Introduction to Causal Inference

# Problem Set 1

Professor: Teppei Yamamoto

**Due Friday, July 15 (at beginning of class)**

Only the required problems are due on the above date. The optional problems will not directly count toward your grade, though you are encouraged to complete them as your time permits. If you choose not to work on the optional problems, use them for your self-study during the summer vacation.

# Required Problems

## Problem 1

Supposed that you have a random sample of size $n$ from the population of interest. Answer the following questions that are designed to help you get familiar with potential outcomes. Try to keep your answers brief and your language precise. Throughout the problem, assume that the Stable Unit Treatment Value Assumption (SUTVA) holds.

(a) Explain the notation $Y_i(0)$.

(b) Contrast the meaning of $Y_i(0)$ with the meaning of $Y_i$.

(c) Contrast the meaning of $Y_i(0)$ with the meaning of $Y_i(1)$. Is it ever possible to observe both at the same time? Why?

(d) Explain the notation $\mathbb{E}[Y_i(0)|D_i = 1]$, where $D_i$ is a binary variable that gives the treatment status for subject $i$, 1 if treated, 0 if control.

(e) Contrast the meaning of $\mathbb{E}[Y_i(0)]$ with the meaning of $\mathbb{E}[Y_i|D_i = 0]$.

(f) Contrast the meaning of $\mathbb{E}[Y_i(0)|D_i = 1]$ with the meaning of $\mathbb{E}[Y_i(0)|D_i = 0]$.

(g) Which of the following quantities (that you explained in parts (d) through (f)) can be identified from observed information? Do not make any additional assumptions about the distributions of $Y$ or $D$, except the SUTVA and also that there is at least one observation with

$D_i = 1$, and at least one with $D_i = 0$ in the observed data.

$$\mathbb{E}[Y_i(0)|D_i = 1] \tag{1}$$
$$\mathbb{E}[Y_i(0)] \tag{2}$$
$$\mathbb{E}[Y_i|D_i = 0] \tag{3}$$
$$\mathbb{E}[Y_i(0)|D_i = 0] \tag{4}$$

(h) Now, assume that $D_i$ is randomly assigned to the units in this sample. Which of the above quantities can be identified from the observed data?

## Problem 2

This problem is for those of you who are unfamiliar with R, which we will use throughout the rest of the course in lectures and problem sets. *If you are already comfortable using R, skip this problem and proceed to Problem 3.*

1. If you haven't, download and install R and RStudio on your own computer. Make sure to choose the right versions for your operating system.

2. Complete Lesson 1 on this website. That is, you should watch each of the video tutorials (Parts 1.1 to 1.6) and complete all the `swirl` quizzes. (You will be able to find out how to install and use `swirl` by following the links on the above website.)

   As proof of your hard work, type the following directly in your console after you complete all the `swirl` quizzes:

   ```
   > savehistory(file = "pset1prob3.txt")
   ```

   and submit the file along with your problem set answers.

## Problem 3

This problem is designed for those of you who are already familiar with R and need a quick review of key concepts and operations. *If you are new to R and just getting started with it, this problem is optional for you.*

1. Using the data given below, answer the following questions in R.

**Major Foreign Holders of U.S. Treasury Securities**

| Country | Dec2007 | Dec2006 | Dec2005 | Dec2004 | Dec2003 |
|---|---|---|---|---|---|
| Japan | 581.2 | 622.9 | 670.0 | 689.9 | 550.8 |
| China, Mainland | 477.6 | 396.9 | 310.0 | 222.9 | 159.0 |
| United Kingdom | 158.1 | 92.6 | 146.0 | 95.8 | 82.2 |
| Oil Exporters | 137.9 | 110.2 | 78.2 | 62.1 | 42.6 |
| Brazil | 129.9 | 52.1 | 28.7 | 15.2 | 11.8 |
| Caribbean Banking Centers | 116.4 | 72.3 | 77.2 | 51.1 | 47.3 |
| All Others | – | – | – | – | – |
| Total | 2353.2 | 2103.1 | 2033.9 | 1849.3 | 1523.1 |

**Total Public Debt Outstanding**

| Year | Debt Held by the Public | Intra-governmental Holdings | Total |
|---|---|---|---|
| Dec2003 | 4044.2 | 2953.7 | 6997.9 |
| Dec2004 | 4408.4 | 3187.8 | 7596.2 |
| Dec2005 | 4714.8 | 3455.6 | 8170.4 |
| Dec2006 | 4901.0 | 3779.2 | 8680.2 |
| Dec2007 | 5136.3 | 4092.9 | 9229.2 |

(a) How much did the debt held by the public increase from 2003 to 2007? What proportion of this new debt was bought up by foreigners? What proportion of the new debt was bought up by the Chinese?

(b) What proportion of U.S. Treasury Securities held by foreigners in 2007 were held by the top six countries/groupings?

(c) As a percentage, what is the nominal increase in holdings from 2003 to 2007 of these six countries/groupings as a group? What if we exclude Japan?

