Stat 504 Midterm

Dongyang Wang

2/17/2022

Q1

1.1

Taking derivative of $argmin_\beta E((Y_i - X_i^T\beta)^2)$ we obtain $E(X_i(Y_i - X_i\beta)) = 0$. Then we have

$$\begin{aligned} E(X_i(Y_i - X_i\beta)) &= 0 \\ E(X_iY_i) - E(X_iX_i^T\beta) &= 0 \\ E(X_iY_i) &= E(X_iX_i^T)\beta \\ \beta_{ols} &= (E(X_iX_i^T))^{-1}E(X_iY_i) \end{aligned}$$

Let $E_n[f(X_i)] = \frac{1}{n}\sum_i f(X_i)$, then we have a plug in estimator for $\beta_{OLS} = E[X_iX_i^T]^{-1}E[X_iY_i]$:

$$\begin{aligned} \hat{\beta} &= argmin_\beta \frac{1}{n}\sum_i (Y_i - X_i^T\beta) \\ &= argmin_\beta E_n[(Y_i - X_i^T\beta)^2] \\ &= E_n[X_iX_i^T]^{-1}E_n[X_iY_i] \\ &= [\frac{1}{n}\sum_i X_iX_i^T]^{-1}[\frac{1}{n}\sum_i X_iY_i] \\ &= (X^TX)^{-1}X^TY \end{aligned}$$

1.2

1.2a Let $Y_i = X_i\beta + e_i$,

$$\begin{aligned} \hat{\beta}_{OLS} &= [\frac{1}{n}\sum_i X_iX_i^T]^{-1}[\frac{1}{n}\sum_i X_iY_i] \\ &= [\frac{1}{n}\sum_i X_iX_i^T]^{-1}[\frac{1}{n}\sum_i X_i(X_i^T\beta + e_i)] \\ &= [\frac{1}{n}\sum_i X_iX_i^T]^{-1}[\frac{1}{n}\sum_i X_i(X_i^T\beta)] + [\frac{1}{n}\sum_i X_iX_i^T]^{-1}[\frac{1}{n}\sum_i X_ie_i] \\ &= \beta + [\frac{1}{n}\sum_i X_iX_i^T]^{-1}[\frac{1}{n}\sum_i X_ie_i] \end{aligned}$$

Since the other parameters are constants, only e_i is random variable, $E(\hat{\beta}_{OLS}|X) = \beta + [\frac{1}{n}\sum_i X_iX_i^T]^{-1}[\frac{1}{n}\sum_i X_iE(e_i|X_i)]$.

1.2b No, it is not in general an unbiased estimator. Because e_i is only uncorrelated with X_i but not mean independent, $E(e_i|X_i)$ is usually not 0. Therefore, $\hat{\beta}_{OLS} \neq \beta$. The exception to this is when CEF is linear, meaning $e_i = \epsilon_i$ such that $E(e_i|X_i) = E(\epsilon_i|X_i) = 0$ and $\hat{\beta}_{OLS} = \beta$.

1.2c If $E(Y_i|X_i) = X_i^T\beta_{OLS}$, $\epsilon_i = Y_i - E(Y_i|X_i) = Y_i - X_i^T\beta_{OLS} = e_i$. Therefore, $E(e_i|X_i) = E(\epsilon_i|X_i) = 0$ by decomposition properties and $\hat{\beta}_{OLS} = \beta + 0 = \beta$. Therefore, $\hat{\beta}_{OLS}$ is conditionally unbiased. Also, by tower property,

$$E(\hat{\beta}_{OLS}) = E(E(\hat{\beta}_{OLS}|X)) = E(\beta) = \beta$$

1.2d Since $Var(\hat{\beta}_{OLS}|X) = [\frac{1}{n}\sum_i X_i X_i^T]^{-1}[\frac{1}{n}\sum_i X_i X_i^T Var(e_i|X_i)][\frac{1}{n}\sum_i X_i X_i^T]^{-1}$, and we know that $Var(e_i|X_i) = Var(\epsilon_i|X_i) = Var(Y_i|X_i) = \sigma^2$, we have

$$Var(\hat{\beta}_{OLS}|X) = \frac{\sigma^2}{n}\sum_i X_i X_i^T]^{-1} = \sigma^2(X^TX)^{-1}$$

1.2e Since $\hat{\beta}_{OLS} = \beta + [\frac{1}{n}\sum_i X_i X_i^T]^{-1}[\frac{1}{n}\sum_i X_i e_i]$ and e_i is normally distributed, by properties of normal distribution we know that $\hat{\beta}_{OLS}|X_i$ is also normally distributed, where the mean and variance are calculated above. Therefore, $\hat{\beta}_{OLS}|X_i \sim N(\beta, \sigma^2(X^TX)^{-1})$.

