Problem Set 5: Observational Studies

GOV 2003

Due at 11:59 pm (ET) on Oct 20, 2021

Instruction

Before you begin, please read the following instructions **carefully**:

- **No late submission is allowed** without prior approval from the instructors.
- **All answers should be typed up**. We recommend the use of `Rmarkdown`. A `Rmarkdown` template for this problem set is provided. Answers to analytical solutions should also be typed up.
- **A PDF copy of your answer** including your computer code should be uploaded to `Gradescope` before the deadline. **Do not submit the markdown file itself**.

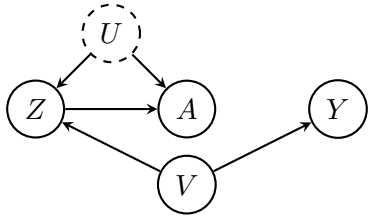
Introduction

This problem set consists of two parts. In `Part A`, we will investigate **directed acyclic graphs** (DAGs) to describe the causal structure of variables. In `Part B`, we will investigate regression analysis for observational data.

Part A: Directed acyclic graphs (DAGs)

Question 1: D-separation (5 pts; 1 pt for each)

For this question you will use the following causal directed acyclic graph:

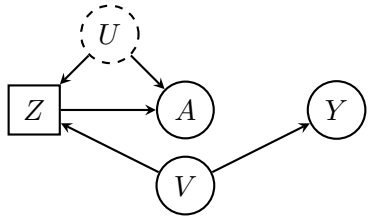


(a) List all of the paths between A and Y.

(b) If the above causal DAG is correct, would you expect there to be an association between A and Y? Briefly explain your answer.

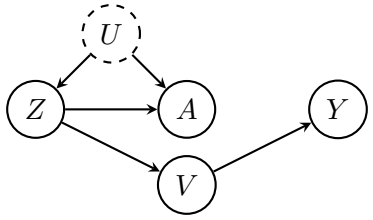
Hint: Identify confounder(s) and mediator(s) in each path.

(c) Suppose that we control for Z (either in a regression or by running our analysis within levels of Z). In this analysis, does A **d-separated** from Y by Z? Briefly explain your answer.

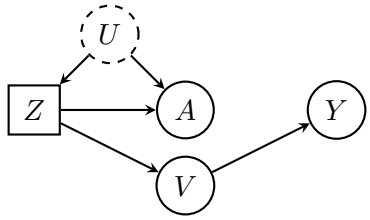


Hint: Think about whether Z is a confounder *or* mediator *or* collider in each path between A and Y.

(d) Suppose now that we flip the direction of the arrow from V to Z, so that $Z \rightarrow V$. Would you then expect to see an association between A and Y in this revised DAG? Briefly explain your answer.



(e) Suppose in this revised DAG we now control for Z. Does A **d-separated** from Y by Z in this revised DAG? Briefly explain your answer.



Answer 1

(a) Two paths:

- $A \leftarrow Z \leftarrow V \rightarrow Y$.
- $A \leftarrow U \rightarrow Z \leftarrow V \rightarrow Y$

(b) Yes, because of the path 1 above: there is a path with no colliders means there should be a marginal association.

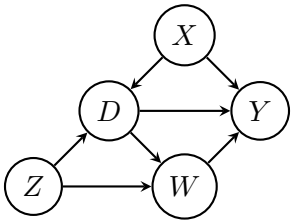
(c) No, because of the path 2 above. Controlling for a collider on that path "opens" the association between A and Y.

(d) Yes, the first path becomes $A \leftarrow Z \rightarrow V \rightarrow Y$ and still creates a marginal association between A and Y because there are no colliders on this path.

(e) Yes. Now A is independent of Y conditional on the value of Z because Z is not a collider on any path, so conditioning on it blocks the association over both paths.

