

## Stat 265 - HW 2 Solutions/Comments (Winter 2010)

### 1. **Observational study – regression** – Sample R code is provided separately on the website.

- (a) The estimated treatment effect is  $-.065$  with standard error  $.030$ . This suggests electronic voting leads to negative impact on the Bush vote which is significant at the  $.05$  level ( $p=.0365$ ). I think it is important to report estimates and standard errors, not just p-values.
- (b) Using the standardized distance between means (mean of treatment group - mean of control divided by pooled standard deviation) we find major covariate imbalance. The foreign born percentage and the percent registered as Democrat/Republican have standardized differences greater than 1. Counties with electronic voting are substantially different than other counties on those variables. If these variables are important then it could change the conclusion in (a). Many people just showed graphs – this is helpful but some quantitative measure is a good idea. I prefer the standardized difference to a t-test. Testing is OK but the real question is not whether the means are statistically significantly different (in large samples they always will be); the real question is whether there are large differences.
- (c) I included some of the different voting and registration percentages. This reduced the estimated “treatment” effect to  $-.003$  (std.err =  $.008$ , not significant). In my opinion it is not enough to report that the effect is not significant. Not significant does not mean zero effect; it means based on the current sample size that we can’t reject the hypothesis of zero effect. These are not the same thing.
- (d) I kept the significant predictors from (c) and added the ethnicity variables. This yielded an estimated effect of  $-.008$  (std.err  $.007$ ). Finally I added the low education and foreign born variables which led to an estimate of  $-.006$  (std.err  $.007$ ). Interestingly the treatment effect in all 3 of my multivariate regressions was quite similar (and quite reasonable given the matching result in the next question). (Note: I inadvertently omitted income from all of my regressions; don’t know if it changes things.)
- (e) Regression is not an ideal way to control for covariates because it often (usually!) relies on some kind of extrapolation using the linear function that is being used to approximate the relationship of the mean outcome and the covariates. Thus we must rely on having the model correct. This means the choice of covariates (true for matching also) and the linearity of the model.

### 2. **Observational study - matching.** Again sample code and detailed output are provided separately.

- (a) The matches are reported on the output.
- (b) The covariates look a bit better than they did in (1)(b). Naturally the variable that we matched on is much improved and nearly perfectly balanced.
- (c) The estimated treatment effect on the treated counties is  $-0.071$  which is quite similar to our first (naive) regression. The standard error is  $.037$  which is also similar (slightly larger) to the first regression. It was surprising to see many errors in computing this standard error. The estimate is the mean of the 15 pair differences. The standard error is the s.d. of the 15 pair differences divided by the square root of the sample size!
- (d) My suggestion of using  $.02$  as a threshold was not a good one. If you omit pairs with distance  $> .01$ , then you go from the full 15 treatment cases to 11 treatment cases but the estimated does not change much (now  $-.079$  with std.err  $.043$ ). Restricting to even better matches, omitting pairs with distance  $> .005$ , yields an estimate of  $-.074$  with std.err  $.060$ . Of course omitting some treatment units makes it hard to describe the population on which we are drawing conclusions.

This first part of the problem was to demonstrate matching. I thought this variable would be more helpful. It turns out not to be.

- (e) It is critical to describe what you did to measure distance. Otherwise the reader can’t judge your choice. I used Mahalanobis distance matching using all 13 variables. (See code). The variables are now better balanced than they were earlier. Variable 10 is not balanced as well but most of the others have improved. Unfortunately the standardized differences are still fairly large by traditional standards ( $> .25$ ). The estimated treatment effect is reduced to  $-0.041$  with std.err  $.025$ . Here restricting to good matches does have a bigger impact; using only cases with matches having distance less than 12 leads to an estimated treatment effect of  $-0.004$  (std.err.  $.018$ ) which is similar to the multiple regression estimates.

NOTE: I had seen another analysis of these data using propensity scores in which the propensity analysis estimated a positive effect of the electronic voting on Bush vote which is a major change from the regression. I had thought a matching analysis would produce the same outcome. It did not for me or most of you. Note that both matching and multiple regression to show that the crude estimate (negative impact on Bush vote) is certainly wrong and the actual effect is quite minimal.

### 3. I hope you enjoyed the article you read. Some brief comments.

- (a) Angrist - The author attempts to determine the effect of voluntary military service on income. Clearly comparing veterans to a general collection of non-veterans will have big problems. He is able to avoid that by comparing veterans to others that enlisted and were not selected. This is still problematic because selection to serve is based on characteristics that are likely related to future income. To handle this selection effect he carries out several analyses including a matching analysis. The matching has a big impact and leads to more plausible inferences than the naive analysis. One thing to note is that the instrumental variables analysis (we will talk about this soon) leads to similar results; because it relies on a different population it is hard to compare the results.
- (b) Gilligan and Sergenti - Here the authors attempt to evaluate the effectiveness of UN peacekeeping by matching “crises” one of which the UN intervened and one in which they didn’t. The matching seems to work well and provide a convincing answer for what had been very difficult to answer. Notice that after matching G&S ran a more sophisticated analysis on the matched pairs rather than just take the differences. This is totally fine – the matching just creates comparable treatment and control groups; after that you can do whatever analysis you like! I found it interesting that they chose to match with replacement since this led to the same control being used several times. They don’t say how they would adjust the standard error.