

Handed out: Wednesday February 3, 2010

Due: Wednesday February 17, 2010

NOTE: Note that there is no school on Monday February 15. I will hold an office hour on Tuesday February 16 but you will likely still want to plan on working on this ahead of that day. **NOTE:** As with last week's assignment there is some computing required. R hints provided but any software is fine.

1. **Observational study on effect of electronic voting:** J. Sekhon (UC Berkeley) examined purported voting irregularities in the 2004 presidential election in Florida. A technical report describing the work is on the course website. The data are also on the course website as either a .Rdata file or a comma delimited data file .csv. You can load the .Rdata file by clicking on it or by entering R and using the load command

```
load("z:\\HAL\\Courses\\Stat265\\fla2004.Rdata")
```

with suitable directory given. You can read the .csv file with the read.csv command

```
a <- read.csv("z:\\HAL\\Courses\\Stat265\\fla2004.csv")
```

The data file contains the following columns:

```
idno - Identification number for county
county - County name
bush04 - Percentage Vote for Bush
etouch - Dummy variable for whether or not the county has electronic voting
income - median income
votePer96.dem - vote percentage for the democratic candidate in '96
votePer96.rep - vote percentage for the republican candidate in '96
votePer00.dem - vote percentage for the democratic candidate in '00
votePer00.rep - vote percentage for the republican candidate in '00
regPer00.dem - percent registered as Democrat
regPer00.rep - percent registered as Republican
turnout00 - turnout in 2000
hisp00 - percentage hispanic in 2000
white00 - percentage white in 2000
black00 - percentage black in 2000
lowEduc00 - percentage with low Education level in 2000
foreignBorn00 - percentage foreign born in 2000
```

We analyze these data to ascertain the effect of electronic voting on the Bush vote.

- (a) Though we have not yet talked about regression analysis as an approach to causal inference, I assume most people are familiar with this idea. We regress the Bush vote (bush04) on the “treatment” indicator (etouch) and covariates to determine the effect of the treatment. As a first step, regress Bush vote on only the treatment indicator. Describe the conclusion you draw.
- (b) Compare the distribution of the covariates in the treatment and control group. Describe your results. Does this impact your faith in the answer from (a)?
- (c) Hopefully “controlling” for the covariates will get a more reliable answer. Choose a set of covariates that you think are likely to impact the outcome and/or the treatment assignment. Regress the Bush vote on the treatment indicator and the selected covariates. How does your answer compare to (a)?
- (d) Repeat (c) for two other choices of the covariate set. The estimated treatment effect varies depending on the covariates included. Discuss.
- (e) The usual regression interpretation of the coefficient of the treatment indicator is the effect of treatment with covariates held fixed. That sounds good. Why is this not an ideal way to estimate the causal effect?

R hints: If the data are in a matrix a, then you can run the regression for part (a) with the command

```
model1 <- lm(a$bush04 ~ a$etouch)
```

or

```
model1 <- lm(a[,3] ~ a[,4])
```

To add other variables to the model formula just change “a\$etouch” to, for example, “a\$etouch + a\$white00 + a\$lowEduc00”. After running a regression then you can get the usual regression summaries with the “summary(model1)” command.

2. **Matching.** We next explore matching. There are only 15 counties with electronic voting and 52 without.

- (a) To start easy, suppose we decide to match on only a single variable, the percent of registered Republicans in 2000. Find the nearest match to each county with electronic voting. List the matches you have found, the distances, and compare the full set of covariates on the matched samples. (I’d prefer matching without replacement; but with limited time to prepare the HW I’ve given you sample code below that does it with replacement. You can use that or improve it.)
- (b) How do the matched samples compare on the covariates compared to the full samples (i.e., compare with 1b)?
- (c) Estimate the average treatment effect on the treated units by comparing the Bush vote in the matched pairs. What is the standard error for the estimate? How does it compare to the regression estimates?
- (d) We might decide that if the best match is not good enough (say the difference is greater than .02) then that pair should be excluded. Repeat the previous part excluding poorly matched pairs. How does this effect the conclusion? How does this effect the interpretation?
- (e) Now try to match on a different distance measure – include at least five covariates. Repeat (a)-(c) for your new distance measure.

R hints: This will require some programming. Here’s a sketch of appropriate logic for the univariate matching on x:

```
index <- c(1:67)
trt <- x[a$etouch == 1]
trt.ix <- index[a$etouch == 1]
ctl <- x[a$etouch == 0]
ctl.ix <- index[a$etouch == 0]
matches <- matrix(0,length(trt),2)
for (i in (1:length(trt))) {
  dis <- abs(ctl - trt[i])
  match.ix <- which(dis == min(dis))
  matches[i,] <- c(trt.ix[i], ctl.ix[match.ix]) }
```

3. There are two articles in which matching is applied posted on the course website with this assignment. One, by Gilligan and Sergenti uses matching to improve causal inference in a study of UN peacekeeping. The other, by Angrist uses matching to study the effect of military service on the labor market. Pick one and read it. Briefly summarize how matching was used and the benefits of matching. Also identify any methodological concerns you may have. (NOTE: The Angrist paper also includes an instrumental variable approach; you can ignore this for now.)