Question 2: Backdoor criterion (5 pts; 1 pt for each)

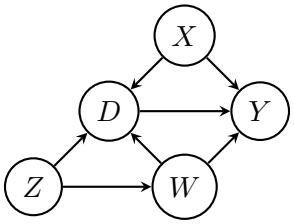
For this question you will use the following causal directed acyclic graph:



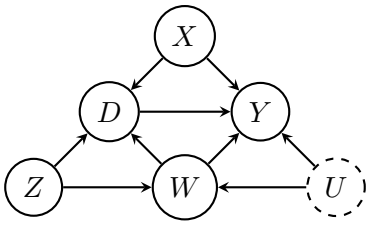
(a) List all of the **backdoor** paths from D to Y.

(b) Consider a set of nodes, $\mathbf{S} \in \{\emptyset, \{X\}, \{Z\}, \{W\}, \{X, Z\}, \{X, W\}, \{Z, W\}, \{X, Z, W\}\}$. List all the $\mathbf{S}$ such that **blocks** all the backdoor paths from D to Y. Is there any $\mathbf{S}$ which satisfies the **backdoor criterion**? Briefly explain your answer.

(c) Suppose now that we flip the direction of the arrow from D to W, so that $W \rightarrow D$. Is there any $\mathbf{S}$ which satisfies the **backdoor criterion** in this revised DAG? Briefly explain your answer.



(d) Suppose now that there exist an unobserved variable U in our DAG which is a confounder of W and Y as below. Note that we cannot condition on U since it is unobserved. Is there any $\mathbf{S}$ given which we can identify the average treatment effect of D on Y?



(e) Now, we use the `R` package `dagitty` (or `ggdag`) to answer the same question of (b). Visualize the original DAG with `dagitty`. Then, list all of the sets of variables that satisfy the backdoor criterion to identify the average treatment effect of D on Y using the package. Please include the codes in your answer.

Answer 2

(a) Two backdoor paths:

- $D \leftarrow X \rightarrow Y$
- $D \leftarrow Z \rightarrow W \rightarrow Y$

(b) Two sets of nodes block all the backdoor paths:

- $\{X, Z\}$: backdoor criterion is met since it includes no node that is a descend of D.
- $\{X, Z, W\}$: backdoor criterion is not met since W is a descend of D. (post-treatment bias)
- $\{X, W\}$: backdoor criterion is not met since W is a descend of D. (post-treatment bias)

(c) Note that there exist three backdoor paths:

- $D \leftarrow X \rightarrow Y$
- $D \leftarrow Z \rightarrow W \rightarrow Y$
- $D \leftarrow W \rightarrow Y$

Now only one set of nodes blocks all the backdoor paths:

- $\{X, Z, W\}$: backdoor criterion is met since it includes no node that is a descend of D.

(d) The backdoor paths are now as follows:

- $D \leftarrow X \rightarrow Y$
- $D \leftarrow Z \rightarrow W \rightarrow Y$
- $D \leftarrow W \rightarrow Y$
- $D \leftarrow Z \rightarrow W \leftarrow U \rightarrow Y$

