PSET6

Sarah Van Alsten

2/24/2020

Question 1.

a.

$$\tau = E[Y_{1i}|D_i = 1]Pr(D_i = 1) + E[Y_{1i}|D_i = 0]Pr(D_i = 0) - E[Y_{0i}|D_i = 1]Pr(D_i = 1) - E[Y_{0i}|D_i = 0]Pr(D_i = 0)$$

```
#create a table using real data
my.table <- as.data.frame(rbind(c(1, 500/800, 73, NA),
                                c(0, 300/800, NA, 56 )))

#rename the cols
names(my.table) <- c("d", "Pr_Di_d", "E_Y1_given_Di_d", "E_Y0_given_Di_d")

#print
my.table %>%
  knitr::kable() %>% kableExtra::kable_styling("striped")
```

d	Pr_Di_d	E_Y1_given_Di_d	E_Y0_given_Di_d
1	0.625	73	NA
0	0.375	NA	56

b.

```
#'Best' possible outcome = rating of 100
#'worst' possible outcome = rating of 0

sharp.lower <- as.data.frame(rbind(c(1, 500/800, 73, 100),
                                c(0, 300/800, 0, 56 )))

sharp.upper <- as.data.frame(rbind(c(1, 500/800, 73, 0),
                                c(0, 300/800, 100, 56 )))

lower.bound <- (73 - 100)*(.625) + (0 - 56)*(.375)
upper.bound <- (73 - 0)*(.625) + (100 - 56)*(.375)
```

The sharp bounds of the ATE would be -37.875 to 62.125. These bounds represent the range of all possible treatment effects given the data that we observed.

c. This assumption implies monotone treatment selection, meaning that treated units are expected to have better outcomes/higher self-efficacy than control units under either treatment assignment:

$$E[Y_{0i}|D = 1] \geq E[Y_{0i}|D = 0]$$

$$E[Y_{1i}|D = 1] \geq E[Y_{1i}|D = 0]$$

and then:

$$\tau \leq E[Y_i|D_i = 1]Pr(D_i = 1)+E[Y_{1i}|D_i = 0]Pr(D_i = 0)-E[Y_{0i}|D_i = 1]Pr(D_i = 1)-E[Y_i|D_i = 0]Pr(D_i = 0)$$

```
#the actual treatment effect should thus be the upper bound:
#expected value for D = 1 - expected value for D = 0
73-56
```

```
## [1] 17
```

The new bounds would thus be -37.875 to 17 .

d. This assumption states that:

$$E[Y_{1i}|D = 1, X = 3] = E[Y_{1i}|D = 0, X = 3]$$

$$E[Y_{0i}|D = 1, X = 3] = E[Y_{0i}|D = 0, X = 3]$$
$$E[Y_{1i}|D = 1, X = 2] \geq E[Y_{1i}|D = 0, X = 2]$$
$$E[Y_{0i}|D = 1, X = 2] \geq E[Y_{0i}|D = 0, X = 2]$$
$$E[Y_{1i}|D = 1, X = 1] \geq E[Y_{1i}|D = 0, X = 1]$$
$$E[Y_{0i}|D = 1, X = 1] \geq E[Y_{0i}|D = 0, X = 1]$$

For the upper bound this then becomes:

$$\begin{aligned}\tau \leq \sum_{X \in x}(&E[Y_i|D_i = 1, X_i = x]Pr(D_i = 1, X_i = x)+\\ &E[Y_{1i}|D_i = 0, X_i = x]Pr(D_i = 0, X_i = x)-\\ &E[Y_{0i}|D_i = 1, X_i = x]Pr(D_i = 1, X_i = x)-\\ &E[Y_i|D_i = 0, X_i = x]Pr(D_i = 0, X_i = x))\end{aligned}$$

And for the lower bound,

$$\begin{aligned}\tau \geq \sum_{X \in x}(&E[Y_i|D_i = 1, X_i = x]Pr(D_i = 1, X_i = x) - \underline{Y}(Pr(D_i = 1, X_i = x))+\\ &\overline{Y}(Pr(D_i = 0, X_i = x)) - (E[Y_i|D_i = 0, X_i = x]Pr(D_i = 0, X_i = x))\end{aligned}$$

