POL-GA 1251
Quantitative Political Analysis II
Homework 5

1. (15 points) As discussed in MHE (pp. 80-91), CCI (pp. 98-105, 152-155), and in lecture, an alternative to matching for estimating causal effects is to use *inverse-probability weighting* (IPW), where the probabilities that we use are the propensity scores: $e(X_i) = \Pr[D_i = 1|X_i]$. To estimate the ATT via IPW, one constructs the following weights for each i,

$$w_i = \begin{cases} 1 & \text{if } D_i = 1 \\ \frac{e(X_i)}{1-e(X_i)} & \text{if } D_i = 0 \end{cases} \quad \text{or, equivalently, } w_i = D_i + (1 - D_i)\left(\frac{e(X_i)}{1 - e(X_i)}\right),$$

and then one runs a *weighted* regression of the outcome on the treated, specifying the vector of w_i's as the weights. In Stata, this is done using the "`[pweight = ...]`" option, and in R this is done by using the "`weights=...`" option.

In randomized experiments, we know the true values of the $e(X_i)$'s. But in observational studies, treatment is assigned through some natural or social process, and so we have to estimate the $e(X_i)$'s. One way to do so is to fit a logistic regression of D_i on the X_i vectors, and then use the predicted probabilities from that regression as $\widehat{e}(X_i)$ estimates. Then, one uses the $\widehat{e}(X_i)$'s to construct the weights. Because the $\widehat{e}(X_i)$'s are estimates, we may want to propagate our estimation uncertainty about the $\widehat{e}(X_i)$'s into our standard errors for our treatment effects estimates. One could try to work out the math for doing so analytically. But as we've seen, a simpler way to proceed would be to use the bootstrap.

Here is what I'd like you to do:

- (Non-bootstrapped estimate) Use the unmatched Gibson data from last week, and estimate propensity scores for your treatment variable via logistic regression, using the same control variables that you had used for matching. Construct weights as above, run the weighted regression, and report your IPW estimate of the ATT. Discuss any differences between this estimate and your matching estimate (don't worry about the subgroup analysis for this). Obtain the usual robust standard error from this estimate (ignoring the fact that you estimated the propensity scores).

- (Bootstrapped estimate) Now, bootstrap the above. Specifically, write a looping bootstrap function that,

  1. Draws a bootstrap sample (with replacement), with the bootstrap sample size equal to the actual sample size of the complete-case dataset (3,321). This could be done by first sampling the ID1 variable with replacement, and then keeping all units who match the sampled ID1 values. Remember to ensure that if the same ID1 value is sampled multiple times, then you include that unit's data in the bootstrap sample dataset multiple times.
  2. With each bootstrap sample, you estimate the propensity scores as above, construct the weights, then estimate the ATT via weighted regression, and store the ATT estimate from each bootstrap draw.

Have the function produce 1,000 bootstrap estimates of the ATTs. Take the standard deviation of the the bootstrap estimates as your "vanilla" bootstrap estimate of the standard error. How does this standard error compare to the robust standard error from the non-bootstrapped estimate?

2. You are going to consider clustering issues that arise in the paper,

Arceneaux, Kevin. 2007. "I'm Asking for Your Support: The Effects of Personally Delivered Campaign Messages on Voting Decisions and Opinion Formation." *Quarterly Journal of Political Science*. 2(1):43-65.

Answer the following questions:

- Read Arceneaux's description of the experimental protocol on pp. 48-52. What does the protocol suggest about the need to account for clustering in the canvassing treatments? What about for the phone treatment? (5 points)
- Using Arceneaux's replication materials, reproduce the first two columns from Table 3, which estimate the ITT effects of the canvassing treatments and phone treatments on preference for voting for the target candidate. Do this with and without the cluster-robust standard error correction. How do the estimates differ? What does this suggest about the level of within-precinct correlation in the outcomes? (10 points)