Q2

2.1

```
rm(list = ls())
#install.packages('sensemakr')
library('sensemakr')
```

2.1a

```
## See details in:
## Carlos Cinelli and Chad Hazlett (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias
df <- darfur
#str(df)
```

$ATE = E(Y_i(1) - Y_i(0))$. $Y_i(1)$ is the outcome in the world where we set by intervention the treatment to $D_i = 1$, $Y_i(0)$ is the outcome in the world where we set by intervention the treatment to $D_i = 0$.

2.2

2.2a With the claim that village and female are sufficient for control of confounding, $Y_i(d) \perp D_i|V_i, F_i$, such that $E(Y(d)) = E(E(Y(d)|V_i, F_i)) = E(E(Y(d)|D = d, V_i, F_i))$ where $\perp$ means independence, and $ATE = E(Y_i(1) - Y_i(0)) = E(E(Y(1)|D = 1, V_i, F_i)) - E(E(Y(0)|D = 0, V_i, F_i))$.

2.2b With consistency, we have $D_i = d$, $Y_i(d) = Y_i$. Thus, $E(Y(d)) = E(E(Y(d)|V_i, F_i)) = E(E(Y(d)|D = d, V_i, F_i)) = E(E(Y|D = d, V_i, F_i))$. $ATE = E(Y_i(1) - Y_i(0)) = E(E(Y_i|D_i = 1, V_i, F_i)) - E(E(Y_i|D_i = 0, V_i, F_i))$. Yes, as shown in the formula that we can identify the ATE from the observaed data. For estimation, we can use the plug-in estimator $\hat{ATE} = \frac{1}{n}\sum_i(\hat{E}(Y_i|D_i = 1, V_i, F_i) - \hat{E}(Y_i|D_i = 0, V_i, F_i))$.

2.2c X_i is not necessary for identification of the ATE, since it is not even in the formula. Based on our assumption, village and female are sufficient for control of confounding, so the inclusion of X_i is not necessary. However, we still want to include it in our regression because our assumption is possibly wrong and the variables in X may help us identify the problems. We also do not want to create an omitted variable biased by not including something.

2.3

2.3a $ATE = E(Y_i(1) - Y_i(0)) = E(\tau_1 + \beta_{1f}F_i + V_i^T\beta_{1v} + X_i^T\beta_{1x}) - E(\beta_{1f}F_i + V_i^T\beta_{1v} + X_i^T\beta_{1x}) = E(\tau_1) = \tau_1$

```
library(tidyverse)
```

2.3b

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ols <- lm(peacefactor ~ directlyharmed + female + village + herder_dar
          + farmer_dar + age + pastvoted, data = df)
coefficients(ols)[2]
```

```
## directlyharmed
##     0.09751154
```

```
# Verify using plug in estimate
pred_df = select(df, c(female, village, herder_dar, farmer_dar, age, pastvoted))

Ey1 <- mean(predict(ols, newdata = data.frame(directlyharmed = 1, pred_df)))
#mean(predict(ols, newdata = data.frame(directlyharmed = 1, female = df$female, village =df$village, he
Ey0 <- mean(predict(ols, newdata = data.frame(directlyharmed = 0, pred_df)))
ATE <- Ey1 - Ey0
ATE
```

```
## [1] 0.09751154
```

Per the specification above, the ATE can be identified by τ_1, which is equal to 0.09751154. This is consistent with the method using the plug-in estimate.

```
ols1 <- lm(peacefactor ~ female + village + herder_dar + farmer_dar
           + age + pastvoted, data = df)
ols2 <- lm(directlyharmed ~ female + village + herder_dar + farmer_dar
           + age + pastvoted, data = df)
y_tilda <- residuals(ols1)
d_tilda <- residuals(ols2)

# FWL
ols3 <- lm(y_tilda ~ d_tilda)
c(coefficients(ols3)[2], coefficients(ols)[2])
```

2.3c

```
##        d_tilda directlyharmed
##     0.09751154     0.09751154
```

Indeed, the estimates from FWL and regression are identical.

