Linear Regression with Interactions

Amos Okutse

21 December, 2022

Contents


```
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")
```

0.1 INTERACTED LINEAR REGRESSION MODEL

- The linear model fitted in this context has all two-way interactions between the treatment variable and the baseline covariates and is supposed to be misspecified.

0.1.1 PART A: FULL DATA

```
## create the function to return the desired estimates from the model
lm_interact_one <- function(df = NULL){
  # fit random forest model for all individuals
  lm_all <- linear_reg() %>%
  set_mode("regression") %>%
  set_engine("lm") %>%
  fit(formula = y ~ A + x1 + x2 + x3 + x4 + A*x1 + A*x2 + A*x3 + A*x4, data = df)
## set A = 0 and generate predictions for everyone
  df_A0 <- df
  df_A0$A <- 0
  pred_A0 <- predict(lm_all, df_A0)
## set A = 1 and generate predictions for everyone
  df_A1 <- df
  df_A1$A <- 1
  pred_A1 <- predict(lm_all, df_A1)
## compute the ATE
  ATE_adjusted = mean(pred_A1$.pred - pred_A0$.pred)
## compute the biases in absolute values
  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
  onea <- onea %>% map_dfr(data.frame) # n = 500, sd = 1
  oneb <- oneb %>% map_dfr(data.frame) # n = 500, sd = 45
  onec <- onec %>% map_dfr(data.frame) # n = 2000, sd = 1
  oned <- oned %>% map_dfr(data.frame) # n = 2000, sd = 45
```

0.1.2 PART B: OBSERVED DATA

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

```
## create the function to return the desired estimates from the linear model with
 ↪  analysis restricted to observed
lm_interact_two <- function(df = NULL){
## fit random forest model for all individuals with R=1
  df=dplyr::filter(df, R==1)
  lm_two <- linear_reg() %>%
  set_mode("regression") %>%
  set_engine("lm") %>%
  fit(formula = y ~ A + x1 + x2 + x3 + x4 + A*x1 + A*x2 + A*x3 + A*x4, data = df)
## set A=0 and generate predictions for those with R=1
  df_A0 <- df
  df_A0$A <- 0
  pred_A0 <- predict(lm_two, df_A0)
## set A=1 and generate predictions for those with R=1
```

```
  df_A1 <- df
  df_A1$A <- 1
  pred_A1 <- predict(lm_two, df_A1)
## compute the ATE
  ATE_adjusted = mean(pred_A1$.pred)-mean(pred_A0$.pred)
## compute the biases
  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) #n = 500, sd = 1
  twob <- twob %>% map_dfr(data.frame) # n = 500, sd = 45
  twoc <- twoc %>% map_dfr(data.frame) # n = 2000, sd = 1
  twod <- twod %>% map_dfr(data.frame) # n = 2000, sd = 45
```

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

```
## create the function to return the desired estimates from the linear model fitted on
 ↪  those with R==1 and predict for everyone
lm_interact_three <- function(df = NULL){

  # fit random forest model for all individuals with R=1
  lm_three <- linear_reg() %>%
  set_mode("regression") %>%
  set_engine("lm") %>%
  fit(formula = y ~ A + x1 + x2 + x3 + x4 + A*x1 + A*x2 + A*x3 + A*x4, data =
   ↪  dplyr::filter(df, R == 1))
## set A = 0 and generate predictions for everyone
  df_A0 <- df
  df_A0$A <- 0
  pred_A0 <- predict(lm_three, df_A0)
## set A = 1 and generate predictions for everyone
  df_A1 <- df
  df_A1$A <- 1
  pred_A1 <- predict(lm_three, df_A1)
## compute the ATE
  ATE_adjusted = mean(pred_A1$.pred) - mean(pred_A0$.pred)
## compute the biases
  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 data frames
  threea<- threea %>% map_dfr(data.frame)
  threeb<- threeb %>% map_dfr(data.frame)
```

```
  threec<- threec %>% map_dfr(data.frame)
  threed<- threed %>% map_dfr(data.frame)
```