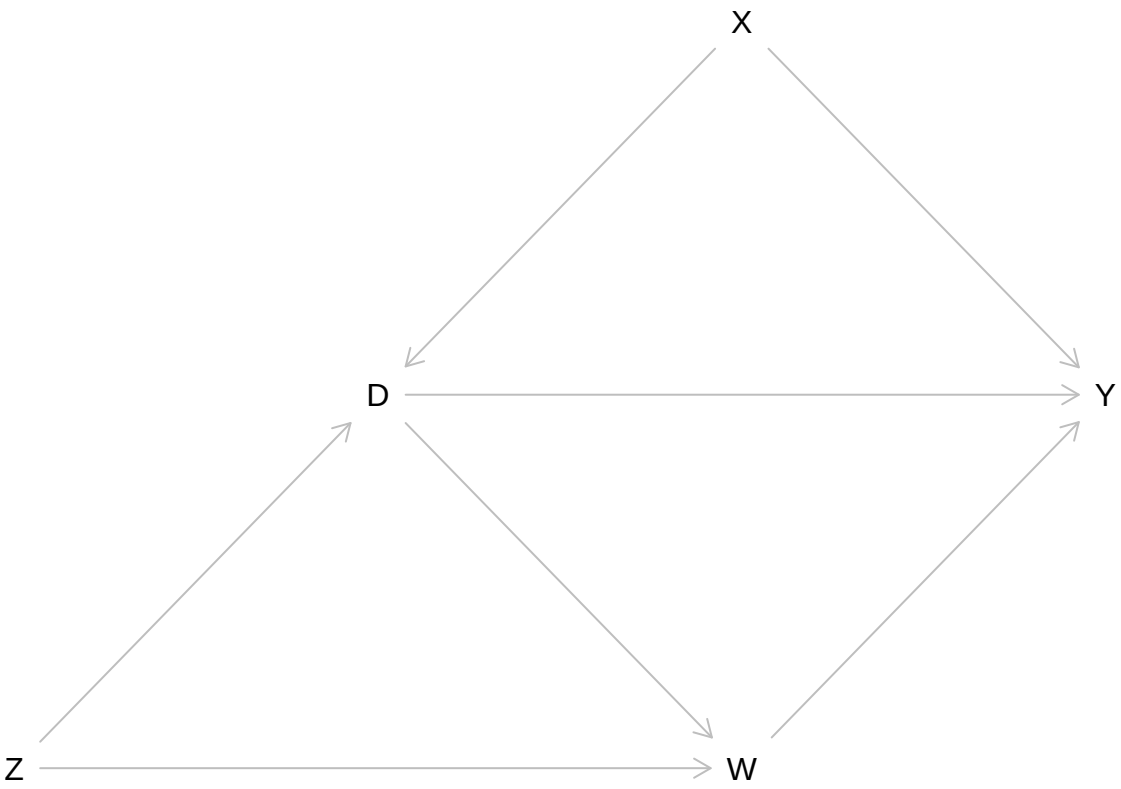
Now, conditioning on W "opens" the association between Z and U, and thus D and Y are dependent. However, we can block this association once we condition on Z as well. Thus, $\{X, Z, W\}$ still satisfies the backdoor criterion for identification.

(e)

```
library(dagitty)
g <- dagitty('dag {
D [pos="0,0"]
Y [pos="2,0"]
Z [pos="-1,1"]
```

```
W [pos="1,1"]
X [pos="1,-1"]
D -> Y
D -> W
W -> Y
Z -> D
Z -> W
X -> D
X -> Y
}')

plot(g)
```




```
adjustmentSets(g, "D", "Y", type="all")
```

```
## { X, Z }
```

```
# isAdjustmentSet(g, c("X","Z","W"), "D", "Y")
```

Part B: Regression estimation of the ATE

In this part, you will replicate and extend the analysis of Washington (2008), who estimated the effect of legislators having daughters on their votes on issues related to gender. The idea behind the paper is that, conditional on the number of children a couple has, a couple having one (or more) girls is randomly assigned. Thus, Washington argues that the effect of daughters is identified. She then looks at the effect of having daughters on members' of Congress scores from the National Organization for Women (NOW) and the American Association of University Women (AAUW). You will use `girls105`, a subset of data for the 105th Congress, from `Washington2008.RData`. The table at the end of this pdf describes the variables.

Question 3 (5 pts; 1 pt for each)

(a) Read through the paper. Replicate the findings for the 105th Congress that are found in Table 2, column 2. Note equations 1 and 2 on p. 315, which describe the specification.

Hint: Be aware that equation 1 includes fixed effects for the total number of children, and equation 2 additionally includes census region fixed effects.

