Sim_setting_binary

Hun

2/11/2022

For dichotomous outcomes, we use a logit link i.e., $log\frac{E[Y|X]}{1-E[Y|X]} = \sum_{i=0}^{p}\theta_i X_i$ and the true average treatment effect is $E[Y^1 = 1] - E[Y^0 = 1]$.

In this project, we want to simulate the binary outcome so that the treatment (i.e. smoking status) causes an increases in the probability of the occurrence of the outcome (i.e. lung cancer) by 15 percent. For our population without treatment (smoking) history, we assume that the probability of getting lung cancer is 0.2. Namely, E[Y^{0} = 1] = 0.2. From this fact, we know that our true log odds ratio (β_1) for our binary logistic regression is approximately 0.767.

We set our desired proportion of treatment is 0.14 given that 14% of U.S. adults were current smokers in 2019.

Based on this simulation design, we can obtain our α_0 coefficient for our propensity score model and β_0 and β_1 for our logistic model with some algebra computation.

With logit model and standardization method (creating pseudo population),

$E[Y^1 = 1] - E[Y^0 = 1]$

$\Leftrightarrow\ E[Y^1 = 1|A = 1] - E[Y^0 = 1|A = 0]$ (by exchangeabiltiy)

$\Leftrightarrow\ E[Y = 1|A = 1] - E[Y = 1|A = 0]$

$\Leftrightarrow\ \frac{e^{\beta_0+\beta_1 A+\beta_2 L_1+\beta_3 L_2}}{1+e^{\beta_0+\beta_1 A+\beta_2 L_1+\beta_3 L_2}} - 0.2 = 0.15$

$\Leftrightarrow\ e^{\beta_0+\beta_1 A+\beta_2 L_1+\beta_3 L_2} = 0.35 + 0.35e^{\beta_0+\beta_1 A+\beta_2 L_1+\beta_3 L_2}$

$\Leftrightarrow\ \beta_0 + \beta_1 A + \beta_2 L_1 + \beta_3 L_2 = log\frac{0.35}{0.65}$

Now we use,

$E[Y^0 = 1] = e^{\beta_0+\beta_2 L_1+\beta_3 L_2} = 0.2$

$\Leftrightarrow\ \beta_0 + \beta_2 L_1 + \beta_3 L_2 = log\frac{0.2}{0.8}$

$\Leftrightarrow\ \beta_2 L_1 + \beta_3 L_2 = log\frac{0.2}{0.8} + \beta_0$

Using this equation, we get the following:

$\Leftrightarrow\ \beta_0 + \beta_1 A + log\frac{0.2}{0.8} + \beta_0 = log\frac{0.35}{0.65}$

$\Leftrightarrow\ \beta_0 \approx (-0.619 + 1.386 + 0.767A)/2$

$\Rightarrow\ E[\hat{\beta_0}] \approx E[-0.619 + 1.386 + 0.767A]$

$\Rightarrow\ \hat{\beta_0} \approx (-0.619 + 1.386 + 0.767E[A])/2$

$\Rightarrow\ \hat{\beta_0} \approx (-0.619 + 1.386 + 0.767 * 0.14)/2$

$\Rightarrow\ \hat{\beta_0} \approx 0.437$

Finding α_0 and coefficients for L1, L2 ,L3 in propensity score model

This process allows us to find the values of coefficients for our propensity score model in order to achieve the expected proportion of the treated in our data simulation.

```
pre_data <- defData(varname = "L1", formula = "0", variance = 1,
                dist = "normal")
pre_data <- defData(pre_data, varname = "L2", formula = "0", variance = 1,
                dist = "normal")
pre_data <- defData(pre_data, varname = "L3", formula = "0", variance = 1,
                dist = "normal")

set.seed(777)
df <- genData(5000, pre_data)


data_gen <- function(desired_prop) {

   alpha_0 = log(desired_prop/(1 - desired_prop))

   coef_L1 <- sample(df$L1, 1000, replace = TRUE)

   coef_L2 <- sample(df$L2, 1000, replace = TRUE)

   coef_L3 <- sample(df$L3, 1000, replace = TRUE)

   A_logit <- vector(mode = "list",length = length(coef_L2))
   p_A <-  vector(mode = "numeric",length = length(coef_L2))

   p_A[1] <-  mean(alpha_0 + coef_L2[1]*df$L1 + coef_L2[1]*df$L2 - coef_L3[1]*df$L3)

   tol <- 1e-2

   i = 1

   while (abs(p_A[i] - 0.14) > tol) {

     i = i + 1

     A_logit[[i]] <-
       alpha_0 + coef_L2[i]*df$L1 + coef_L2[i]*df$L2 - coef_L3[i]*df$L3

     p_A[i] <- mean(exp(A_logit[[i]])/(1 + exp(A_logit[[i]])))

     if (abs(p_A[i] - 0.14) < tol) {
       mean_treated_proportion <- p_A[i]
       desired_coef_L1 <- coef_L1[i]
       desired_coef_L2 <- coef_L2[i]
       desired_coef_L3 <- coef_L3[i]
     }

     if (i > length(coef_L2)) {
       stop("You need better coverage, mate.")
    }
   }

   return(tibble(alpha_0,  mean_treated_proportion,
            desired_coef_L1, desired_coef_L2, desired_coef_L3))
}

set.seed(777)
data_gen(0.14)
```

```
## # A tibble: 1 x 5
##   alpha_0 mean_treated_proporti~ desired_coef_L1 desired_coef_L2 desired_coef_L3
##     <dbl>                  <dbl>           <dbl>           <dbl>           <dbl>
## 1   -1.82                  0.146          -0.877          -0.118           0.352
```

Generating the treatment and outcome data

Here, our intention is to introduce confouders in the data generation process in order to see how using propensity matching and standardization can tackle the problems caused by confounders in predicting the average treatment effect.

```
pre_data <- defData(pre_data, varname = "A",
                    formula = "-1.815 + -0.877*L1 - 0.118*L2 + 0.352*L3",
                dist = "binary", link = "logit")
pre_data <- defData(pre_data, varname = "Y",
                    formula = "0.437 + 0.767*A + 1.89*L2 -0.7*L3" ,
                dist = "binary", link = "logit")


df <- genData(5000, pre_data)

fit <- glm(Y~ A + L2 + L3, data = df, family = "binomial")

summary(fit)
```

```
##
## Call:
## glm(formula = Y ~ A + L2 + L3, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7349  -0.7168   0.2675   0.7127   2.9461
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.43993    0.04110  10.703   <2e-16 ***
## A            0.97355    0.10146   9.595   <2e-16 ***
## L2           1.85634    0.05590  33.208   <2e-16 ***
## L3          -0.75397    0.04083 -18.467   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6782.8  on 4999  degrees of freedom
## Residual deviance: 4502.7  on 4996  degrees of freedom
## AIC: 4510.7
##
## Number of Fisher Scoring iterations: 5
```

```
sampl.treated <- df %>%
  mutate(A = 1)

sampl.untreated <- df %>%
  mutate(A = 0)

mean(predict(fit, sampl.treated, type = "response") - predict(fit, sampl.untreated, type = "response"))
```

```
## [1] 0.1384354
```

Even though this is one simulation, we can check that the estimated ATE is somewhat reasonable in comparison to 0.15 (True ATE) considering we add two more covariates (L2 and L3) for the outcome data generation and haven't used propensity score matching to tackle confounders. We expect propensity score mathcing + standardization + bootstrap to result in better prediction.