Homework 4

MA 590 Special Topics: Causal Inference

Aukkawut Ammartayakun

15 February, 2023

Problem 1

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

```
#Neyman's method,vcoefficient 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
```

```r
#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
```

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?

```r
#OLS with robust standard errors
rbls <- lm_robust(Y~., data = data)
summary(rbls)
```

```
##
## Call:
## lm_robust(formula = Y ~ ., data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##                                 Estimate Std. Error  t value Pr(>|t|)   CI Lower  CI Upper   DF
## (Intercept)                    6.661e-01  0.2138304  3.11508  0.00185  2.469e-01 1.085e+00 4903
## Z                             -9.910e-03  0.0141736 -0.69920  0.48446 -3.770e-02 1.788e-02 4903
## sample_count                   1.533e-05  0.0000183  0.83767  0.40226 -2.055e-05 5.122e-05 4903
## avg_attempted                 -2.889e-02  0.1018843 -0.28355  0.77677 -2.286e-01 1.708e-01 4903
## std_attempted                 -5.015e-02  0.1008782 -0.49717  0.61909 -2.479e-01 1.476e-01 4903
## avg_attempted_before_support -3.076e-01  0.1930750 -1.59332  0.11115 -6.861e-01 7.088e-02 4903
## std_attempted_before_support -2.737e-02  0.1267918 -0.21584  0.82912 -2.759e-01 2.212e-01 4903
## avg_support_requested         -1.687e-01  0.1717785 -0.98209  0.32610 -5.055e-01 1.681e-01 4903
## std_support_requested          4.045e-02  0.1274453  0.31736  0.75098 -2.094e-01 2.903e-01 4903
## avg_completed                  1.570e-01  0.1748123  0.89829  0.36907 -1.857e-01 4.997e-01 4903
## std_completed                  1.531e-01  0.1162672  1.31644  0.18809 -7.488e-02 3.810e-01 4903
## avg_attempt_count              2.756e-02  0.0303972  0.90666  0.36463 -3.203e-02 8.715e-02 4903
## std_attempt_count             -5.533e-04  0.0112665 -0.04911  0.96083 -2.264e-02 2.153e-02 4903
## avg_correctness                9.604e-02  0.0780176  1.23106  0.21836 -5.691e-02 2.490e-01 4903
## std_correctness               -7.471e-02  0.1030611 -0.72493  0.46853 -2.768e-01 1.273e-01 4903
```

```
##
## Multiple R-squared:  0.00192 ,   Adjusted R-squared:  -0.00093
## F-statistic: 0.6637 on 14 and 4903 DF,  p-value: 0.8118
```

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

```
##                                    2.5 %       97.5 %
## (Intercept)                 2.468945e-01 1.085301e+00
## Z                          -3.769689e-02 1.787647e-02
## sample_count               -2.055181e-05 5.121784e-05
## avg_attempted              -2.286278e-01 1.708500e-01
## std_attempted              -2.479198e-01 1.476133e-01
## avg_attempted_before_support -6.861444e-01 7.088269e-02
## std_attempted_before_support -2.759354e-01 2.212021e-01
## avg_support_requested      -5.054646e-01 1.680608e-01
## std_support_requested      -2.094036e-01 2.902963e-01
## avg_completed              -1.856777e-01 4.997432e-01
## std_completed              -7.487656e-02 3.809949e-01
## avg_attempt_count          -3.203214e-02 8.715208e-02
## std_attempt_count          -2.264073e-02 2.153412e-02
## avg_correctness            -5.690539e-02 2.489936e-01
## std_correctness            -2.767579e-01 1.273339e-01
```

Problem 3

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

```
#Lin's method
data_lin = data
#scale everything except for Z
data_lin[,2:ncol(data_lin)] <- scale(data_lin[,2:ncol(data_lin)])
reg0 = lm(Y~. - Z, data = data_lin, subset = Z==0)
reg1 = lm(Y~. - Z, data = data_lin, subset = Z==1)
y0 = predict(reg0, newdata=data)
y1 = predict(reg1, newdata=data)
#ATE
ATE = mean(y1-y0)
ATE
```

```
## [1] 24.04541
```

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

```
## [1] 23.98964 24.10118
## attr(,"conf.level")
## [1] 0.95
```

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.

```
#Guo and Basse's method, using random forest
```

Problem 5

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

```
X = model.matrix(~.-Z, 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 f
```

```
## 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 f
```

```
## 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] -0.0003192842
```

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

```
## [1] -0.001216856  0.000578288
```

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~., data = data, subset = Z==0, family = binomial)
#get predicted outcome for the whole sample
ypred = predict(reg0, newdata=data, type="response")
```

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

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

```
## [1] -23.82167
```

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

```
## [1] 3.626544e-12
```

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

```
## [1] -120.01582   72.37248
```

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

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?