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 approximatelyn0.767.

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 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:

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

$\Leftrightarrow\ \beta_0 \approx -0.619 + 1.386 + 0.767A$

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

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

$\Rightarrow\ \hat{\beta_0} \approx -0.619 + 1.386 + 0.767 * 0.14$

$\Rightarrow\ \hat{\beta_0} \approx 0.874$

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")
pre_data <- defData(pre_data, varname = "A",
                    formula = 0.14,
                dist = "binary")
pre_data <- defData(pre_data, varname = "Y",
                    formula = "0.874 + 0.767*A + 0.7*L2 + + 3*L3",
                dist = "binary", link = "logit")

set.seed(1)
df <- genData(5000, pre_data)

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

summary(fit)
```

```
##
## Call:
## glm(formula = Y ~ A, family = "binomial", data = df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.540  -1.333   1.029   1.029   1.029
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.35843    0.03104   11.55  < 2e-16 ***
## A            0.46365    0.08699    5.33 9.82e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6714.4  on 4999  degrees of freedom
## Residual deviance: 6684.8  on 4998  degrees of freedom
## AIC: 6688.8
##
## Number of Fisher Scoring iterations: 4
```

```
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] 0.1060171
```

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 outcome data generation and use a crappy model.