```
# signficance level
alpha <- 0.05
```

```
B <- 1000
ATE.boot <- rep(NA,B)
set.seed(42)

# bootstrap CI
for (i in 1:B){
 idx      <- sample(nrow(df), replace = T)
 d.boot   <- df$directlyharmed[idx]
 x1.boot  <- df$age[idx]
 x2.boot  <- df$female[idx]
 x3.boot  <- df$herder_dar[idx]
 x4.boot  <- df$farmer_dar[idx]
 x5.boot  <- df$pastvoted[idx]
 v.boot   <- df$village[idx]
 y.boot   <- df$peacefactor[idx]
 ols.boot <- lm(y.boot ~ d.boot + x1.boot + x2.boot + x3.boot + x4.boot +x5.boot + v.boot)
 Ey1.boot <- mean(predict(ols.boot, newdata = data.frame(d.boot = 1,
                      x1.boot = x1.boot, x2.boot = x2.boot, x3.boot = x3.boot,
                      x4.boot = x4.boot, x5.boot = x5.boot, v.boot = v.boot)))
 Ey0.boot <- mean(predict(ols.boot, newdata = data.frame(d.boot = 0,
                      x1.boot = x1.boot, x2.boot = x2.boot, x3.boot = x3.boot,
                      x4.boot = x4.boot, x5.boot = x5.boot, v.boot = v.boot)))
 ATE.boot[i] <- Ey1.boot - Ey0.boot
}

# Quantile
CI_quantile <- quantile(ATE.boot, c(alpha/2, 1-alpha/2))
CI_quantile
```

2.3d

```
##       2.5%      97.5%
## 0.04707136 0.14167567
```

```
# Classical
norm_z <- c(qnorm(0.025),qnorm(0.975))
Ci_helper <- norm_z * sd(ATE.boot)
CI_classical <- Ci_helper + coefficients(ols)[2]
CI_classical
```

```
## [1] 0.04945219 0.14557088
```

```
#c(CI_quantile, CI_classical)
```

The quantile CI, [0.04707136, 0.14167567], is more toward the left than the classical CI, [0.04945219, 0.14557088], but they are pretty similar and both include our coefficient from OLS: 0.09751154.

2.3e Since the CEF can be reasonably approximated with a linear function of covariates, the finding correlation-wise is that being injured in the war is associated with a 0.09751154 increase in the prediction of people's willingness to make peace. The counterfactual-wise explanation, under our assumption that village and female are sufficient for control of confounding, is that we may claim that being injured in the war causes an individual a 0.09751154 increase in the prediction of their willingness to make peace.

2.4

2.4a $ATE = E(Y_i(1) - Y_i(0)) = E(\tau_2 + \beta_{2f}F_i + V_i^T\beta_{2v} + \beta_{2fd}F_i + X_i^T\beta_{2x}) - E(\beta_{2f}F_i + V_i^T\beta_{2v} + X_i^T\beta_{2x}) = \tau_2 + \beta_{2fd}E(F_i)$. It is more complicated than a single regression coefficient.

```
ols_inter <- lm(peacefactor ~ directlyharmed * female + village + herder_dar + farmer_dar + age + pastv

# Numeric sanity check
# coef(ols_inter)
# mean(df$female)
 0.4561129 * 0.0130072482 + 0.0918138865
```

2.4b

```
## [1] 0.09774666
```

```
0.4561129 * 0.0130072482
```

```
## [1] 0.005932774
```

```
# Verify using plug in estimate
pred_df = select(df, c(female, village, herder_dar, farmer_dar, age, pastvoted))

Ey1_inter <- mean(predict(ols_inter, newdata = data.frame(directlyharmed = 1, pred_df)))
Ey0_inter <- mean(predict(ols_inter, newdata = data.frame(directlyharmed = 0, pred_df)))
ATE1 <- Ey1_inter - Ey0_inter
ATE1
```

```
## [1] 0.09774666
```

We have ATE of 0.09774666.

