Homework 4

MA 590 Special Topics: Causal Inference

Aukkawut Ammartayakun

16 February, 2023

Problem 1

Estimate the ATE, with a 95% confidence interval, using Neyman's method, without any covariate adjustment.

```
#Neyman's method, coefficient of Z is the ATE
sls <- lm(Y~., data = data)
summary(sls)
```

```
##
## Call:
## lm(formula = Y ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7401 -0.5576  0.4140  0.4397  0.6037
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   6.661e-01  2.031e-01   3.279  0.00105 **
## Z                            -9.910e-03  1.417e-02  -0.699  0.48444
## sample_count                  1.533e-05  1.816e-05   0.844  0.39859
## avg_attempted                -2.889e-02  1.009e-01  -0.286  0.77470
## std_attempted                -5.015e-02  1.000e-01  -0.501  0.61612
## avg_attempted_before_support -3.076e-01  1.896e-01  -1.622  0.10484
## std_attempted_before_support -2.737e-02  1.263e-01  -0.217  0.82850
## avg_support_requested        -1.687e-01  1.699e-01  -0.993  0.32066
## std_support_requested         4.045e-02  1.258e-01   0.321  0.74786
## avg_completed                 1.570e-01  1.649e-01   0.952  0.34108
## std_completed                 1.531e-01  1.136e-01   1.348  0.17779
## avg_attempt_count             2.756e-02  3.088e-02   0.893  0.37217
## std_attempt_count            -5.533e-04  1.139e-02  -0.049  0.96126
## avg_correctness               9.604e-02  7.744e-02   1.240  0.21495
## std_correctness              -7.471e-02  1.038e-01  -0.720  0.47170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4964 on 4903 degrees of freedom
## Multiple R-squared:  0.00192,    Adjusted R-squared:  -0.00093
## F-statistic: 0.6737 on 14 and 4903 DF,  p-value: 0.8023
```

```
#find confidence interval
confint(sls)
```

```
##                                      2.5 %       97.5 %
## (Intercept)                   2.678701e-01 1.064326e+00
## Z                            -3.769566e-02 1.787524e-02
## sample_count                 -2.027383e-05 5.093985e-05
## avg_attempted                -2.267382e-01 1.689603e-01
## std_attempted                -2.462525e-01 1.459460e-01
## avg_attempted_before_support -6.794230e-01 6.416125e-02
## std_attempted_before_support -2.750157e-01 2.202824e-01
## avg_support_requested        -5.016975e-01 1.642936e-01
## std_support_requested        -2.062021e-01 2.870947e-01
## avg_completed                -1.663026e-01 4.803681e-01
## std_completed                -6.957458e-02 3.756929e-01
## avg_attempt_count            -3.297769e-02 8.809763e-02
## std_attempt_count            -2.288627e-02 2.177966e-02
## avg_correctness              -5.577584e-02 2.478641e-01
## std_correctness              -2.782077e-01 1.287837e-01
```

```
#find difference of confidence interval of ATE
confint(sls)[2,2] - confint(sls)[2,1]
```

```
## [1] 0.0555709
```

Problem 2

Estimate the ATE, with a 95% confidence interval, using OLS regression with "robust" standard errors (i.e. lm_robust). Include covariates in the regression–your choice which ones, or if you want to do anything fancy to them (e.g. include interactions between covariates, non-linear terms, etc.). Why is or isn't it OK to us OLS with a binary outcome?

```
#OLS with robust standard errors, with covariates of avg_attempted
sls <- lm_robust(Y~Z + avg_attempted, data = data)
summary(sls)
```

```
##
## Call:
## lm_robust(formula = Y ~ Z + avg_attempted, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)    0.57917    0.04735 12.2321 6.487e-34  0.48635  0.67200 4915
## Z             -0.01107    0.01415 -0.7819 4.343e-01 -0.03881  0.01668 4915
## avg_attempted -0.01447    0.05672 -0.2552 7.986e-01 -0.12567  0.09672 4915
##
## Multiple R-squared:  0.000138 ,  Adjusted R-squared:  -0.0002688
## F-statistic: 0.3392 on 2 and 4915 DF,  p-value: 0.7124
```

```
#find confidence interval
confint(sls)
```

```
##                     2.5 %     97.5 %
## (Intercept)    0.48634758 0.67199526
## Z             -0.03881371 0.01668118
## avg_attempted -0.12567459 0.09672472
```

```r
#find range of confidence interval of ATE
confint(sls)[2,2] - confint(sls)[2,1]
```

```
## [1] 0.0554949
```

Problem 3

Estimate the ATE, with a 95% confidence interval, using Lin (2012)'s method. Same deal with covariates as in part 2.

```r
#Lin's method
data_lin = data
#scale avg_attempted
data_lin$avg_attempted = scale(data_lin$avg_attempted)
#truncate the data to only include response, covariates, and treatment
data_lin = data_lin[,c(1,2,4)]
reg0 = lm(Y~avg_attempted, data = data_lin, subset = Z==0)
reg1 = lm(Y~avg_attempted, data = data_lin, subset = Z==1)
#regression only on covariates
y0 = predict(reg0, newdata=data_lin)
y1 = predict(reg1, newdata=data_lin)
#ATE
ATE = mean(y1-y0)
ATE
```

```
## [1] -0.01106611
```

```r
#CI
mean(y1-y0) + t.test(residuals(reg1), residuals(reg0))$conf.int
```

```
## [1] -0.03881032  0.01667810
## attr(,"conf.level")
## [1] 0.95
```

```r
#diff of CI
(mean(y1-y0) + t.test(residuals(reg1), residuals(reg0))$conf.int)[2] -
(mean(y1-y0) + t.test(residuals(reg1), residuals(reg0))$conf.int)[1]
```

```
## [1] 0.05548843
```