(d) Create a vector of the percentage changes for each country from 2003 to 2007. Be sure to include names indicating which value corresponds to which country.

(e) Repeat the exercises in parts (b) and (c) using the vectors you have just created. Use `sum()` and indices.

(f) What was the ratio of debt held by the public relative to intra-governmental government holdings, for each year 2003-2007? When did this ratio hit it's high point and low point? Make sure to include year labels.

2. (a) Download the Middle East dataset available on the course website and save it to some convenient location. Change your working directory via `setwd()` or using the pull-down menu to the directory in which you saved the data. Read the data using the `read.csv("filename.csv")` command. How many observations (countries) are there? How many variables? [Don't just count off the screen. Imagine we were working with a much larger dataset.]

(b) Create a new variable in the data frame that equals GDP per capita in each country, and calculate the mean GDP per capita in the entire region.

(c) What is the class of the religion variable? If it is not already a character string, coerce the plurality religion for each country to a character string.

(d) You can use "negative indices" to remove rows or columns from objects. For example, `X[-1,]` will give you the matrix `X` less its first row. Using this trick, remove the capital and currency variables from the dataset.

(e) How many of these countries have an HDI value above 0.75? (Hint: The `sum(x)` function counts up the number of `TRUE`s in `x` if `x` is a logical vector.)

(f) How many of these countries have both a density equal to or above 35 and a population above 10 million?

(g) Create a new variable within the Middle East dataset indicating whether each country has a high (above 0.75) or low HDI value.

(h) Create a new dataset consisting of just the high HDI countries.

(i) The `save.image` function allows you to save your workspace into a file so you can retrieve your work later on. Use the following syntax to save your objects in your working directory: `save.image(file = "nameyoulike.RData")`. (You can also do the same via RStudio's drop-down menu.) Now, close RStudio (don't forget to save your code as well!), reopen it, and load the saved workspace using the `load` function.

# Optional Problems

## Problem A

You and some colleagues are conducting an intervention in Ghana's 2016 election[1]. Your goal is to assess the effect of deploying a new biometric voting machine on the incidence of electoral fraud at polling stations. Though Ghana is made up of 275 constituencies, due to the political context you are only allowed to perform your experiment within one constituency. Unconcerned, you and your team randomly select eight polling stations in the constituency, and among the eight, randomly assign half to receive the new voting machines and the other to serve as a control group. $D_i$ gives the resulting treatment status for each polling station $i$, for $i \in \{1, ..., N\}$, where $D_i \in \{0, 1\}$ and $N = 8$. The outcome of interest is the percentage of votes in a polling station attributable to fraud, $Y_i$.

a) Assume that both parts of SUTVA hold. Calculate, explaining your answer:

i) For each polling station $i$, the number of potential outcomes that can be defined;

ii) For each polling station $i$, the number of unit treatment effects that can be defined; and,

[1] This problem is loosely inspired by true events. If you are interested in the substantive or methodological issues, see Asunka et al. 2015.

iii) For the sample of polling stations, the number of (unconditional) average treatment effect estimands that can be defined.

b) A Ghanaian political insider gets wind of your study, and gives you some information. She says that because all of the polling stations in your study are within one small constituency, there will be interference between units. She explains how "interference" might occur in this case: In Ghana, the political operatives in each polling station who are responsible for committing fraud will move elsewhere if their efforts are frustrated. Their range of movement is predetermined by the geographic influence of the party bosses. Local conditions are such that the operatives in each polling station can only move to one, and only one, other polling station, as shown in Figure 1.

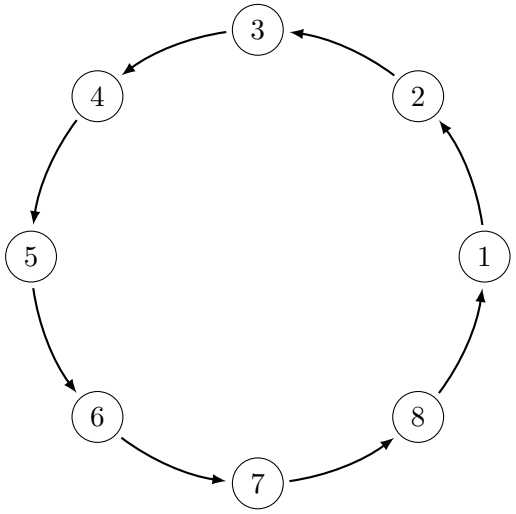


Figure 1: Operatives may move to only one other polling station, as indicated by the arrows (for example, from the second to third polling station, but not the reverse).

Given this structure, answer again questions (i) - (iii).

c) Your funder is keen to spend their budget, and demands that you field the intervention as described originally: $N = 8$, and for $D_i = \{0, 1\}$, $\sum_{i=1}^{N} D_i = 4$. You set aside your concerns, and observe the data in Table 1 on the percentage of votes in each polling station attributable to fraud.