```
# signficance level
alpha <- 0.05
B <- 1000
ATE.boot1 <- rep(NA,B)
set.seed(42)

# bootstrap CI
for (i in 1:B){
 idx      <- sample(nrow(df), replace = T)
 d.boot   <- df$directlyharmed[idx]
 x1.boot  <- df$age[idx]
 x2.boot  <- df$female[idx]
 x3.boot  <- df$herder_dar[idx]
 x4.boot  <- df$farmer_dar[idx]
 x5.boot  <- df$pastvoted[idx]
 v.boot   <- df$village[idx]
 y.boot   <- df$peacefactor[idx]
 ols.boot <- lm(y.boot ~ x1.boot + d.boot * x2.boot + x3.boot + x4.boot +x5.boot + v.boot)
 Ey1.boot <- mean(predict(ols.boot, newdata = data.frame(d.boot = 1,
                  x1.boot = x1.boot, x2.boot = x2.boot, x3.boot = x3.boot,
                  x4.boot = x4.boot, x5.boot = x5.boot, v.boot = v.boot)))
 Ey0.boot <- mean(predict(ols.boot, newdata = data.frame(d.boot = 0,
                  x1.boot = x1.boot, x2.boot = x2.boot, x3.boot = x3.boot,
                  x4.boot = x4.boot, x5.boot = x5.boot, v.boot = v.boot)))
```

```
 ATE.boot1[i] <- Ey1.boot - Ey0.boot
}

# Quantile
CI_quantile1 <- quantile(ATE.boot1, c(alpha/2, 1-alpha/2))
CI_quantile1
```

2.4c

```
##       2.5%      97.5%
## 0.04751783 0.14238240
```

```
# Classical
norm_z <- c(qnorm(0.025),qnorm(0.975))
Ci_helper1 <- norm_z * sd(ATE.boot1)
CI_classical1 <- Ci_helper + ATE1
CI_classical1
```

```
## [1] 0.04968732 0.14580600
```

```
#c(CI_quantile1, CI_classical1)
```

Using quantiles, confidence interval is [0.04751783, 0.14238240]. Classical method: [0.04968732, 0.14580600].

2.4d It adds to the previous finding. With this new model, we can observe that injured women are predicted to feel stronger about peace than injured men by 0.005932774 (calculated using the formula in (a)). But the trend is still that being injured in the war is associated with an increase in the prediction of their willingness to make peace, and the average total effect is 0.09774666. The explanation, under our assumption that village and female are sufficient for control of confounding, counterfactual-wise is that we nay claim that being injured in the war causes an individual a 0.09774666 increase in the prediction of their willingness to make peace.

2.5

2.5a With the claim that village, female, and center are sufficient for control of confounding, $Y_i(d) \perp D_i|V_i, F_i, C_i$, such that $E(Y(d)) = E(E(Y(d)|V_i, F_i, C_i)) = E(E(Y(d)|D = d, V_i, F_i, C_i))$ where $\perp$ means independence, and $ATE = E(Y_i(1) - Y_i(0)) = E(E(Y(1)|D = 1, V_i, F_i, C_i)) - E(E(Y(0)|D = 0, V_i, F_i, C_i))$.

2.5b Mathematically, $bias = \tau_1 - \tau = \gamma\delta$. That means introducing the hypothetical confounder C_i changes our inference by giving us an increase in the coefficient of D_i by $\gamma\delta$.

γ is the predictive impact of the omitted variable C on Y after adjusting for W, and namely it is the regression coefficient of C in the regression $E(Y_i|D_i, F_i, V_i, X_i, C_i) = \tau D_i + \beta_f F_i + V_i^T \beta_v + X_i^T \beta_x + \gamma C_i$.

δ is the imbalance of C across levels of D, after adjusting for W where $W_i = [F_i, V_i, X_i]$. Namely, it is the regression coefficient of D on the regression $C \sim D + W$. It is called the imbalance because δ represents the difference between the treated (D = 1) and the non-treated (D = 0).

2.5c For this question we simply need to estimate τ_1 using the model in specification 1 (Q2.3) and minus $0.2 * 0.2 = 0.04$.

```
# signficance level
alpha <- 0.05
B <- 1000
tau.boot <- rep(NA,B)
set.seed(42)
```

```
# bootstrap CI
for (i in 1:B){
 idx      <- sample(nrow(df), replace = T)
 d.boot   <- df$directlyharmed[idx]
 x1.boot  <- df$age[idx]
 x2.boot  <- df$female[idx]
 x3.boot  <- df$herder_dar[idx]
 x4.boot  <- df$farmer_dar[idx]
 x5.boot  <- df$pastvoted[idx]
 v.boot   <- df$village[idx]
 y.boot   <- df$peacefactor[idx]
 ols.boot <- lm(y.boot ~ d.boot + x1.boot + x2.boot + x3.boot
                + x4.boot +x5.boot + v.boot)
 tau.boot[i] <- coefficients(ols.boot)[2] -0.04
}