(b) Replicate the results for the 105th Congress using an imputation estimator rather than a simple regression estimator, using `anygirls` as the treatment variable. Note that you may have to subset the data to enforce *common support*, which isn't necessary with the usual regression. Bootstrap this whole process to get standard errors.

Hints:

- First, subset the data to enforce *common support* (specifically, you will have to restrict the number of children and the religious groups).
- Run a separate regression model (of AAUW score on the covariates) in two subsets of the data (treated, `anygirls == 1`, and control, `anygirls == 0`).
- Then use predictions from these regressions to estimate the ATE. You can use the `predict()` function to do this.
- Use nonparametric bootstrap to get estimates of the variance (set the number of simulations to be 1000).

(c) How does the imputation estimator differ from the regression estimator in this case? Why features of the treatment effect might make these two estimators similar or different in this case? Which estimator should we prefer?

(d) Look at equation 2 of Washington (2008) and think about post-treatment bias. Identify any variables you deem post-treatment and re-run your analysis from part 1 above without those variables. How do your results change?

(e) Identify some reason(s) why Washington may have wanted to include those post-treatment variables. Which specification do you believe more accurately represents the effect of daughters and why?

Answer 3

(a)

```r
load("Washington2008.RData")

# STATA code for Table 2 Model 2 (anything similar to this specification counts as correct ans
# reg aauw ngirls white female repub  age agesq srvlng srvlngsq
#   reld1 reld3-reld5 chid2-chid11 regd* demvote  if congress==105
mod105 <- lm(aauw~ngirls + white + female + repub +
               age + I(age^2) + srvlng +I(srvlng^2) + demvote +
               factor(totchi) + factor(rgroup) +
               factor(region),
             data = girls105)
summary(mod105)
```

```
##
## Call:
## lm(formula = aauw ~ ngirls + white + female + repub + age + I(age^2) +
##     srvlng + I(srvlng^2) + demvote + factor(totchi) + factor(rgroup) +
##     factor(region), data = girls105)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.930 -11.232  -0.809   8.842  74.928
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       42.121424  24.611845   1.711 0.087773 .
## ngirls             2.384821   1.123848   2.122 0.034448 *
## white              0.143934   3.675846   0.039 0.968785
## female             9.193665   2.910140   3.159 0.001702 **
## repub            -60.468440   2.280175 -26.519  < 2e-16 ***
## age                0.854420   0.860012   0.993 0.321065
## I(age^2)          -0.006414   0.008232  -0.779 0.436377
## srvlng            -0.207689   0.323641  -0.642 0.521416
## I(srvlng^2)        0.003581   0.010908   0.328 0.742865
## demvote           62.147959  11.567940   5.372 1.32e-07 ***
## factor(totchi)1   -3.497354   3.738651  -0.935 0.350113
## factor(totchi)2   -4.408748   3.038459  -1.451 0.147565
## factor(totchi)3   -7.397466   3.526217  -2.098 0.036542 *
## factor(totchi)4  -10.871754   4.061698  -2.677 0.007740 **
## factor(totchi)5  -13.035875   5.093865  -2.559 0.010859 *
## factor(totchi)6  -17.865026   8.953045  -1.995 0.046672 *
## factor(totchi)7  -11.856900  10.376659  -1.143 0.253864
## factor(totchi)8  -43.702778  12.225549  -3.575 0.000393 ***
## factor(totchi)9  -15.928693  19.124351  -0.833 0.405395
## factor(totchi)10 -26.659436  18.646598  -1.430 0.153574
## factor(rgroup)1   -5.670655   7.605983  -0.746 0.456374
## factor(rgroup)2  -10.175492   7.595589  -1.340 0.181113
## factor(rgroup)3   -2.466313   8.860067  -0.278 0.780877
```

```
## factor(rgroup)4    4.011928   8.223420   0.488 0.625911
## factor(region)2   -9.286950   4.408594  -2.107 0.035775 *
## factor(region)3  -12.576558   4.421819  -2.844 0.004680 **
## factor(region)4   -7.869625   5.086779  -1.547 0.122632
## factor(region)5  -10.056063   4.565364  -2.203 0.028184 *
## factor(region)6  -19.685949   5.393746  -3.650 0.000297 ***
## factor(region)7  -15.546480   4.887668  -3.181 0.001583 **
## factor(region)8  -16.451153   5.450929  -3.018 0.002706 **
## factor(region)9   -8.656789   4.432920  -1.953 0.051532 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.44 on 402 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:   0.84,  Adjusted R-squared:  0.8276
## F-statistic: 68.06 on 31 and 402 DF,  p-value: < 2.2e-16
```

