DS-UA 201: Midterm Exam

Professor Anton Strezhnev

October 30, 2020

Instructions

You should submit your writeup (as a knitted .pdf along with the accompanying .rmd file) to the course website before 11:59pm EST on Friday October 30th. Please upload your solutions as a .pdf file saved as `Yourlastname_Yourfirstinitial_midterm.pdf`. In addition, an electronic copy of your .Rmd file (saved as `Yourlastname_Yourfirstinitial_midterm.Rmd`) should accompany this submission.

Late finals will not be accepted, **so start early and plan to finish early**. Remember that exams often take longer to finish than you might expect.

This exam has **4** questions and is worth a total of **50 points**. Show your work in order to receive partial credit. Also, I will not accept un-compiled .rmd files.

I general, you will receive points (partial credit is possible) when you demonstrate knowledge about the questions we have asked, you will not receive points when you demonstrate knowledge about questions we have not asked, and you will lose points when you make inaccurate statements (whether or not they relate to the question asked). Be careful, however, that you provide an answer to all parts of each question.

You may use your notes, books, and internet resources to answer the questions below. However, you are to work on the exam by yourself. You are prohibited from corresponding with any human being regarding the exam (unless following the procedures below).

I will answer clarifying questions during the exam. I will not answer statistical or computational questions until after the exam is over. If you have a question, send email to me. If your question is a clarifying one, I will remove all identifying information from the email and reply on Piazza. Do not attempt to ask us questions in person (or by phone), and do not post on Piazza.

Problem 1 (10 points)

For this problem, you will need to install an R packages to draw causal DAGs called `dagitty`. You can install these directly from CRAN using the command `install.packages("dagitty")` Do this *outside* of this .Rmd file (otherwise it will be installed each time you knit the file). The code to generate the DAG below will depend on this package. However, you will not need to draw any graphs yourself, only discuss the one provided.

Consider the following causal directed acyclic graph:

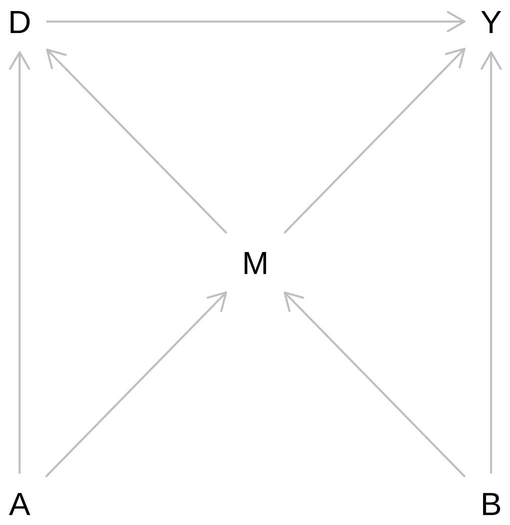

Figure 1: Directed Acyclic Graph 1

Question 1 (3 points)

List all of the paths from D to Y and identify those paths as causal or non-causal.

---

1. *D->Y*: There is a direct effect between D and Y.
2. *D<-M->Y*: This is a mutual dependence through M–a non-causal path. D and Y are not causally related but M is a common cause of both.
3. *D<-A->M->Y*: This is a non-causal path.
4. *D<-M<-B->Y*: This is a non-causal path.
5. *D<-A->M<-B->Y*: This is a backdoor path blocked by a collider.

Question 2 (3 points)

Suppose we conditioned on A and B Under the "back-door" criterion, is this conditioning set sufficient to identify the causal effect of D on Y? Explain why or why not.

No. When A and B are conditioned on, the paths of D<-A->M->Y and D<-M<-B->Y become blocked. However, the path of D<-M->Y is still open so this conditioning set won't be sufficient to identify the causal effect of D on Y.

Question 3 (4 points)

Suppose we conditioned on M only. Under the "back-door" criterion, is this conditioning set sufficient to identify the causal effect of D on Y? Explain why or why not.