# Quantile
CI_quantile2 <- quantile(tau.boot, c(alpha/2, 1-alpha/2))
CI_quantile2
```

```
##       2.5%      97.5%
## 0.00707136 0.10167567
```

The confidence interval is [0.00707136, 0.10167567]. Since the interval does not contain 0 yet, we may still maintain the results found in previous part of this study. But since the lower bound is quite close to 0, we need to be careful of other potential confounders. If they exist, it is likely to drag the confidence interval down to contain 0, which makes the result not significant.

2.6

2.6a $Bias^2 = (\gamma\delta)^2 = \frac{R^2_{Y\sim Z|D,W} R^2_{D\sim Z|W}}{1-R^2_{D\sim Z|W}} \frac{Var(Y^{\perp D,W})}{Var(D^{\perp W})}$ where $W_i = [F_i, V_i, X_i]$. The second part $\frac{Var(Y^{\perp D,W})}{Var(D^{\perp W})}$ can be estimated from the data because they only include known variables. While Z is unobserved, the first part, $\frac{R^2_{Y\sim Z|D,W} R^2_{D\sim Z|W}}{1-R^2_{D\sim Z|W}}$, of the expression needs to be limited by hypothesis regarding the strength of confounding.

2.6b Based on the formula, we know that the upper bound for the bias is $\sqrt{\frac{R^2_{Y\sim Z|D,W} R^2_{D\sim Z|W}}{1-R^2_{D\sim Z|W}} \frac{Var(Y^{\perp D,W})}{Var(D^{\perp W})}} = \sqrt{\frac{0.12\cdot 0.01}{0.99}} \frac{SD(Y^{\perp D,W})}{SD(D^{\perp W})}$

```
ols_1 <- lm(peacefactor ~ directlyharmed + female + village + herder_dar + farmer_dar
          + age + pastvoted, data = df)
ols_2 <- lm(directlyharmed ~ female + village + herder_dar + farmer_dar
          + age + pastvoted, data = df)

# Calcualte upper bound
sd_ratio = sd(residuals(ols_1))/sd(residuals(ols_2))
r2 <- sqrt(0.12 * 0.01 /0.99)
bias_bound <- sd_ratio * r2

# tau
tau_estimate <- coefficients(ols_1)[2] - bias_bound

library(sensemakr)
```

```r
sens_estimate <- adjusted_estimate(ols_1, 'directlyharmed', r2dz.x = 0.01, r2yz.dx = 0.12)

c(tau_estimate, sens_estimate)
```

```
## directlyharmed directlyharmed
##      0.0748508      0.0748508
```

Since the direction of bias reduces the magnitude of τ, the result from two methods are less than original coefficient but identical: 0.0748508.

```r
# signficance level
alpha <- 0.05
B <- 1000
tau.boot1 <- rep(NA,B)
set.seed(42)
r2 <- sqrt(0.12 * 0.01 /0.99)

# bootstrap CI
for (i in 1:B){
 idx      <- sample(nrow(df), replace = T)
 d.boot   <- df$directlyharmed[idx]
 x1.boot  <- df$age[idx]
 x2.boot  <- df$female[idx]
 x3.boot  <- df$herder_dar[idx]
 x4.boot  <- df$farmer_dar[idx]
 x5.boot  <- df$pastvoted[idx]
 v.boot   <- df$village[idx]
 y.boot   <- df$peacefactor[idx]
 ols.boot <- lm(y.boot ~ d.boot + x1.boot + x2.boot + x3.boot
                + x4.boot +x5.boot + v.boot)
 ols.boot1 <- lm(d.boot ~ x1.boot + x2.boot + x3.boot
                + x4.boot +x5.boot + v.boot)
 sd_ratio = sd(residuals(ols.boot))/sd(residuals(ols.boot1))
 bias_bound <- sd_ratio * r2
 tau.boot1[i] <- coefficients(ols.boot)[2] - bias_bound
}

# Quantile
CI_quantile3 <- quantile(tau.boot1, c(alpha/2, 1-alpha/2))
CI_quantile3
```

2.6c

```
##       2.5%      97.5%
## 0.02411385 0.11851678
```

The confidence interval for τ is [0.02411385, 0.11851678]. The strength of confounding is still insufficient to change the main conclusions of the study, since 0 is not in the interval.