STATISTICS 265 – Spring 2018 – Homework 2

Handed out: Tuesday May 1, 2018
Due: Tuesday May 15, 2018

1. **Observational study on effect of electronic voting: Regression**

   Many people will remember that there was a great deal of controversy around the 2000 United States presidential election results in Florida (there was a confusing ballot design and a controversial recount). Less well known is controversy that flared up after the 2004 presidential election in Florida. Some people were concerned that there was a difference between results obtained in counties with scanner-read paper ballots and in counties with direct digital (touchscreen) voting. The theory was that the digital machines led to surprisingly high Bush votes. You can get an idea of the controversy associated with this issue by reading this message which was posted prior to the election: http://gnosis.cx/voting-project/October.2004/0212.html. J. Sekhon (UC Berkeley) examined the purported voting irregularities. A technical report describing his 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.

   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 (touchscreen) 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 - precent 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. In this question we use regression to estimate the treatment effect.

   (a) As a first step, regress the Bush vote on the treatment indicator with no other covariates. Describe the conclusion you reach.

   (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 for these data?

   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. **Observational study on effect of electronic voting: Propensity Scores - Design**

   Here we analyze the same data using propensity scores to estimate the causal effect of touchscreen 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 data into a matrix a (as above). 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) Note that the data set includes 13 covariates that can be used in building the propensity score model. Because of the small sample size (only 15 counties have touchscreen voting), it is possible with careful model selection (including interactions and quadratic terms) 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 be a very effective logistic regression model (... it would also be a textbook example of overfitting ...) but would not be terribly useful for causal inference. Explain why.

3. **Observational study on effect of electronic voting: Propensity Scores - Analysis**

   Given the small sample size here we will not apply the algorithm described in class to create blocks. Instead let's just agree a priori to create three equal-sized blocks (in terms of the treated units).

   (a) Use the propensity scores for the treated units that you obtained above to define 3 equal-sized blocks (i.e., put 5 treated units in each block). Identify the number of control units in each block. Assess covariate balance in the blocks.

   (b) Estimate the average treatment effect within each of the blocks. Also estimate the standard error of the block-level treatment effect.

   (c) Combine these estimates to produce a single ATE estimate and standard error.

   (d) This is a setting where it may make more sense to estimate the average treatment effect on the treated units. Explain why.

   (e) Repeat (b) and (c) for the ATT.

   (f) Summarize your findings. Consider the regression results from 1a, 1c, 1d and the propensity score results from 3c, 3e. What conclusion do you draw about the effect of electronic voting on the Bush vote?

4. **Reading** – Dehejia and Wahba (JASA, 1999) is a classic paper illustrating 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). They use a matching approach to estimate the average treatment effect on the treated rather than the subclassification approach that we have focused on.

   (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.