# POLC 4325 Assignment #2

Posted: September 30, 2024
Due: October 14, 2024 (11:59pm)

George Johnson

## 1 Potential Outcomes Model (10%)

In the standard setup we discussed, which of the following is NOT true? Explain why.

A. An individual treatment effect $\tau_i = Y_i(1) - Y_i(0)$ is always not identifiable

B. An average treatment effect $E[\tau_i] = E[Y_i(1)] - E[Y_i(0)]$ is always identifiable The ability to identify a treatment effect depends on the assumptions / experimental design.

C. The value of $\tau_i$ depends on unit $i$

D. If all units from a subgroup are not treated, we cannot estimate the average treatment effect for the group

## 2 SUTVA (10%)

The stable-unit-treatment-value assumption (SUTVA) is a critical assumption in causal identification. Which of the following does NOT violate the SUTVA? Explain why.

A. Some control subjects received the treatment

B. Some control and treated units switched their treatment status by chance

C. Some treated subjects contacted their fellow treated units If treated subjects contact other treated subjects there is no violation since they are part of the same group meaning outcomes are not affected.

D. Some treated units contacted their friends in the control group

## 3 Natural Experiments (10%)

Read McCauley and Posner (2017) "The Political Sources of Religion Identification: Evidence from the Burkina Faso-Cˆote d'Ivoire Border" published in the *British Journal of Political Science*.

1. What is **endogenous sorting**? (5%)
This is where individuals/groups sort themselves into different categories based on personal characteristics or choices.

2. How does endogenous sorting threaten the identification strategy taken in the study? Explain while using the word "potential outcomes." (5)

Endogenous sorting threatens the potential outcomes of the study because there is a violation of selection bias, random assignment, and other unobserved cofounders such as prior knowledge or motivations. Due to these problems, there is poor internal validity.

# 4 The Difference-in-Means Estimator (10%)

Read McCauley and Posner (2017) "The Political Sources of Religion Identification: Evidence from the Burkina Faso-Cˆote d'Ivoire Border" published in the *British Journal of Political Science*.

1. Find and download the replication dataset of McCauley and Posner (2017). Do not change the name of the original file. As a proof, **report the DOI** of the dataset. DOI is an URL starting from " https://doi.org/..." (5%)

Data - https://doi.org/10.7910/DVN/S9MKYH,

paper - https://doi.org/10.1017/S0007123416000594

2. Reproduce the first two rows of Table 1 (see below, you can ignore the standard errors). This means that you need to guess what the authors did to produce these reported values. You do not have to recreate the same table format. Please also note that Columns 3-4 in Row 2 cannot be exactly reproduced due to an error in the original paper.

When reproducing the ATE (Column 4), use two approaches:

- The simple difference-in-means estimator (i.e., compute two averages and compute the difference between them).

  When using the difference in means values and adjusting the percent by *100 to get a whole number you get values of 19.89691 for row 1 and 26.48454 for row 2

- The OLS. Compare the result to the one in 1.

  When doing an OLS, you get a value of .19897 and .26485 so you get near the same values when adjusted them to whole numbers of 19.897 for row 1 and 26.485 for row 2.

**Submit your code** and make sure that your code is reproducible. It means that your code should generate the same result when the grader executes your code. If the grader encounters any errors while running your code, you will receive 0 point. (5%)

430 MCCAULEY AND POSNER

TABLE 1 *The Salience of Religious Identity*

| | Full sample (%) | Burkina Faso (%) | Côte d'Ivoire (%) | CI – BF difference |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Lists religion as most important identity | 18.8 | 10.0 | 27.8 | 17.8* (0.04) |
| Lists religion among top two identities | 36.0 | 23.0 | 49.7 | 26.7 (0.14) |
| Willing to marry across religious lines | 66.3 | 75.0 | 57.1 | −17.9 (0.08) |
| Feels closer to co-nationals than co-religionists | 59.7 | 77.3 | 40.3 | −37.0* (0.02) |

# 5 Stratification or Subclassification (35%) All code to get answers in R file

As everyone knows, the Titanic ocean cruiser hit an iceberg and sank on its maiden voyage. Slightly more than 700 passengers and crew survived out of the 2,200 people on board. It was a horrible disaster. One of the things about it that was notable, though, was the role that wealth and norms played in passengers' survival.

Imagine that we wanted to know whether or not being seated in first class made someone more likely to survive. Given that the cruiser contained a variety of levels for seating and that wealth was highly concentrated in the upper decks, it's easy to see why wealth might have a leg up for survival. But the problem was that women and children were explicitly given priority for boarding the scarce lifeboats. If women and children were more likely to be seated in first class, then maybe differences in survival by first class is simply picking up the effect of that social norm.