Problem 4

Choose a model other than OLS to model potential outcomes as a function of covariates, and use it to estimate the ATE with a 95% confidence interval, following Guo and Basse (2020)'s method.

```r
#Guo and Basse's method
dat=data[,c(1,2,4)]
reg1zero=glm(Y~avg_attempted,data=data,subset=Z==1,family=binomial)
reg1=lm(Y~avg_attempted,data=data,subset=Z==1&Y>0)
y1hat = predict(reg1zero,newdata=dat,type='response')*predict(reg1,newdata=dat)
y1hat=y1hat-mean(y1hat[dat$Z==1])+mean(dat$avg_attempted[dat$Z==1])
reg0zero=glm(Y~avg_attempted,data=data,subset=Z==0,family=binomial)
reg0=lm(Y~avg_attempted,data=data,subset=Z==0&Y>0)
y0hat=predict(reg0zero,newdata=dat,type='response')*predict(reg0,newdata=dat)
```

```
y0hat=y0hat-mean(y0hat[dat$Z==0])+mean(dat$avg_attempted[dat$Z==0])

resid1=dat$avg_attempted[dat$Z==1]-y1hat[dat$Z==1]
resid0=dat$avg_attempted[dat$Z==0]-y0hat[dat$Z==0]

(tau.hat=mean(y1hat-y0hat))
```

```
## [1] 0.001007091
```

```
(ci=tau.hat+t.test(resid1,resid0)$conf.int)
```

```
## [1] -0.006093175  0.008107357
## attr(,"conf.level")
## [1] 0.95
```

```
(ci[2]-ci[1])
```

```
## [1] 0.01420053
```

Problem 5

Estimate the ATE, with a 95% confidence interval using LOOP with the default "random forest" predictions. Use $p = Pr(Z = 1) = 0.5$.

```
#define design matrix with covariates
X = model.matrix(~avg_attempted, data = data)
#estimate the proportion of Z=1
p = mean(data$Z)
tau.loop = loop(data$Y, data$Z, X, p=p) #tau, var
```

```
## Warning in randomForest.default(Z[Tr == 1, , drop = FALSE], Y[Tr == 1, , : The response has five or 
## regression?

## Warning in rfout$mse/(var(y) * (n - 1)/n): Recycling array of length 1 in vector-array arithmetic is
##   Use c() or as.vector() instead.

## Warning in randomForest.default(Z[Tr == 0, , drop = FALSE], Y[Tr == 0, , : The response has five or 
## regression?

## Warning in rfout$mse/(var(y) * (n - 1)/n): Recycling array of length 1 in vector-array arithmetic is
##   Use c() or as.vector() instead.
```

```
tau.loop[1]
```

```
## [1] -0.01099899
```

```
#CI
CI = c(tau.loop[1] - 1.96*sqrt(tau.loop[2]),tau.loop[1] + 1.96*sqrt(tau.loop[2]))
CI
```

```
## [1] -0.03875704  0.01675906
```

```
#range of CI
CI[2] - CI[1]
```

```
## [1] 0.0555161
```

Problem 6

Estimate the the number of correct responses attributable to assignment to hints (vs explanations) using Hansen & Bowers (2008) method. We barely discussed this one in class, but check it out in the lecture notes and at the bottom of a newly-revised covariateAdjustment.r.

```
#Hansen and Bowers's method
#logistic regression to model Y as a function of covariates in Z = 0
reg0 = glm(Y~Z+avg_attempted, data = data, subset = Z==0, family = binomial)
#get predicted outcome for the whole sample
#truncate the data
data = data[,c(1,2,4)]
ypred = predict(reg0, newdata=data, type="response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == : prediction from a rank
```

```
#estimate attributable effect
(ae = sum(data$Y-ypred))
```

```
## [1] -27.20875
```

```
resids=ypred[data$Z==0]-data$Y[data$Z==0]
sum(resids)
```

```
## [1] -1.176836e-14
```

```
#confidence interval
ae + nrow(data)*t.test(resids)$conf.int
```

```
## [1] -123.6185   69.2010
## attr(,"conf.level")
## [1] 0.95
```

```
#range of CI
(ae + nrow(data)*t.test(resids)$conf.int)[2] -
(ae + nrow(data)*t.test(resids)$conf.int)[1]
```

```
## [1] 192.8195
```

Problem 7

Comment on what you found–did the estimates largely agree? Did covariate adjustment seem to help? Do you believe some answers more than others? If you had to choose one estimate of all six to include in a report, which would you choose, and why?

Solution

The result here are largely agree with each other that hint negatively impact to students. The covariate adjustment help tighten the confidence interval and since there are 3 majorly agree with each other, so three of those would be the reasonable one to trust. It is quite clear that the tighest confidence interval (that largely agree) would be the one to pick because it shows that the model is tightly fitted to the data. The confidence interval of the Lin's method is the tighest one, so I would choose the Lin's method.