Untuned Random Forest Regression Model

Amos Okutse

25 December, 2022

Contents

0.0.1 LOAD DATA

```
rm(list = ls())
## load the saved single data files
load("C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\data\\df_one.RData")
load("C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\data\\df_two.RData")
load("C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\data\\df_three.RData")
load("C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\data\\df_four.RData")


## load the saved list data files
load("C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\data\\dsets1.RData")
load("C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\data\\dsets2.RData")
load("C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\data\\dsets3.RData")
load("C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\data\\dsets4.RData")
```

0.0.2 RANDOM FOREST MODELS

- The random forest models in this notebook are not tuned in any way and the results are based on

0.0.3 PART A: FULL DATA

```r
## create the function to return the desired estimates from the model
rf_one <- function(df = NULL){
  # fit random forest model for all individuals
  rf_all <- rand_forest(trees = 500) %>%
    set_mode("regression") %>%
    set_engine("ranger") %>%
    fit(formula = y ~ A + x1 + x2 + x3 + x4, data = df)
  ## set A = 0 and generate predictions for everyone
  df_A0 <- df
  df_A0$A <- 0
  pred_A0 <- predict(rf_all, df_A0)
  ## set A = 1 and generate predictions for everyone
  df_A1 <- df
  df_A1$A <- 1
  pred_A1 <- predict(rf_all, df_A1)
  ## compute the ATE
  ATE_adjusted = mean(pred_A1$.pred - pred_A0$.pred)
  ## compute the bias
  bias_adjusted = ATE_adjusted - 50
  ## return the results as a data frame
  rslt = data.frame("ATE_adjusted" = ATE_adjusted, "bias_adjusted" = bias_adjusted)
  return(rslt)
}
```

```r
# combine the results into a data frame
onea <- onea %>% map_dfr(data.frame)
oneb <- oneb %>% map_dfr(data.frame)
onec <- onec %>% map_dfr(data.frame)
oned <- oned %>% map_dfr(data.frame)
```

0.0.4 PART B: OBSERVED DATA ONLY

- Analysis restricted on the observed data alone, that is, where $R = 1$. Predictions are then made to only those individuals with observed outcomes.

```r
## create the function to return the desired estimates from the model
rf_two <- function(df = NULL){
  ## filter the data to have only individuals with R = 1
  df = dplyr::filter(df, R == 1)
  # fit random forest model for all individuals with R=1
  rf_two <- rand_forest(trees = 500) %>%
    set_mode("regression") %>%
    set_engine("ranger") %>%
    fit(formula = y ~ A + x1 + x2 + x3 + x4, data = df)
  ## set A=0 and generate predictions for those with R=1
  df_A0 <- df
  df_A0$A <- 0
  pred_A0 <- predict(rf_two, df_A0)
  ## set A=1 and generate predictions for those with R=1
  df_A1 <- df
  df_A1$A <- 1
```

```
  pred_A1 <- predict(rf_two, df_A1)
  ## compute the ATE
  ATE_adjusted = mean(pred_A1$.pred)-mean(pred_A0$.pred)
  ## compute the bias
  bias_adjusted = ATE_adjusted - 50
  ## return the results as a data frame
  rslt = data.frame("ATE_adjusted" = ATE_adjusted, "bias_adjusted" = bias_adjusted)
  return(rslt)
}
```

```
# combine the results into a data frame
twoa <- twoa %>% map_dfr(data.frame)
twob <- twob %>% map_dfr(data.frame)
twoc <- twoc %>% map_dfr(data.frame)
twod <- twod %>% map_dfr(data.frame)
```

0.0.5 PART C: MODIFIED AS IN PART B WITH PREDICTIONS FOR EVERYONE

```
## create the function to return the desired estimates from the model
rf_three <- function(df = NULL){
  # fit random forest model for all individuals with R=1
  rf_three <- rand_forest(trees = 500) %>%
    set_mode("regression") %>%
    set_engine("ranger") %>%
    fit(formula = y ~ A + x1 + x2 + x3 + x4, data = dplyr::filter(df, R == 1))
  ## set A = 0 and generate predictions for everyone
  df_A0 <- df
  df_A0$A <- 0
  pred_A0 <- predict(rf_three, df_A0)
  ## set A = 1 and generate predictions for everyone
  df_A1 <- df
  df_A1$A <- 1
  pred_A1 <- predict(rf_three, df_A1)
  ## compute the ATE
  ATE_adjusted = mean(pred_A1$.pred) - mean(pred_A0$.pred)
  ## compute the bias
  bias_adjusted = ATE_adjusted - 50
  ## return the results as a data frame
  rslt = data.frame("ATE_adjusted" = ATE_adjusted, "bias_adjusted" = bias_adjusted)
  return(rslt)
}
```

```
# combine the results into a data frame
threea <- threea %>% map_dfr(data.frame)
threeb <- threeb %>% map_dfr(data.frame)
threec <- threec %>% map_dfr(data.frame)
threed <- threed %>% map_dfr(data.frame)
```

Table 1: Untuned random forest regression model results averaged across n = 1000 datasets under full, observed, and observed modified analysis

Data generating values	n	ate	sd	bias
n = 500, SD = 1	500	47.17470	1.1979156	-2.8252963
n = 500, SD = 1	244	45.20914	1.9245677	-4.7908564
n = 500, SD = 1	244	44.89195	2.2606900	-5.1080539
n = 500, SD = 45	500	45.58716	4.5536492	-4.4128367
n = 500, SD = 45	258	41.95029	7.0187998	-8.0497138
n = 500, SD = 45	258	41.59113	7.4554034	-8.4088655
n = 2000, SD = 1	2000	49.04485	0.4016603	-0.9551542
n = 2000, SD = 1	997	48.42142	0.6261312	-1.5785787
n = 2000, SD = 1	997	48.24541	0.8353125	-1.7545887
n = 2000, SD = 45	2000	49.04864	0.3997893	-0.9513560
n = 2000, SD = 45	1003	47.72648	3.0813689	-2.2735209
n = 2000, SD = 45	1003	47.51083	3.4683437	-2.4891737

0.1 EXTRACT RESULTS

0.2 EXPORT RESULTS TO TABLE

```
untuned_rf = bind_rows(list("n = 500, SD = 1" = full, "n = 500, SD = 1" = obs, "n = 500,
↪  SD = 1"= obs_m, "n = 500, SD = 45" = full2, "n = 500, SD = 45" = obs2, "n = 500, SD =
↪  45" = obs_m2, "n = 2000, SD = 1" = full3, "n = 2000, SD = 1" = obs3, "n = 2000, SD =
↪  1" = obs_m3, "n = 2000, SD = 45" = full4, "n = 2000, SD = 45"=  obs4, "n = 2000, SD =
↪  45" = obs_m4), .id = "Data generating values")
kable(untuned_rf, format = "latex", caption = "Untuned random forest regression model
↪  results averaged across n = 1000 datasets under full, observed, and observed modified
↪  analysis")
```

```
## the order of the rows starts with n = 500
write.csv(untuned_rf, file = "C:\\Users\\aokutse\\OneDrive - Brown
↪  University\\ThesisResults\\[4]_random_forest\\untuned_rf\\untuned_rf_results.csv",
↪  row.names = FALSE)
```