Given what we know about individuals with high incomes, we can also simplify further, such that the upper bound becomes

$$\begin{aligned}\tau \leq\ &E[Y_i|D_i = 1]Pr(X_i = 3)-\\ &E[Y_i|D_i = 0]Pr(X_i = 3)+\\ \sum_{X \in (1,2)}(&E[Y_{1i}|D_i = 1, X_i = x]Pr(D_i = 1, X_i = x)+\\ &E[Y_{1i}|D_i = 0, X_i = x]Pr(D_i = 0, X_i = x)-\\ &E[Y_{0i}|D_i = 1, X_i = x]Pr(D_i = 1, X_i = x)-\\ &E[Y_{0i}|D_i = 0, X_i = x]Pr(D_i = 0, X_i = x))\end{aligned}$$

and the lower bound becomes

$$\begin{aligned}\tau \geq\ &(E[Y_i|D_i = 1, X_i = 3] - E[Y_i|D_i = 0, X_i = 3])(Pr(X_i = 3)+\\ \sum_{X \in 1,2}(&E[Y_i|D_i = 1, X_i = x]Pr(D_i = 1, X_i = x) - \underline{Y}(Pr(D_i = 1, X_i = x))+\\ &\overline{Y}(Pr(D_i = 0, X_i = x)) - (E[Y_i|D_i = 0, X_i = x]Pr(D_i = 0, X_i = x))\end{aligned}$$

The resulting bounds would be less credible than those expressed in part c, because we are making an additional assumption that further narrows our range of possible expected treatment effects. Instead of the less strict assumption that all treated units have expected outcomes greater than or equal to control units (irrespective of income), we added the constraint for those with high incomes that the potential outcomes for treated units if they got control is equal to the potential outcome for control under control and the potential outcomes for control units if they got treatment is equal to the potential outcome for treated units that got treatment. This makes the bound smaller, because we have imputed the "missing" outcomes for high income units rather than assuming the worst case scenario. This also follows the Law of Decreasing Credibility, in that as we make stronger assumptions, our inference is less credible.

e.

```
#calculate new bounds
new.bound <-
#stratum for high income
(83 - 70)*(160/800) +
#middle income stratum: worst case scenario
(73)*(163/800) + (0)*(77/800) - (100)*(163/800) - (55)*(77/800) +
#low income stratum : worst case scenario
(69)*(163/800) + (0)*(212/800) - (100)*(188/800) - (51)*(212/800)
```

Now, the estimated ATE is bounded from -31.15125 to‘ 17.

Question 2.