(b)

```
## there's some common support in these groups, but not very much
girls105 <- subset(girls105, rgroup != 0 & rgroup != 3 & rgroup != 4)
## there's almost no common support for children about 4
girls105 <- subset(girls105, totchi > 0 & totchi < 4)

mod1 <- lm(aauw~white + female + repub +
               age + I(age^2) + srvlng +I(srvlng^2) + demvote +
               factor(totchi) + factor(rgroup) +
               factor(region),
           data = girls105, subset = anygirls == 1)
mod0 <- lm(aauw~white + female + repub +
               age + I(age^2) + srvlng +I(srvlng^2) + demvote +
               factor(totchi) + factor(rgroup) +
               factor(region),
           data = girls105, subset = anygirls == 0)

imp.est <- mean(predict(mod1, girls105) - predict(mod0, girls105))

B <- 1000
boots <- rep(NA, B)
set.seed(123)
for (b in 1:B) {
  overlap <- FALSE
  while (!overlap) {
    girls.star <- dplyr::slice_sample(girls105, n = nrow(girls105),
                                      replace = TRUE)
    # check for overlap
    overlap <- setequal(unique(girls.star[girls.star$anygirls==1, "region"]),
                        unique(girls.star[girls.star$anygirls==0, "region"]))
  }
```

```
  mod1 <- lm(aauw~white + female + repub +
               age + I(age^2) + srvlng +I(srvlng^2) + demvote +
               factor(totchi) + factor(rgroup) +
               factor(region),
             data = girls.star, subset = anygirls == 1)
  mod0 <- lm(aauw~white + female + repub +
               age + I(age^2) + srvlng +I(srvlng^2) + demvote +
               factor(totchi) + factor(rgroup) +
               factor(region),
             data = girls.star, subset = anygirls == 0)

  boots[b] <- mean(predict(mod1, girls.star) - predict(mod0, girls.star))
}

summary(boots)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -6.787   3.723   5.915   5.883   8.001  23.393
```

```
imp.bse <- sd(boots)
imp.bci <- sort(boots)[c(25,975)]

# estimate
imp.est
```

```
## [1] 6.525392
```

```
# boostrap
# standard error
imp.bse
```

```
## [1] 3.388193
```

```
# confidence interval
imp.bci
```

```
## [1] -0.8103791 12.3028904
```