When M is the only variable conditioned on, the blocked path of D<-A->M<-B->Y is opened up. So it won't be sufficient to identify the causal effect of D on Y. The conditioning set of A and M, or B and M will be sufficient to identify the causal effect of D on Y.

Problem 2 (15 points)

How do people translate personal experiences into political attitudes? Exploring this question has been frustrated by the non-random assignment of social and economic phenomena such as crime, the economy, education, health care or taxation. In "“Turning personal experience into political attitudes: The effect of local weather on Americans’ perceptions about global warming,” Egan and Mullin (2012) look specifically at the topic of Americans’ beliefs about the evidence for global warming.

They examine whether exposure to abnormally warm temperatures has an effect on whether Americans believe that there is solid evidence that the earth is getting warmer. They use Pew survey data from five months between June 2006 and April 2008.

The variables of interest are:

- `ddt_week` - Average daily departure from normal local temperature (in Fahrenheit) in week prior to survey
- `getwarmord` - Opinion on whether there is “solid evidence” for global warming i.e., the earth getting warmer (no = 1, mixed/some/don’t know = 2, yes = 3).
- `wave` - Month in which survey was conducted (1=June 2006, 2=July 2006, 3=August 2006, 4=January 2007, 5=April 2008).

Below is the code to import the dataset into R

```
### Load in the Egan and Mullin (2013) dataset
gwdataset <- read_dta("gwdataset.dta")
```

Question 1 (5 points)

About 99% of mean daily departure from normal temperatures are between -10 and 20 degrees Fahrenheit. Subset the data down to only those observations (drop the extreme observations with mean daily deviations above 20 or below -10). Use this data for the remainder of this whole problem.

Let’s define our outcome of interest as a binary indicator that takes a value of 1 when a respondent answers that “yes” they believe that there is “solid evidence” that the earth is getting warmer and 0 otherwise.

Let’s define our treatment of interest as exposure to a “heat wave,” defined as a week with an average daily departure from normal local temperature above 10 degrees.

Under the assumption of complete ignorability of treatment, estimate the average treatment effect of exposure to a heat wave on individuals’ belief that there is solid evidence for global warming. Provide an asymptotic 95% confidence interval and interpret your results. Given a rejection threshold of $\alpha = .05$, do we conclude that there is sufficient evidence to reject the null of no average treatment effect?

```
# Subset the data to only observations with mean daily departure from normal temperatures are between $
gwdataset_use <- gwdataset %>% filter(ddt_week >= -10 | ddt_week <= 20)

# create a dummy variable for our outcome of interest
gwdataset_use <- gwdataset_use %>% mutate(warmornot = case_when(getwarmord == 3 ~ 1,
                                                                getwarmord == 2 ~ 0,
                                                                getwarmord == 1 ~ 0,
                                                                TRUE ~ NA_real_))
```

```r
# create a dummy variable for our treatment of interest
gwdataset_use <- gwdataset_use %>% mutate(heatwave = case_when(ddt_week > 10 ~ 1,
                                                               ddt_week <= 10 ~ 0,
                                                               TRUE ~ NA_real_))
# effect of exposure to a heat wave on individuals' belief of global warming
lm_robust(warmornot ~ heatwave, data = gwdataset_use)
```

```
##               Estimate  Std. Error    t value    Pr(>|t|)   CI Lower   CI Upper
## (Intercept) 0.72749692 0.005909899 123.098019 0.000000000 0.71591164 0.73908219
## heatwave    0.04466705 0.014240645   3.136589 0.001716633 0.01675087 0.07258322
##               DF
## (Intercept) 6724
## heatwave    6724
```

On average, people who had been exposed to a heat wave of over 10 degrees are more likely (4.47%) to think that there's strong evidence for global warming. The 95% confidence interval for this estimate is [0.017, 0.073]. Since this confidence level does not include zero, we would reject the null of no average treatment effect at $\alpha = .05$ and conclude that exposure to a heat wave had the effect on individuals' belief that there is solid evidence for global warming.