a.

```
#estimate ATE
edu.mod <- glm(educ ~ abd + age + fthr.ed + mthr.ed + C.ach,
               data = soldier)

#print results
stargazer::stargazer(edu.mod, type = "latex", float = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Mar 01, 2020 - 5:56:24 PM

	Dependent variable:
	educ
abd	-0.554^{***}
	(0.214)
age	0.046^{**}
	(0.021)
fthr.ed	0.159^{***}
	(0.031)
mthr.ed	0.053
	(0.037)
C.ach	-0.716^{**}
	(0.299)
Constant	5.464^{***}
	(0.496)
Observations	741
Log Likelihood	−1,810.666
Akaike Inf. Crit.	3,633.333
Note:	*p<0.1; **p<0.05; ***p<0.01

The necessary identification assumptions to interpret the ATE estimate are that there are no omitted confounders of the association between abduction ane education, that there is conditional ignorability for those who were versus weren't abducted given parents' education, residency in Acholibur, and age, that we have not conditioned on any colliders, that the effect of abduction is the same for all treated units, and that there is no interference between units.

Additionally, in order for the OLS estimator $\widehat{\beta}$ to be an unbiased estimator of the ATE, there needs to be no collinearity between our predictors (confounders), independence of observations, linearity in the relationship between education and predictors, and normally distributed error terms.

b. One potential unobserved confounder would be rurality of childhood home. Individuals growing up in very rural areas were more secluded from villages and were more likely to be abducted. Additionally, rurality could affect educational attainment, perhaps because schools are less accessible to individuals living far from city centers. The DAG would be as follows:

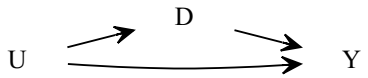


Figure 1: DAG

c.

d. The expected bias would be:

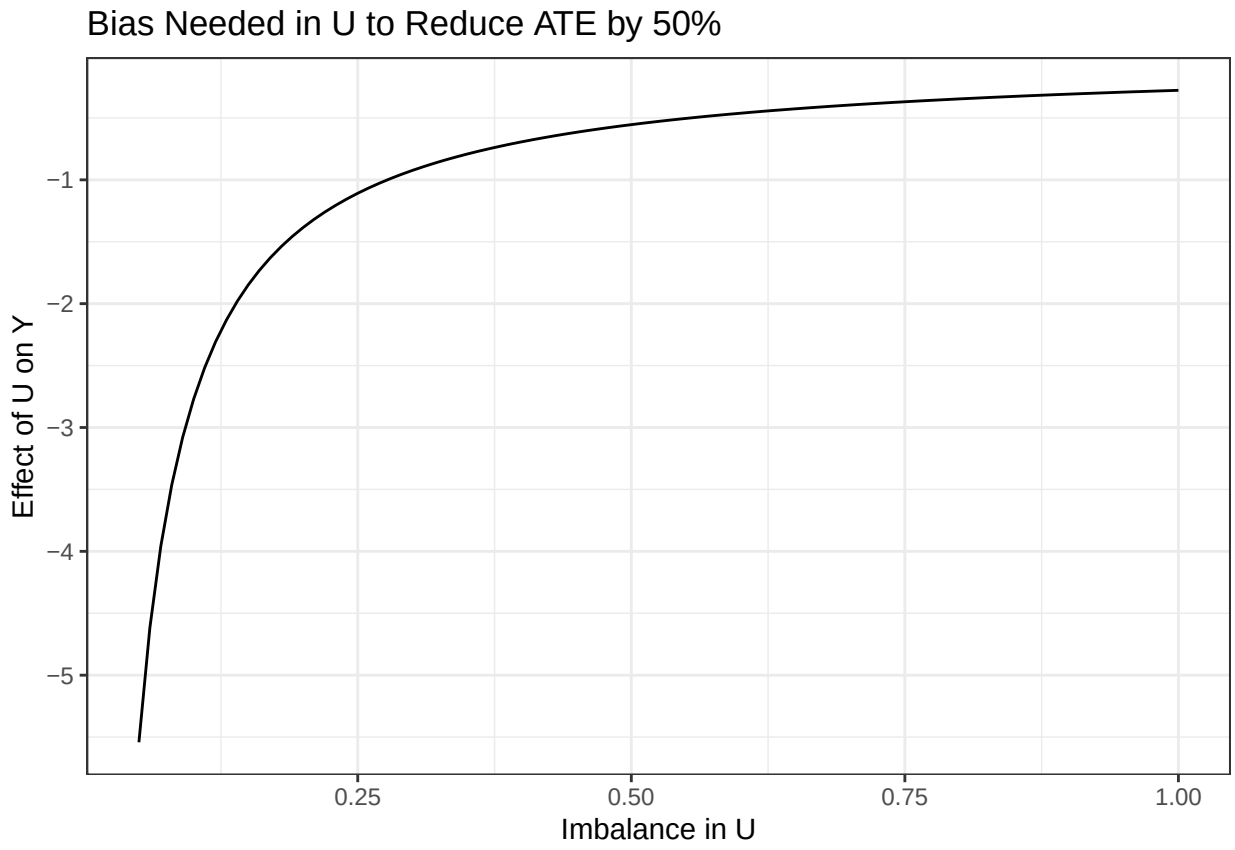
$$E[\hat{\tau}] - \tau = \{Pr(U_i = 1|D_i = 1) - Pr(U_i = 1|D_i = 0)\} * \{E[Y_i|U_i = 1] - E[Y_i|U_i = 0]\}$$

The assumption that the effect of U on Y is the same for treated and untreated units in this analysis is probably not plausible in this context. For instance, living in a rural area may not make as large of a difference to those who were abducted if they are less likely to return to their childhood homes (for instance, because their families do not welcome them after the violence), they may end up living closer to cities where it is easier to access education.

ii.

```
#the actual estimate for abduction = -0.55420
change.edu <- (-.55420)/2

#create vector of U imbalances
u.imbalance <- seq(.05, 1, .01)

#get value of effects of U on Y that would solve to equal change.edu
u.gamma <- change.edu/u.imbalance

