

STATISTICS 265 – Winter 2010 – Homework 3

Handed out: Wednesday February 17, 2010

Due: Wednesday March 3, 2010

NOTE: We use this last homework to examine how propensity scores work using the same observational study of electronic voting that was featured on HW 2.

1. **Observational study on the effect of electronic voting - creating propensity scores:** 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 saving it to your local disk, entering R and using the load command

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

with suitable directory given. The data matrix will be read in with the name "data". You can read the .csv file by saving it to your local disk, entering R and using 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) You can use the correlation command ('cor(x,y)' for vectors x and y or 'cor(x)' for matrix x) to identify variables that seem to be related the treatment (etouch) and/or the outcome (bush04). Which variables are most highly correlated with etouch? Which variables are most highly correlated with bush04?
- (b) Use logistic regression to estimate a propensity score for assignment to the treatment (etouch). This can be done with the 'glm' command. Assume you have read the into a matrix a. Then you can run a logistic regression and save the resulting model with

```
prop1 <- glm(a$etouch ~ a$regPer00.rep, family=binomial)
summary(prop1)
plot(a$etouch, prop1$linear.predictor)
```

where the last two commands provide summary information for the logistic regression and plot the logit of the propensity scores (also known as the linear predictor component of the logistic regression) by group. Your final logistic regression should include important predictors (including the one used in the example code above) and there should be some overlap between the propensity scores of the two groups (control and treatment). Include with your assignment the summary for your final logistic regression and the overlap plot for your final logistic regression.

- (c) There are 13 covariates that can be included in building the propensity score. With careful model selection (including interactions and quadratic terms) it is possible to create a logistic regression that perfectly predicts group membership (i.e., counties with $etouch = 1$ have high propensity scores and counties with $etouch = 0$ have low propensity scores). This would seem like a very effective logistic regression model but is not useful for causal inference. Explain why a result of this type is problematic for a propensity score analysis.

2. Electronic voting - propensity score analysis

- (a) Propensity scores and matching
 - i. Use the propensity score you estimated in the previous problem, or more precisely the linear predictor, to identify matched controls for the 15 counties with electronic voting. (The matching code from HW 2 may be helpful here.)
 - ii. Compare the covariates on the matched samples.
 - iii. Estimate the average treatment effect on the treated units and its standard error.
- (b) Propensity scores and subclassification
 - i. Use the propensity scores for the treated units to define 3 equal-sized blocks (i.e., put 5 treated units in each block). Identify the number of control units in each block. (Reminder: Only use control units whose propensity scores overlap with the treatment group.)
 - ii. Estimate the average treatment effect on the treated units within each of the blocks. Also estimate the standard error.
 - iii. Combine these estimates to produce a single ATT estimate and standard error.
- (c) Summarize your findings. What is the estimated effect of electronic voting on the Bush vote? (Recall that the linear regression models found a negative impact while matching (HW 2) led to a positive estimated effect for electronic voting.)

3. Reading – Dehejia and Wahba (JASA, 1999) illustrates the potential for propensity scores to assist in causal inference. The setting is one in which a randomized experiment was carried out to evaluate the impact of a training program on income (LaLonde 1986). Dehejia and Wahba set out to evaluate the program using observational control groups (while ignoring the real control group). The paper is posted on the course website. Please read the paper (you should be able to skip Section 3).

- (a) Dehejia and Wahba first restrict attention to a subset of the LaLonde data for which there are two years of pre-training income. Explain why this is important. (See Section 2 and especially Section 5.2.)
- (b) Large amounts of the “observational” control data are ignored in the propensity score analysis (see first paragraph of Section 4). Is this a good thing or a bad thing in your opinion? Explain.
- (c) Comment on the effectiveness of propensity scores in this setting.