Using the data set on the Titanic, we calculate a simple difference in mean outcomes (SDO), which finds that being seated in first class raised the probability of survival by 35.4%. But note, since this does not adjust for observable confounders age and gender, it is a biased estimate of the ATE. So next we use subclassification weighting to control for these confounders. Here are the steps that will entail:

1. Assess the covariate balance across the treatment and control groups with respect to the two conditioning variables (age and gender). In this dataset, Age == 0 if young and Age == 1 if old, whereas Sex == 0 if woman and Sex == 1 if man. Discuss the result and what is the main problem with respect to causal identification (5%).

   When assessing balance it can be seen that the first class has the highest mean age suggesting age as a confounder when assessing this group. There is also a large difference in gender when related to crew being a much higher proportion of male suggesting gender as a possible confounder for crew. The main problem with causal identification in the titanic data set is from the imbalance in covariates across the treatment and control groups.

2. Stratify the data into four groups: young males, young females, old males, old females (5%).

3. Calculate the difference in survival probabilities for each group (5%).

```
1 Old.Female     0.744
2 Young.Female   0.622
3 Old.Male       0.203
4 Young.Male     0.453
```

4. Calculate the proportion of each of the four groups. These are our strata-specific weights (5%).

```
1 old.Female   0.193
2 Young.Female 0.0204
3 old.Male     0.757
4 Young.Male   0.0291
```

5. Calculate the weighted average survival rate using the strata weights.

```
> print(survival_weight)
[1] 0.323035
```

6. Compare the above to the naive difference in mean survival rates in the data (5%).

```
1 1 [1st class]          0.625
2 2 [2nd class]          0.414
3 3 [3rd class]          0.252
4 4 [crew]               0.240
```

When comparing the values seen here, we can say that the lower classes had a much lower chance of survival when compared to that of the higher classes. The 1st class can possibly cause a skew in the overall survival rate since they had a high survival rate relative to the classes below it.

7. In order to claim that the effect in 4. is a *causal* effect, what assumption do you have to make? Discuss also when such an assumption may break (5%).

You must assume all people had the same chance of living in their respective class (1st 2nd 3rd or crew etc.) and that the gender and age ratios were equally distributed through the classes and that they had

# 6 Matching (25%)

Continue working on the Titanic dataset.

1. What is the causal estimand that one can identify with matching? (5%)

   Class and its effect on survival rate

2. Estimate the estimand via the Coarsened Exact Matching. Report its estimated value (5%).

```
[1] 0.2574016
```

3. Estimate the estimand via the Nearest-Neighbor Matching. Report its estimated value (5%).

```
> print(ate_nn)
[1] 0.1384615
```

4. Discuss when the matching-based identification strategy (overall) breaks (5%)

   The matching based identification strategy overall breaks when it begins. For example these strategies are and do reduce bias and approximate causal effect, but they rely on assumptions of no unobserved confounding, common support, and correct model specification. These assumptions can break the identification strategy and lead to type 1 or type 2 errors.

5. Discuss the plausibility of the assumption required for matching to recover the causal estimand with the running example (hint: discuss whether there are possible unobserved confounders and what they may be) (5%).

   Unobserved factors in this example would likely be items such as wealth status, health status, race, and location on the ship when it was struck. These confounders are unmeasured and possibilities because of the diversity that was aboard the ship, so while matching may provide insight into the treatment effect the causal identification cannot determined outright do to possible confounders.