#plot it
as.data.frame(cbind(u.imbalance, u.gamma)) %>%
  ggplot(aes(x = u.imbalance, y = u.gamma)) +
```

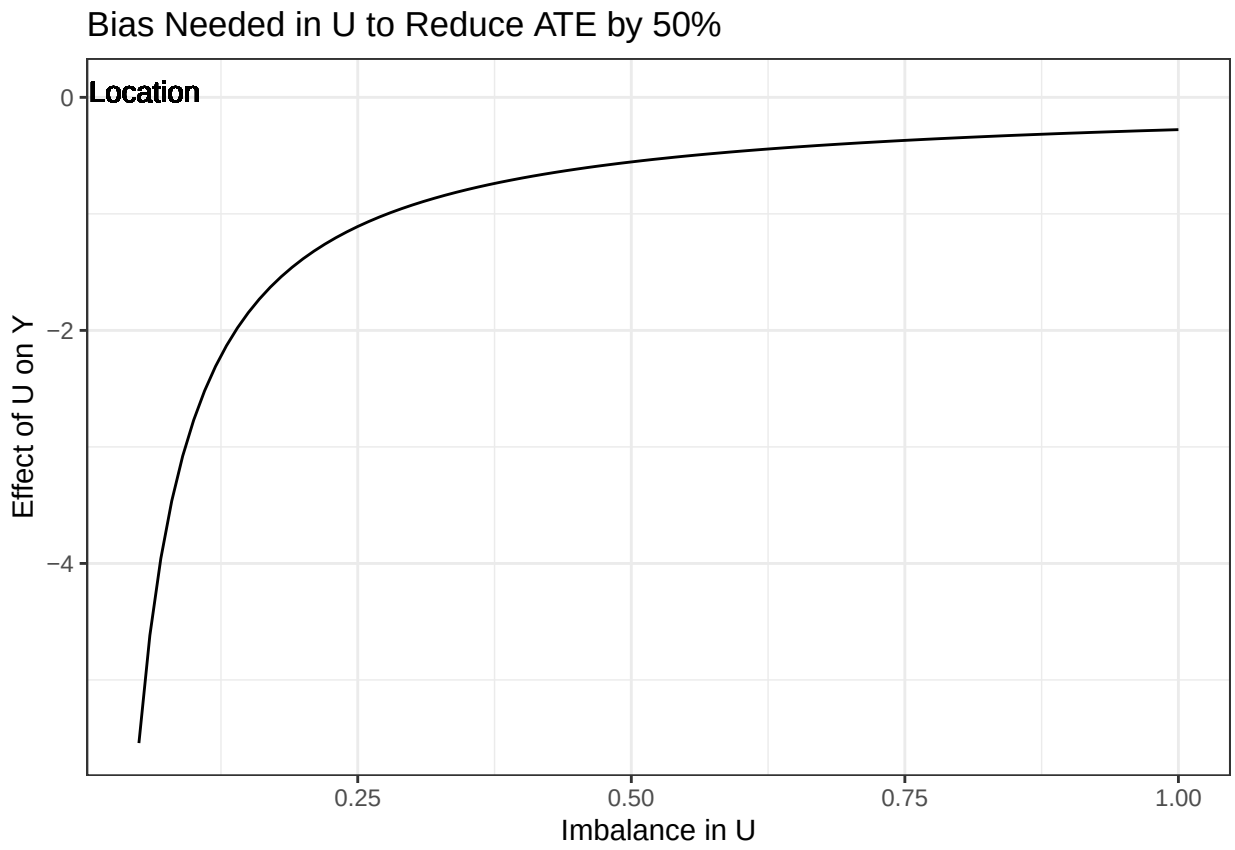
```
  geom_path() +
  theme_bw() +
  labs(x = "Imbalance in U",
       y = "Effect of U on Y")+
  ggtitle("Bias Needed in U to Reduce ATE by 50%")
```



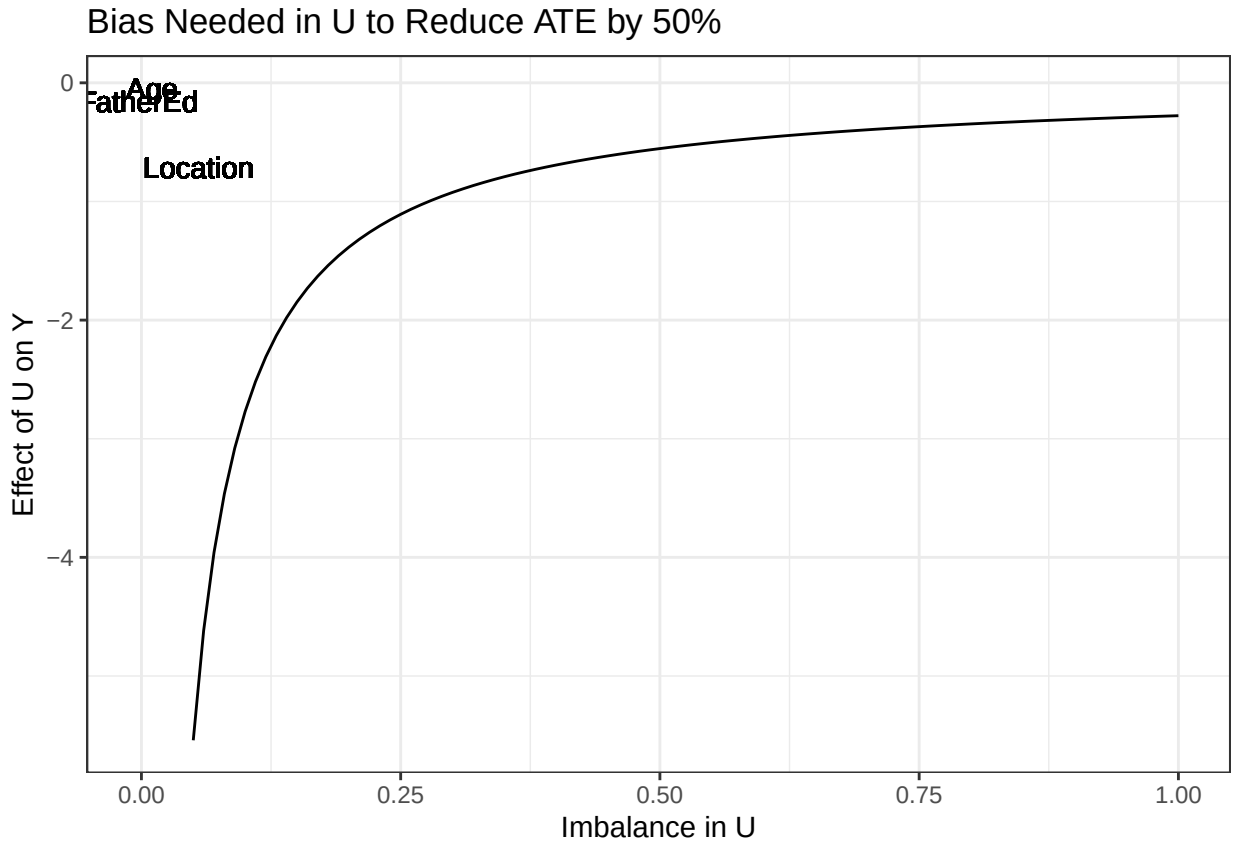
iii.