0.2 Summary of the results from each case

```
## Extract analysis results from each data file
##------------------------------------------------------------------------------
## case 1 [n = 500, sd = 1]
##------------------------------------------------------------------------------
options(scipen = 999)
## full
full <- c(n = nrow(df_one), ate = mean(onea$ATE_adjusted), sd = sd(onea$ATE_adjusted),
 ↪  bias = mean(onea$bias_adjusted))
full
```

```
##           n         ate          sd        bias
## 500.00000000  50.09880381   1.50836408   0.09880381
```

```
## observed
obs <- c(n = nrow(subset(df_one, R == 1)), ate = mean(twoa$ATE_adjusted), sd =
 ↪  sd(twoa$ATE_adjusted), bias = mean(twoa$bias_adjusted))
obs
```

```
##            n          ate           sd         bias
## 244.000000000  49.996095782   1.888891798  -0.003904218
```

```
## observed modified
obs_m <- c(n = nrow(subset(df_one, R == 1)), ate = mean(threea$ATE_adjusted), sd =
 ↪  sd(threea$ATE_adjusted), bias = mean(threea$bias_adjusted))
obs_m
```

```
##           n          ate          sd         bias
## 244.00000000  49.94235813   3.10794308  -0.05764187
```

```
##------------------------------------------------------------------------------
## case 2 [n = 500, sd = 45]
##------------------------------------------------------------------------------
full2 <- c(n = nrow(df_two), ate = mean(oneb$ATE_adjusted), sd = sd(oneb$ATE_adjusted),
 ↪  bias = mean(oneb$bias_adjusted))
full2
```

```
##          n         ate          sd        bias
## 500.0000000  50.1447344   4.2006700   0.1447344
```

```
## observed
obs2 <- c(n = nrow(subset(df_two, R == 1)), ate = mean(twob$ATE_adjusted), sd =
 ↪  sd(twob$ATE_adjusted), bias = mean(twob$bias_adjusted))
obs2
```

```
##           n         ate          sd        bias
## 258.0000000  50.1352389   6.0406221   0.1352389
```

```
## observed modified
obs_m2 <- c(n = nrow(subset(df_two, R == 1)), ate = mean(threeb$ATE_adjusted), sd =
 ↪  sd(threeb$ATE_adjusted), bias = mean(threeb$bias_adjusted))
obs_m2
```

```
##            n          ate           sd         bias
## 258.00000000  50.06601146   7.26245700   0.06601146
```

```
##------------------------------------------------------------------------------
## case 3 [n = 2000, sd = 1]
##------------------------------------------------------------------------------
full3 <- c(n = nrow(df_three), ate = mean(onec$ATE_adjusted), sd = sd(onec$ATE_adjusted),
 ↪  bias = mean(onec$bias_adjusted))
full3
```

```
##             n           ate            sd          bias
## 2000.00000000   49.98054763    0.74097558   -0.01945237
```

```
## observed
obs3 <- c(n = nrow(subset(df_three, R == 1)), ate = mean(twoc$ATE_adjusted), sd =
 ↪  sd(twoc$ATE_adjusted), bias = mean(twoc$bias_adjusted))
obs3
```

```
##             n           ate            sd          bias
## 997.000000000  49.992417436   0.904396719  -0.007582564
```

```
## observed modified
obs_m3 <- c(n = nrow(subset(df_three, R == 1)), ate = mean(threec$ATE_adjusted), sd =
 ↪  sd(threec$ATE_adjusted), bias = mean(threec$bias_adjusted))
obs_m3
```

```
##            n          ate           sd         bias
## 997.00000000  49.97716299   1.52957594  -0.02283701
```

```
##------------------------------------------------------------------------------
## case 4 [n = 2000, sd = 45]
##------------------------------------------------------------------------------
full4 <- c(n = nrow(df_four), ate = mean(oned$ATE_adjusted), sd = sd(oned$ATE_adjusted),
 ↪  bias = mean(oned$bias_adjusted))
full4
```

```
##             n           ate            sd          bias
## 2000.00000000   50.04515567    2.14408135    0.04515567
```

Table 1: Interacted linear regression model results averaged across n = 1000 data sets under full, observed, and observed modified analyses

Data generating values	n	ate	sd	bias
n = 500, SD = 1	500	50.09880	1.5083641	0.0988038
n = 500, SD = 1	244	49.99610	1.8888918	-0.0039042
n = 500, SD = 1	244	49.94236	3.1079431	-0.0576419
n = 500, SD = 45	500	50.14473	4.2006700	0.1447344
n = 500, SD = 45	258	50.13524	6.0406221	0.1352389
n = 500, SD = 45	258	50.06601	7.2624570	0.0660115
n = 2000, SD = 1	2000	49.98055	0.7409756	-0.0194524
n = 2000, SD = 1	997	49.99242	0.9043967	-0.0075826
n = 2000, SD = 1	997	49.97716	1.5295759	-0.0228370
n = 2000, SD = 45	2000	50.04516	2.1440813	0.0451557
n = 2000, SD = 45	1003	50.13979	2.9709491	0.1397917
n = 2000, SD = 45	1003	50.13138	3.5657452	0.1313826

```
## observed
obs4 <- c(n = nrow(subset(df_four, R == 1)), ate = mean(twod$ATE_adjusted), sd =
 ↪  sd(twod$ATE_adjusted), bias = mean(twod$bias_adjusted))
obs4
```

```
##            n          ate           sd         bias
## 1003.0000000   50.1397917    2.9709491    0.1397917
```

```
## observed modified
obs_m4 <- c(n = nrow(subset(df_four, R == 1)), ate = mean(threed$ATE_adjusted), sd =
 ↪  sd(threed$ATE_adjusted), bias = mean(threed$bias_adjusted))
obs_m4
```

```
##            n          ate           sd         bias
## 1003.0000000   50.1313826    3.5657452    0.1313826
```

0.3 Table of Interacted Linear Regression Results

```
interacted_linear = 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(interacted_linear, format = "latex", caption = "Interacted linear regression model
 ↪  results averaged across n = 1000 data sets under full, observed, and observed
 ↪  modified analyses")
```

```
## save the results file as .csv
write.csv(interacted_linear, file = "C:\\Users\\aokutse\\OneDrive - Brown
 ↪  University\\ThesisResults\\[3]_interacted\\interacted_linear_results.csv", row.names
 ↪  = FALSE)
```