Stat 265 - HW 2 Solutions/Comments (Spring 2018)

1. **Observation study on the effect of electronic voting: regression.**

    (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). 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 the differences are large in magnitude (and using the pooled s.d. as a reference is one way to judge this).

    (c) I included the three covariates with the largest standardized differences (foreignborn00, regper00.dem, regper00.rep) in a regression model. This reduced the estimated "treatment" effect to -.023 (std.err = .027, 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) (the two registration percentages) and added income as well as the two vote percentages from 2000. This yielded an estimated effect of +.004 (std.err .008). The sign is different but the result is still far from significant. Then I added the hispanic and black percentages to the model. This yielded an estimated effect of -.002 (std.err .006). Once again the effect of touchscreen voting is not significant.

    (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 expected outcome and the covariates. Thus we must rely on having the model correct. This means we have to have the correct choice of covariates (true for any causal modeling) and the form of the model must be correct.

2. **Observational study on the effects of electronic voting: propensity scores - design.**

    (a) The correlation of each covariate with the treatment assignment (etouch) and the response (bush04) are listed below. The variables most highly correlated (+ or -) with the Bush vote are previous years democratic and republican vote proportions. The variables most highly correlated with treatment assignment are the proportion of registered democrats and republicans and the percentage of foreign born in the county.

```
              bush04 etouch
bush04          1.00  -0.26
etouch         -0.26   1.00
income         -0.06   0.33
votePer96.dem  -0.90   0.17
votePer96.rep   0.73  -0.04
votePer00.dem  -0.97   0.22
votePer00.rep   0.97  -0.21
regPer00.dem    0.15  -0.45
regPer00.rep   -0.04   0.42
turnout00      -0.18   0.29
hisp00         -0.25   0.20
white00         0.42   0.20
black00        -0.43  -0.19
lowEduc00       0.12  -0.16
foreignBorn00  -0.39   0.40
```

(b) I considered a sequence of three models. I started by including the variable regPer00.rep which has a high correlation with etouch (but not the highest) and reason to believe this might be relevant. Then I tried each other variable in turn to see what would be the most significant addition. This led to the 2nd model including regPer00.rep and votePer96.dem. I then tried each remaining variable in turn to see what would be the most significant addition. This led to a 3rd model including votePer00.rep. This was also a significant predictor. In looking at the plots of the logit propensity scores I found that the 2nd model provided the best overlap. The third provides a better logistic model but seemed to do worse in terms of overlap. Thus I report results for the 2nd model here.

```
glm(formula = fla$etouch ~ fla$regPer00.rep + fla$votePer96.dem, family = binomial)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -20.110      6.609  -3.043  0.00234 **
fla$regPer00.rep    22.456      7.720   2.909  0.00363 **
fla$votePer96.dem   22.709      8.145   2.788  0.00530 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 71.258  on 66  degrees of freedom
Residual deviance: 44.069  on 64  degrees of freedom
AIC: 50.069
```

(c) Note that I did not want you to try and show that you could get perfect separation. Perfect separation is not helpful for causal inference because it is not possible to identify matching units. There is some subtlety here – on the one hand our goal is not to predict treatment assignment but rather to create comparable groups. This would argue for the approach that we used of taking an "inferior" logistic regression to get better balance. On the other hand however, if it is in fact possible to completely identify the treatment assignment based on covariates, then that is strong evidence that we don't have a regular assignment mechanism since things are deterministic given X. If, as seems likely here, the perfect separation is really just statistical overfitting, then we may be OK as we proceeded here.

3. **Observational study on the effect of electronic voting: propensity scores - analysis.**

(a) The question here was not terribly clear. I apologize for that. I created 3 blocks, each consisting of 5 treatment observations. One key issue is that there are many controls with very low propensity scores. I did not include them in my lowest block ... I made the cut point a bit below the lowest propensity score of the treated group. For my analysis there were only two controls in the two higher propensity score blocks and 21 controls in the bottome of the three blocks. Given the small numbers of units it probably makes the most sense just to look at the means of the covariates within blocks to assess balance. Balance achieved here is good but definitely not perfect – given the small number of counties per block it is not too surprising that there are some notable differences.

(b) The table below shows the average propensity score (logit), the average response, and the standard deviation within each block/group. This is followed by an estimate of the treatment effect and standard error (using the Neyman two sample approach) within each block. There is problem if you have only one control observation in a block as you can't estimate the standard deviation in that group and it probably doesn't make sense to assume it is zero.

```
Subclassification
         Treatment                                Control
Block    N   avgprop    avgy       sdy        N    avgprop    avgy       sdy
1 (hi)   5   0.9920483 0.4538984 0.08843032  2   0.9402516 0.5321991 0.004116263
2 (med)  5   0.2384116 0.5898852 0.04303992  2   0.1195223 0.5382275 0.039561287
3 (low)  5 -1.2194370 0.5904136 0.09540234 21 -1.5050534 0.5722421 0.121609413
Block       effect        se
1 (hi)      -0.07830067 0.03965421
2 (med)      0.05165778 0.03395636
3 (low)      0.01817156 0.05024492
```

(c) The combined ATE estimator is obtained by averaging the three block estimates (and standard errors) using the proportions of the total population (after deleting the unmatched controls) which are 7/40, 7/40, 26/40. The combined estimate is .007 and the standard error is .034.

(d) Here there are a set of counties that moved to touchscreen voting. One can argue that the relevant election question for this specific election is what would have happened to the Bush04 vote if these counties had not used touchscreen voting. The ATE estimates the treatment effect if all of Florida counties (except those we deleted!!) had made a change. This is not relevant to the Bush04 vote outcome but could be relevant if we focused on a different outcome like voter turnout. The state might like to know the impact on turnout if all counties implemented touchscreen voting and would use the ATE.

(e) Note that all that is required for the ATT is to combine the same block estimates with new weights based only on the number of treatment units in each block, $(N_{Tj}/N_T) = 5/15$ in my case. Here the combined estimate is -.003 and the standard error is .024.

(f) The regression approach and the propensity score all suggest a non-significant (essentially zero) effect of electronic voting on the Bush vote. I hope that you found it valuable to try out the techniques. Unfortunately the example that we focused on was not terribly interesting from the causal perspective.

4. **Dehejia and Wahba article**

(a) Dehejia and Wahba argue in Section 2 that one should use more than one pre-treatment earnings measurement to get a good evaluation of a treatment of this type. Later sensitivity analyses show that the unconfoundedness assumption is questionable without the second year of pre-treatment earnings.

(b) Of course it is a good thing to ignore the large amounts of control data that don't seem to resemble the treatment group. That's the whole point of our approach. We are interested in those units that look like they might have received treatment.

(c) This is an interesting paper in that it shows how the propensity score approach can work. It also however shows that there are lots of choices to be made and no single definitive "causal estimate". One interesting feature of the article is the use of the control groups (PSID and CPS) to assess the unconfoundedness assumption. Section 5.2 finds that without the 1974 earnings data the two control groups give different effect size estimates which suggests unconfoundedness is not plausible. (This is one of the ideas in Chapter 21 of Imbens and Rubin for assessing unconfoundedness.)