Write out an appropriate estimator for the average treatment effect (ATE) given SUTVA. Drawing on your insights so far and your knowledge of causal inference, under what conditions will this estimator be unbiased? Apply it to the data and compute an estimate of the effect of the new biometric voting machines on the incidence of fraud.

d) Given that you know the structure of the interference network, you believe you may be able to rescue something from the study. Using Figure 1, translate Table 1 into a table formatted like Table 2, where the latter columns give $Y_i(D_i, D_{i-1})$ for the combinations of possible

Table 1: Observed Data: Treatment and Ballot Stuffing

| Unit | $D_i$ | $Y_i$ |
|---|---|---|
| 1 | 1 | 4% |
| 2 | 1 | 2% |
| 3 | 0 | 8% |
| 4 | 1 | 9% |
| 5 | 0 | 12% |
| 6 | 0 | 13% |
| 7 | 0 | 4% |
| 8 | 1 | 1% |

values for $D_i$ and $D_{i-1}$. (And for $i = 1$, $Y_1(D_1, D_8)$, per Figure 1). Where you can, fill the cells of your table with values from Table 1; leave blank the cells representing unobserved potential outcomes.

Table 2: Observed and Unobserved Potential Outcomes

| Unit | $D_i$ | $D_{i-1}$ | $Y_i(1,1)$ | $Y_i(1,0)$ | $Y_i(0,1)$ | $Y_i(0,0)$ |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| ⋮ | | | | | | |
| 8 | | | | | | |

e) Formally express each of the estimands described below using potential outcomes; propose appropriate estimators for each; and finally estimate them with the help of your new table.

   i) The ATE, conditional on a neighbor taking treatment.

   ii) The ATE, conditional on a neighbor taking control.

   iii) The magnitude of effect modification due to assignment of treatment to a neighboring unit.

f) Having heard about your complicated experimental design, a colleague approaches you for advice about a potential field experiment. Their experiment, which randomizes the provision of information about the quality of schooling, will be conducted in a dense urban area. Treatment is to be assigned at the household level. After questioning your colleague extensively, you ascertain that the structure of the urban area can be described by a ring lattice network with the number of households given by $N$, and the interconnectedness of the households given by degree $d$:

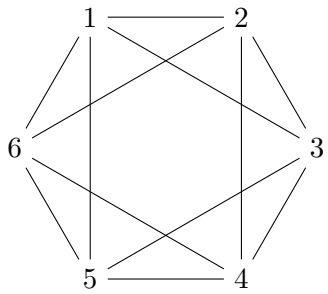


Figure 2: Households and connections between households in the village. This example has $N = 6$, where $N$ represents the "nodes" or units, and $d = 4$, where $d$ represents the degrees of the network, or how many connections each node/unit has to other nodes/units

Figure 2 implies that any household in the village is connected to the four ($d = 4$) nearest households, so that interference can be ruled out between households that are three or more lots apart. We can rule out interference between any households not immediately connected by an edge. Given this setup, and given $N > d$, answer the following:

i) For each household $j$, give the number of potential outcomes that can be defined.

ii) For each household $j$, give the number of unit treatment effects that can be defined.

iii) Imagine now that the experimental intervention is no longer binary, but ternary. It comprises two distinct treatments and a control. In this scenario, how many potential outcomes are there for each $j$? And how many unit treatment effects?

iv) Assume that as N increases the structure of the households continues to follow an expanding ring lattice as shown in figure 2, where the number of connections is still defined by $d$, for any $d < N$. In terms of $k$, the number of levels of treatment (e.g., $k = 2$ if treatment is binary), and $d$, the number of degrees in the ring lattice network, write down a general expression for $M_{k,d}$, the number of unique unit treatment effects for any given unit $j$. An increase in which parameter, $k$ or $d$, has worse implications for the tractability of a study?

## Problem B

This problem will introduce you to directed acyclic graphs (DAGs), and reasoning about them. Consider the following two DAGs, in which $X$, $Y$, and $Z$ are observed, and $U_a$ and $U_b$ are hypothesized but unobserved.

a) List all of the nodes in Graph A.

b) List all of the edges in Graph A.

c) List each path terminating at $Y$ that exists in Graph A, and repeat for Graph B. Interpret the differences between graphs.

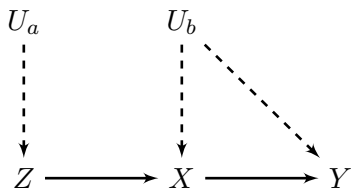


Figure 3: Causal Graph A

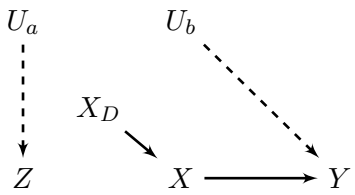


Figure 4: Causal Graph B

d) Explain the relationship between $X_D$ and $X$ in Graph B.

e) Is the relationship between $X$ and $Y$ *identified* in Graph B? Why?

f) Assuming $D_i \in \{0, 1\}$, write down the ATE of $X$ on $Y$ for Causal Graph B using the notation introduced in class.