(Y|A,L)\} - \hat{E}[g\{\hat{E}(Y|A,L)\}|L].$$

4. Fit a linear regression of $\mu(Y,A,L)$ on the sole predictor $A-\hat{E}(A|L)$ (without an intercept) using OLS in order to obtain an estimate $\hat{\beta}$ of β.

The variance is estimated through the sandwich estimator.

```
VD(data, formula = formula,
   Y_var = "weight", A_var = "first_step")

HulC(data, FUN = VD,
     formula = formula, Y_var = "weight", A_var = "first_step")
```