Question 2 (5 points)

This paper combines data from 5 different Pew surveys from 2006-2008 taken at different times during the year. It may be the case that there is something different across survey waves such that complete ignorability is an unreasonable assumption. Choose an appropriate set of analyses to evaluate whether survey wave confounds treatment and outcome. Interpret your results and discuss whether complete ignorability of our treatment is a reasonable assumption.

Note: Remember that `wave` is a discrete indicator variable for survey month/year - it is *not* a meaningful numeric value. You may want to convert it to a character vector.

```r
# create a dummy variable to convert the discrete indicator variable to characters
gwdataset_use <- gwdataset_use %>% mutate(stratum_month = case_when(wave == 1 ~ "June 2006",
                                                            wave == 2 ~ "July 2006",
                                                            wave == 3 ~ "August 2006",
                                                            wave == 4 ~ "January 2007",
                                                            wave == 5 ~ "April 2008"))

gwdataset_use %>% group_by(stratum_month) %>% summarize(heatwaveMean = mean(heatwave),
                                                        warmornotMean = mean(warmornot),
                                                        N=n(), .groups = "keep")
```

```
## # A tibble: 5 x 4
## # Groups:   stratum_month [5]
##   stratum_month heatwaveMean warmornotMean     N
##   <chr>                <dbl>         <dbl> <int>
## 1 April 2008           0.135         0.695  1395
## 2 August 2006        0.00863         0.761  1390
## 3 January 2007       0.515           0.765  1602
## 4 July 2006          0.00533         0.770   938
## 5 June 2006          0.0136          0.689  1401
```

We find that on average, more people believe there is a strong evidence for global warming in the surveys during August 2006 (0.761), January 2007 (0.765), and July 2006 (0.770). They are about 8 percent more than the avaerage in the other two surveys. This is not consistent with the assumption of complete ignorability.

Question 3 (5 points)

Suppose instead that we assume that excess temperatures are *conditionally* ignorable given survey wave. Using a stratified difference-in-means estimator, stratifying on survey wave, estimate the average treatment effect of exposure to a heat wave on individuals' belief that there is solid evidence for global warming. Provide an asymptotic 95% confidence interval and interpret your results. Compare your findings to your answer from Question 1 and discuss any differences you find.

```r
# first, write the function to calculate difference in means
diff_in_means <- function(treated, control){
  # Point Estimate
  point <- mean(treated) - mean(control)

  # Standard Error
  se <- sqrt(var(treated)/length(treated) + var(control)/length(control))

  # Asymptotic 95% CI
  ci_95 <- c(point - qnorm(.975)*se,
             point + qnorm(.975)*se)

  # P-value
  pval <- 2*pnorm(-abs(point/se))

  # Return as a data frame
  output <- data.frame(meanTreated = mean(treated), meanControl = mean(control), est = point, se = se,

  return(as_tibble(output))

}

# then write a function that does stratified estimation
strat_diff_in_means <- function(outcome, treatment, stratum){

  # For each stratum
  strat_ests <- bind_rows(map(unique(stratum),
                              function(x) diff_in_means(outcome[treatment == 1&stratum == x],
                                                         outcome[treatment==0&stratum == x])))

  # Normalize weights to sum to 1
  strat_ests$weight <- strat_ests$N/sum(strat_ests$N)

  # Point estimate
  point = sum(strat_ests$est*strat_ests$weight)

  # Standard error
  se = sqrt(sum(strat_ests$se^2*strat_ests$weight^2))

  # Asymptotic 95% CI
  ci_95 <- c(point - qnorm(.975)*se,
             point + qnorm(.975)*se)

  # P-value
  pval <- 2*pnorm(-abs(point/se))

  # Return as a data frame
  output <- data.frame(est = point, se = se, ci95Lower = ci_95[1],
                       ci95Upper = ci_95[2], pvalue = pval, N= length(outcome))

  return(as_tibble(output))

}

# diff_in_means(gwdataset_use$warmornot[gwdataset_use$heatwave==1],
#               gwdataset_use$warmornot[gwdataset_use$heatwave==0])

# calculate the stratified difference-in-means estimate
strat_diff_in_means(gwdataset_use$warmornot, gwdataset_use$heatwave,
                    gwdataset_use$stratum_month)
```

```
## # A tibble: 1 x 6
##      est     se ci95Lower ci95Upper  pvalue     N
##    <dbl>  <dbl>     <dbl>     <dbl>   <dbl> <int>
## 1 0.0915 0.0309    0.0309     0.152 0.00310  6726
```