(c)

```
# Regression analysis with similar setup (binary treatment)
mod105_bin <- lm(aauw~anygirls + white + female + repub +
               age + I(age^2) + srvlng +I(srvlng^2) + demvote +
               factor(totchi) + factor(rgroup) +
               factor(region),
             data = girls105)
summary(mod105_bin)
```

```
##
## Call:
## lm(formula = aauw ~ anygirls + white + female + repub + age +
```

```
##     I(age^2) + srvlng + I(srvlng^2) + demvote + factor(totchi) +
##     factor(rgroup) + factor(region), data = girls105)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.950  -9.572  -0.916   9.405  66.242
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      18.932923  36.061979   0.525  0.60010
## anygirls          4.867317   2.775709   1.754  0.08088 .
## white             4.766127   4.633959   1.029  0.30481
## female           10.447154   3.584935   2.914  0.00393 **
## repub           -62.660570   2.785620 -22.494  < 2e-16 ***
## age               0.848723   1.305618   0.650  0.51632
## I(age^2)         -0.007022   0.012577  -0.558  0.57720
## srvlng            0.294198   0.512322   0.574  0.56638
## I(srvlng^2)      -0.014603   0.019208  -0.760  0.44790
## demvote          83.221339  15.226628   5.466 1.23e-07 ***
## factor(totchi)2  -0.557209   3.382807  -0.165  0.86931
## factor(totchi)3  -3.335044   3.683672  -0.905  0.36625
## factor(rgroup)2  -2.966314   2.671111  -1.111  0.26797
## factor(region)2 -13.713045   6.351344  -2.159  0.03191 *
## factor(region)3 -14.959427   6.138064  -2.437  0.01558 *
## factor(region)4  -5.957789   6.859449  -0.869  0.38602
## factor(region)5 -10.699502   6.394141  -1.673  0.09566 .
## factor(region)6 -20.315997   7.592810  -2.676  0.00801 **
## factor(region)7 -16.759290   6.696039  -2.503  0.01303 *
## factor(region)8 -16.658220   7.196197  -2.315  0.02153 *
## factor(region)9 -11.498130   6.525205  -1.762  0.07941 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.06 on 224 degrees of freedom
## Multiple R-squared:  0.8489, Adjusted R-squared:  0.8354
## F-statistic: 62.91 on 20 and 224 DF,  p-value: < 2.2e-16
```

The imputation estimate (6.5253919) is larger than that of the regression estimate (4.8673166). The treatment effect might not be constant across units, which would cause the two estimates to be different.

(d) Answers will depend, but generally `service length` and `demvote` will be post-treatment. When you drop these, the effect of girls gets larger.

(e) Washington was likely trying to control for characteristics of the members and/or districts that might influence their vote on women's issues. That is, we might want to control for the fact that a member comes from a very conservative district or that the member has seniority in the House. With the conservative district, it could be that conservative districts have a preference for conservative candidates on women's issues and also candidates with fewer girls.

Then this preferences drives the correlation, not a direct effect of having girls on member preferences. But it is easy to see that, for a given district, having a member with a girl causes a different vote than having a member with all boys.

References

Washington, E. L. (2008). Female socialization: How daughters affect their legislator fathers' voting on women's issues. *American Economic Review*, 98(1):311–332.

Name	Description
year	Year
congress	Congress number
party	Party 1: dem 2: rep 3:ind
district	District number
statenam	State of MC
name	Name of MC
ngirls	Number of female children
nboys	Number of male children
totchi	Total children
anygirls	Indicator for any female children
propgirls	proportion female children
rgroup	Religious groups 0-none 1-prot 2-cath/orth 3-othchr 4-jewish
statabb	MC State Abbreviation
statalph	State alph codes
region	MC district region
repub	MC a Republican?
srvlng	MC length of service
female	Gender of MC (Female = 1)
white	Race of MC (White = 1, Other = 0)
bday	Birthday of MC
age	Age of MC
demvote	Demoratic share in most recent presidential election
medinc	Median income
perf	Percent female of voting age population
perw	Percent white (total population)
perhs	Percent high school grad rate (age 25p)
percol	Percent college grad rate (age 25p)
alabort	Proportion in state who favor allowing abortion
moreserv	Proportion in state who favor more spending on services
moredef	Proportion in state who favor more spending on defense
morecrimesp	Proportion in state who favor more crime spending
protgay	Proportion in state who favor laws protecting homosexuals
dr1per	Percent Christian (in state) 2001 CUNY
dr2per	Percent Catholic (in state) 2001 CUNY
dr3per	Percent Mormon/Jehovahs (in state) 2001 CUNY
dr4per	Percent other (in state) 2001 CUNY
dr5per	Percent no religion (in state) 2001 CUNY
aauw	AAUW score
rtl	Right to Life Score