

## Stat 265 - HW 3 Solutions/Comments (Winter 2010)

As usual detailed code and output are provided on the course website. The results there include three different propensity score models; I've only included the best of the three here (good overlap; good explanatory power).

### 1. Creating propensity scores

- (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 on the detailed output. The 2nd model includes regPer00.rep and votePer96.dem. Both are significant predictors and there is good overlap in the sense that all treatment observations appear to be somewhat close to a control observation. Improving the logistic model further made the overlap worse. The overlap plot is in the associated output file but is not reproduced 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) Some people did in fact find that you could get perfect separation for these data. (I did not want you to try and show this.) Everyone observed that such 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 are OK as we proceeded here.

2. **Propensity Score analysis** - Many of you gave very short answers here. In some cases I did not explicitly tell you what to put in the solution but you should know better than give a one line answer. For example, when requested to “compare the covariates on the matched samples”, it is not OK to report that you did this or to list a bunch of numbers without describing what you have done.

- (a) **Matching**

- i. Using the propensity score from question 1 and matching without replacement yields the following matches. The distance is in terms of the logit of the propensity score; this seems to work better than matching on the propensity score. In my opinion the table below should also include the values of the propensity scores and the value of the outcome for each unit. Sorry but I neglected to do that.

Matches

	Trt	ctl	distance
[1,]	43	58	0.1408453356
[2,]	6	26	0.4365010765
[3,]	42	40	0.7490667972
[4,]	50	17	1.0191923706
[5,]	52	5	0.8390635421
[6,]	8	64	0.8106263625
[7,]	30	59	0.7768152516
[8,]	51	27	0.8347253446
[9,]	34	44	0.7352303910
[10,]	11	48	0.7514508490
[11,]	35	9	0.7725819147
[12,]	56	55	0.7810867925
[13,]	60	53	0.0001410482
[14,]	28	41	0.0259993729
[15,]	45	47	0.1373955532

- ii. It is very important to describe what numbers you are looking at. It is generally recommended that you examine the “standardized difference” (equal difference in means divided by pooled standard deviation) for each variable. The t-test (which many of you used) mixes up the size of the imbalance and the sample size. The target is to have covariates within .25 s.d.’s of each other. The table below shows that we fail here on some of the demographic variables (which were not included in my propensity score). I should probably have gone back at this point to include one or more such variable (but I didn’t).

Covariate balance

	Xbar-trt	xbar-ctl	std diff
[1,]	3.928160e+04	3.846020e+04	0.138636075
[2,]	4.558346e-01	4.402390e-01	0.220508284
[3,]	4.392972e-01	4.394142e-01	-0.001582943
[4,]	4.632819e-01	4.546724e-01	0.107714111
[5,]	5.134214e-01	5.191760e-01	-0.072038509
[6,]	3.878838e-01	4.062678e-01	-0.227209723
[7,]	4.347231e-01	4.253083e-01	0.131435643
[8,]	7.133490e-01	7.057115e-01	0.193449388
[9,]	1.219124e-01	8.809752e-02	0.326165554
[10,]	8.683429e-01	8.872598e-01	-0.322646556
[11,]	1.076757e-01	8.681220e-02	0.398053973
[12,]	4.497937e-02	4.051781e-02	0.249549884
[13,]	1.314320e-01	8.437251e-02	0.536932274

- iii. The matching estimator (ATT) of the effect of electronic voting on the Bush 2004 vote for my matching is -.015 with a standard error of .029. Two comments here: first, you should all know better than to report 6-10 decimal places in your final answer; second, everyone had a different answer (of course) but a large positive answer was very difficult to believe. On the subject of reporting results, when the standard error is about .03, which means that you can’t be confident you have the right SECOND digit after the decimal point, then clearly it doesn’t help to report your estimator as -.01490548.

(b) **Subclassification**

- i. I created 3 blocks, each consisting of 5 treatment observations. If your propensity scores are such that there is one or more really outlying propensity score(s), then these should probably not be used. For my analysis there were only two controls in the two higher propensity score blocks and 21 controls in the bottom of the three blocks.
- ii. 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

Block	Treatment				Control			
	N	avgprop	avgy	sdv	N	avgprop	avgy	sdv
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				

iii. The combined ATT is estimator is obtained by averaging the three block estimates (and standard errors). Remember that for ATT the appropriate weight on each block is  $(N_{Tj}/N_T) = 5/15$  in my case. Here the combined estimate is -.003 and the standard error is .024.

(c) The two propensity score approaches both suggest a non-significant (slightly negative) effect of electronic voting on the Bush vote. I hope that you found it valuable to try out the techniques ... I know I did. Unfortunately the example that we focused on was not terribly interesting from the causal perspective in my opinion.

3. **Dehejia and Wahba article** – On the day the assignment was passed out I indicated that this article was not required. This was to allow more time for the project. Apologies to those that went ahead and read it anyway. I’ve provided some answers here for completeness.

(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. A bigger problem is units with large propensity score and no matching unit (as we saw in our FL election example).

(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 20 of Imbens and Rubin for assessing unconfoundedness.)