Assuming conditional ignorability given survey wave, we estimate that the average treatment effect is about 9.1 percentage points. The confidence interval is [0.03087, 0.15211]. Since this confidence level does not include zero, we would reject the null of no average treatment effect at $\alpha = .05$. This result is consistent with the result from Question 1–null rejected. However, after stratification and weighting, the effect is even larger, and the Standard Error also increases from 0.014 to 0.031.

Problem 3 (15 points)

Sometimes when designing an experiment, it is impossible to completely randomize over the entire sample of respondents since respondents arrive in a sequence. For example, experimenters fielding online surveys do not observe the entire sample and sometimes have to randomly assign treatments in a "just-in-time" manner. One approach in this case is to simply flip a fair coin for each individual and assign to treatment or control based on whether that coin comes up heads or tails – a sequence of Bernoulli trials. However, this may result in a sample that has too many treated units and too few control units (or vice-versa).

Efron (1971) suggests an alternative approach that biases the coin depending on how many units have previously been assigned to the treatment group versus the control group.

For this problem, you should use the `problem2.csv` dataset. It contains a simulated dataset with an outcome variable `Y` and a number assigned to each unit `order`.

```r
# Load in problem 2 dataset
problem2 <- read_csv("problem2.csv")
```

You should also be familiar with the `rbinom()` function. The function `rbinom(n, 1, prob)` will generate `n` independent random bernoulli trials (binary 0/1 variable) each with success (1) probability of `prob`. For example, `rbinom(20, 1, .3)` will generate 20 independent bernoulli trials each with probability of 0.3 of being equal to 1.

Question 1 (5 points)

Suppose treatment was assigned via independent Bernoulli trials with a constant probability of treatment $\mathbb{P}(D_i = 1) = .5$ for all units. Given the sharp null hypothesis of no individual treatment effect $(Y_i(1) = Y_i(0))$ approximate via simulation (using 10000 iterations) the randomization distribution of the difference-in-means test statistic for this assignment scheme. Based on your simulation, compute the variance of this randomization distribution.

Hint: In this problem we are posing a series of hypothetical assignment mechanisms. While the dataset does not contain an "observed" treatment indicator, it is not necessary for simulating the possible assignments under the null (so unlike the examples we've done with complete randomization, you won't be using `sample()` to generate permutations). You will not be conducting a hypothesis test by comparing the draws to an observed test statistic, only using simulation to understand the properties of either of the two randomization schemes under the null hypothesis.

```r
# Observed test statistic
set.seed(4503) # Set a random seed
nIter <- 10000 # Number of simulation iterations
N <- 1000
reject <- rep(NA, nIter) # placeholder

# function  for difference-in-means asymptotic hypothesis test
dmean_hypo_test <- function(treated, control){
  point <- mean(treated) - mean(control)
  variance <- var(treated)/length(treated) + var(control)/length(control)
  se <- sqrt(variance)
  return(variance)
}
```

```r
# Start the simulation
for (i in 1:nIter){
  # data generating process
  Y_1 <- rnorm(N, 0, 1)
  Y_0 <- rnorm(N, 0, 1)
  # treatmennt
  D <- rbinom(N, 1, .5)
  # outcome
  Y <- Y_1 * D + Y_0 * (1 - D)
}

dmean_hypo_test(Y_1, Y_0)
```

```
## [1] 0.002041595
```