```
#for each covariate, get imbalance and gamma
#gammas are the regression coefficients
gamm.coef <- coefficients(edu.mod)

#now get rlship btwn covar and abduction
covar.mod <- lm(abd ~ age + fthr.ed + mthr.ed + C.ach,
                data = soldier)

imb.coef <- coefficients(covar.mod)

#plot it with location in Acholibur
as.data.frame(cbind(u.imbalance, u.gamma)) %>%
  ggplot(aes(x = u.imbalance, y = u.gamma)) +
  geom_path() +
  geom_text(aes(x = imb.coef[5], y =gamm.coef[5], label = "Location"))+
  theme_bw() +
  labs(x = "Imbalance in U",
```

```
      y = "Effect of U on Y")+
  ggtitle("Bias Needed in U to Reduce ATE by 50%")
```




```
#plot it with dad education and age
as.data.frame(cbind(u.imbalance, u.gamma)) %>%
  ggplot(aes(x = u.imbalance, y = u.gamma)) +
  geom_path() +
  geom_text(aes(x = imb.coef[5], y = gamm.coef[6], label = "Location"), size = 4)+
  geom_text(aes(x = imb.coef[2], y = -gamm.coef[3], label = "Age"), size = 4)+
  geom_text(aes(x = imb.coef[3], y = -gamm.coef[4], label = "FatherEd"), size = 4)+
  theme_bw() +
  labs(x = "Imbalance in U",
       y = "Effect of U on Y")+
  ggtitle("Bias Needed in U to Reduce ATE by 50%")
```



The observed confounders (age, father's education, and location in Acholibur) have a relatively small effect on the outcome, as shown in the plot. For instance, location in Acholibur, which has the strongest effect, only has a net efect of -.76, with an imbalance of ~5%. The effect of the unobserved confounder would need to be about 7 times as large as the effect of location (with a similar imbalance) to reduce the ATE by 50%. This suggests that it is unlikely that the observed relationship is driven entirely by unobserved confounding, because we would expect that area of residence has a large impact on education and abduction, and it is unlikely that the unobserved confounder would have an association 7 times as large as location. Similar conclusions could also be drawn by comparing the effect of the needed confounder to age or father's education.