Code (just incase R file does not open)

```
#George Johnson

#Advanced Policy Research Methods


#install packages

install.packages("haven")

install.packages('tidyverse')

install.packages('MatchIt')

install.packages('cem')

install.packages("readxl")

install.packages("dplyr")

#load packages

library(tidyverse)

library(haven)

library(MatchIt)

library(cem)

library(readxl)

library(dplyr)


#Question 4


#load data

MPBJPS <- read_xlsx("C:/Users/Tommy_w7c1d3j/Desktop/School/Current Classes/Adv Policy Reasearch Methods/Assignments/Assignment 2/McCauley and Posner, BJPS.xlsx")
```

```
view(MPBJPS)


#percent all

imp_id <- mean(MPBJPS$PrimID == "Religion", na.rm = TRUE) * 100

top_two <- mean(MPBJPS$Relig1st2nd == 1, na.rm = TRUE) * 100


# percent faso

bf_imp_id <- mean(MPBJPS$PrimID[MPBJPS$Country == "Burkina Faso"] == "Religion", na.rm = TRUE)
* 100

bf_top_two <- mean(MPBJPS$Relig1st2nd[MPBJPS$Country == "Burkina Faso"] == 1, na.rm = TRUE) *
100


# percent cote

ci_imp_id <- mean(MPBJPS$PrimID[MPBJPS$Country == "Cote d'Ivoire"] == "Religion", na.rm = TRUE) *
100

ci_top_two <- mean(MPBJPS$Relig1st2nd[MPBJPS$Country == "Cote d'Ivoire"] == 1, na.rm = TRUE) *
100


#fourth column


#difference in means

diff_ci_bf_imp <- ci_imp_id - bf_imp_id

diff_ci_bf_top_two <- ci_top_two - bf_top_two


#ols

MPBJPS$imp_id <- ifelse(MPBJPS$PrimID == "Religion", 1, 0)

MPBJPS$top_two <- ifelse(MPBJPS$Relig1st2nd == 1, 1, 0)

ols_imp <- lm(imp_id ~ Country, data = MPBJPS)

ols_top2 <- lm(top_two ~ Country, data = MPBJPS)

summary(ols_imp)

summary(ols_top2)
```

```
# table 1
summary_table <- data.frame(
  Description = c(
    "Lists religion as most important identity",
    "Lists religion among top two identities"
  ),
  Full_Sample = c(imp_id, top_two),
  Burkina_Faso = c(bf_imp_id, bf_top_two),
  Cote_d_Ivoire = c(ci_imp_id, ci_top_two),
  Difference_CI_BF = c(diff_ci_bf_imp, diff_ci_bf_top_two)
)
print(summary_table)



#Question 5

#load data
dt_titanic <- read_dta("https://github.com/scunning1975/mixtape/raw/master/titanic.dta")
view(dt_titanic)

balance <- dt_titanic %>%
  group_by(class) %>%
  summarize(mean_age = mean(age),
            mean_gender = mean(sex),
            count = n())
print(balance)
```

```
#Groups based on age and sex
dt_titanic <- dt_titanic %>%
  mutate(age_group = ifelse(age > median(age, na.rm = TRUE), "Old", "Young"),
         gender_group = ifelse(sex == 0, "Female", "Male"))

#stratify
dt_titanic <- dt_titanic %>%
  mutate(age_group = ifelse(age == 1, "Old", "Young"),
         gender_group = ifelse(sex == 0, "Female", "Male"))

# cross stratify
dt_titanic$strata <- interaction(dt_titanic$age_group, dt_titanic$gender_group)

# View summary of the stratification
table(dt_titanic$strata)

#Survival
survival <- dt_titanic %>%
  group_by(strata) %>%
  summarize(survival_rate = mean(survived, na.rm = TRUE))
print(survival)

#prop of group
gprop <- dt_titanic %>%
  group_by(strata) %>%
  summarize(prop = n() / nrow(dt_titanic))

print(gprop)
```

```
#combine survival prob and weight
dtamerge <- merge(survival, gprop, by = "strata")
survival_weight <- sum(dtamerge$survival_rate * dtamerge$prop)
print(survival_weight)

#naive diff
naive <- dt_titanic %>%
 group_by(class) %>%
 summarize(survival_rate = mean(survived, na.rm = TRUE))
print(n)

#Question 6

#coarsened matching
cem_match <- cem(treatment = "class", data = dt_titanic, drop = c("survived"))

# View the summary of the CEM result
summary(cem_match)

# Extract the matched data
matched_data <- dt_titanic[cem_match$matched, ]

# Calculate the Average Treatment Effect (ATE)
# Assuming 'survived' is the outcome variable you want to analyze
ate <- mean(matched_data$survived[matched_data$class == 1]) -
 mean(matched_data$survived[matched_data$class == 2])  # Replace '1' and '2' with appropriate treatment/control group codes

# Print the estimated ATE
print(ate)
```

```
#neighbor matching

dt_titanic <- dt_titanic %>%
  mutate(class_bi = ifelse(class == 1, 1, 0))

nn_match <- matchit(class_bi ~ age + sex, data = dt_titanic, method = "nearest", replace = FALSE)
summary(nn_match)

matched_data <- match.data(nn_match)
ate_nn <- mean(matched_data$survived[matched_data$class_bi == 1]) -
  mean(matched_data$survived[matched_data$class_bi == 0])

#ATE Nearest Neighbor
print(ate_nn)
```