Question 2 (5 points)

Consider instead the randomization scheme where treatment is assigned sequentially for units 1 through 100 according to their `order`. In other words, treatment for unit 1 is randomly assigned. Then treatment for unit 2 is randomly assigned depending on the value of the treatment for unit 1, and so on... Let $\tilde{N_{t,i}}$ denote the number units treated prior to unit i, $\tilde{N_{c,i}}$ the number of units under control prior to unit i and $\tilde{Z}_i = \tilde{N_{t,i}} - \tilde{N_{c,i}}$ or the difference in the number of treated and control groups. By definition, $\tilde{Z}_1 = 0$ since there are no treated or control units when the first unit is assigned.

Define the probability of treatment $\mathbb{P}(D_i = 1)$ for the ith unit as

$$\mathbb{P}(D_i = 1) = \begin{cases} \pi & \text{if } \tilde{Z}_i < 0 \\ 0.5 & \text{if } \tilde{Z}_i = 0 \\ (1 - \pi) & \text{if } \tilde{Z}_i > 0 \end{cases}$$

Intuitively, the assignment mechanism biases the probability of receiving treatment upward if there are fewer treated than control and biases it downward if there are more treated than control at the time of assignment.

Let $\pi = .9$. Given the sharp null hypothesis of no individual treatment effect $(Y_i(1) = Y_i(0))$ simulate the randomization distribution of the difference-in-means test statistic under this new treatment assignment mechanism. Based on your simulation, compute the variance of the randomization distribution. How does it compare to your result from Question 1?

```r
# Bootstrapping
set.seed(10003)
boot.iter <- 100 # number of bootstrap iterations
boot_iptw_mis <- rep(NA, boot.iter) # placeholder for boostrap estimates

boot_pi <- 0.9

#iterations
for (i in 1:nIter){
  # data generating process
  Y_1 <- rnorm(N, 0, 1)
  Y_0 <- rnorm(N, 0, 1)

  # outcome
  Y <- Y_1 * D + Y_0 * (1 - D)

    tideZ = length(Y_1) - length(Y_0)
  # conditional probability
  if (tideZ < 0){
  boot_prob <- boot_pi
  } else if (tideZ == 0){
  boot_prob <- .5
  } else (tideZ > 0)
  boot_prob <- (1 - boot_pi)
}

dmean_hypo_test(Y_1, Y_0)
```

```
## [1] 0.001948196
```

The variance here (0.0019) is smaller than that of question 1 (0.0020).

Question 3 (5 points)

Intuitively, what will happen to this randomization process if π is set to be less than .5? What would happen to the variance of the randomization distribution? (You don't need to use a simulation to answer this, but you are welcome to use one if it would help).

When π is set to be less than .5, the randomization distribution would be off balance. When the number of treated is smaller than the number of control, the resulting probability of getting the assignment of treated is less than .5. This results in more and more control and less treated. The sharp null of no average treatment effect would not be true and the variance of the randomization dirstribution would be large as a result.

Problem 4 (10 points)

In "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," Olken (2007) examines whether increased monitoring had the effect of reducing corruption in Indonesian village road projects. At the time of the study, each of the villages was considering building a road. Olken randomly selected some of the villages to be told, prior to the beginning of construction, that their construction project will be audited by the central government. Olken then measured, for each village's road, the actual amount spent using a combination of engineering surveys, market surveys to determine material price and worker surveys to determine wages paid. He then compared this to the amount each village reported that it spent to measure the extent to which funds were diverted to non-construction purposes (corruption). You will analyze some of the data from this experiment here.

The relevant variables you will need in the dataset are:

- `desaid` - Village identifier
- `kecid` - Sub-district identifier
- `audit` - Treatment: Whether a village was assigned to receive an audit.
- `lndiffeall4mainancil` - Outcome: Percent missing in expenditures. Measured as the difference between the log of the reported amount and the log of the actual amount spent on construction (main road + ancillary projects). Note that this can be negative occasionally when the amount reported happens to be *below* what Olken's estimates suggest was actually spent.

Below is the code to import the dataset into R

```r
### Load in the Olken (2007) data
roads <- read_dta("jperoaddata.dta")
```

Question 1 (3 points)

Olken explains that the randomization of audits could not be done village-by-village since there was concern about spillover effects, where knowledge that a neighboring village would be audited might raise concerns among officials in a village that they would be audited too. Discuss which of the identification assumptions the presence of such spillovers would violate and why?

The assumption of ignorability would be violated. Ignoraibility assumes that the treatment is assigned in a way that is independent of the potential outcomes and that the treatment assignment process is randomized. This is not true in this case since the assignment of a treatment will affect the treatment assignment for the neighboring village.

Question 2 (2 points)

Olken randomizes the treatment at the level of the sub-district. Each sub-district contains multiple villages. He then assumes no interference between sub-districts (noting that communication between villages in different sub-districts is less frequent than communication between villages in the same sub-district). What kind of experimental design is this? What are the consequences of using such a design compared to randomization village-by-village?

This would be a cluster-randomized experiment since results for villages in a sub-district are grouped together. Treatment is randomized at the cluster level and all villages in a sub-district get the same treatment. This would potentially solve the problem of the violation of ignorability in the assignment process. However, it also results in fewer independent observations since there is existing dependence when estimating uncertainty, and that there could be higher sampling error as well.

Question 3 (5 points)

Olken measures the discrepancy between each village's reported expenditures and the actual amount spent on the project, generating a measure of "percent missing" expenditures for each village's project. The paper then goes on to estimate the effect of being assigned to the audit treatment on the amount of missing expenditures in each village's project.

Estimate the average treatment effect of assigning a *sub-district* to treatment on average percent missingness in expenditures within the villages in a *sub-district*. Note that the dataset contains observations at the level of the village. Therefore, you will have to aggregate the data up to the level of the sub-district. You can ignore the fact that the sub-districts were also stratified (since the assignment probabilities of treatment were the same in each stratum). Note also that there is missing data (`NA`) in the outcome. You can drop those villages from the analysis (you should have 538 villages with `lndiffeall4mainancil` not `NA`).

Provide a 95% confidence interval and interpret your results substantively. Do we reject the null that treatment had no average effect on the percent of "missing" expenditures at the $\alpha = .05$ level?

```
# subset the data - drop rows with missing data in lndiffeall4mainancil
roads_use <- roads %>% filter(!is.na(lndiffeall4mainancil))

# aggregate the data up to the level of the sub-district
roads_agg <- roads %>% group_by(kecid) %>% summarize(treatment = first(audit),
                                                     meanDifference = mean(lndiffeall4mainancil),
                                                     size = n(),
                                                     .groups = "keep")

# estimate the average treatment effect
lm_robust(meanDifference ~ treatment, data = roads_agg)
```

```
##               Estimate Std. Error   t value     Pr(>|t|)   CI Lower   CI Upper
## (Intercept)  0.2608363 0.03016607  8.646677 1.706311e-14  0.2011563 0.32051621
## treatment   -0.0558659 0.04376983 -1.276356 2.041050e-01 -0.1424593 0.03072747
##              DF
## (Intercept) 130
## treatment   130
```

On average, villages that were audited are reporting 5.59% less than those that were not audited. The 95% confidence interval for this estimate is [-0.142, 0.0317]. Since this confidence level does include zero, we would not reject the null of no average treatment effect at $\alpha = .05$ and conclude that there is not sufficient evidence to support the claim that increased monitoring had the effect of reducing corruption in Indonesian village road projects.
