# *Problem Set 4*

Research Design for Causal Inference
Due: April 28, 2015

## Part I – Project STAR (redux)

Here we return to the Project STAR data that you started analyzing in Problem Set 2. I recommend revisiting the description of the study and variables presented there as well as the R code you used to perform the analysis before you try to answer the questions below.

### *Question 1 – Estimate treatment effects and conduct hypothesis tests*

**Part a** Estimate the treatment effect using a difference-in-means estimator by conducting a t-test.

```
# First, download and read the dataset into a data frame:
d <- read.csv("http://aaronshaw.org/teaching/2015/causal/data/star.csv")

# Create a treatment variable to make things clearer:
d$treat <- d$class.size == "small"

# For the difference-in-means test recall that the dv is read.score:
t.test(d$read.score[d$treat], d$read.score[!d$treat])

##
##  Welch Two Sample t-test
##
## data:  d$read.score[d$treat] and d$read.score[!d$treat]
## t = 5.6498, df = 3597.11, p-value = 1.731e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.852221 7.946771
## sample estimates:
## mean of x mean of y
##  440.5805  434.6810
```

**Part b** Estimate the treatment effect using a linear regression model (OLS) with no covariates.

```
m1 <- lm(read.score ~ treat, data=d)
summary(m1)

##
## Call:
## lm(formula = read.score ~ treat, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.681  -21.681   -3.681   16.319  192.319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 434.6810     0.7089  613.15  < 2e-16 ***
## treatTRUE     5.8995     1.0405    5.67 1.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.7 on 3731 degrees of freedom
## Multiple R-squared:  0.008543,Adjusted R-squared:  0.008277
## F-statistic: 32.15 on 1 and 3731 DF,  p-value: 1.537e-08
```

**Part c** Esimtate the treatment effect using a linear regression model with any pre-treatment covariates you believe may be relevant.

```
m2 <- lm(read.score ~ treat + free.lunch + race + gender + teach.exper, data=d)
summary(m2)

##
## Call:
## lm(formula = read.score ~ treat + free.lunch + race + gender +
##     teach.exper, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -108.06  -20.63   -3.30   15.31  183.45
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   439.15554    1.59790 274.834  < 2e-16 ***
## treatTRUE       5.88319    0.99225   5.929 3.32e-09 ***
## free.lunchyes -14.61353    1.10491 -13.226  < 2e-16 ***
## raceother      -0.20180    6.85067  -0.029 0.976502
## racewhite       4.69633    1.19314   3.936 8.43e-05 ***
## gendermale     -7.16250    0.99122  -7.226 6.01e-13 ***
## teach.exper     0.32636    0.08731   3.738 0.000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.23 on 3726 degrees of freedom
## Multiple R-squared:  0.09967,Adjusted R-squared:  0.09822
## F-statistic: 68.75 on 6 and 3726 DF,  p-value: < 2.2e-16
```

**Part d** Are your estimates of treatment effects in Parts a, b and c different? If they are different, explain (with words and/or equations) why such differences might occur.

> Overall, there are no substantive differences between the estimates of treatment effects generated by these three estimation procedures. The point estimate of the ATE in the OLS model with no covariates is exactly the same as the difference-in-means estimate. The estimate of the ATE in the full model is a little bit smaller than the reduced model and difference of means estimate (5.88 in the full model vs. 5.90 otherwise). This difference could be due to the fact that additional residual variance in the dependent variable has been "explained" or accounted for by the covariates. This is illustrated most clearly by the (small) reduction in the standard error of the point estimate in the full model.

**Part e** Are there any reasons to prefer any of the estimators (difference-in-means; OLS without covariate adjustment; OLS with covariate adjustment) over the others?

> There may be reasons to prefer either the difference-in-means estimate or the multivariate regression estimate. As Gerber & Green discuss, the difference-in-means estimate is conservative unless the causal effect is additive (all individual causal effects are the same).[1] The multivariate regression estimate may have

[1] Or, unless the random assignment occurs in a random sample drawn from an infinite superpopulation of units (which is a theoretical construct anyway, and thus pretty unlikely). See the Little & Rubin, 2000 text listed in "Additional Readings" on the syllabus for more on this point.

> point estimates that are more precise and have smaller standard errors, but these estimates rely on assumptions that may not be met by the data.
>
> For these reasons, more information about the sampling and treatment assignment mechanisms used in this study as well as additional diagnostics of the residuals would be useful to justify a statistical rationale for a preference for one estimate or the other.

*Question 2 – Interpret the findings*

What do you conclude about the effect of this intervention on the basis of these analyses (you may also refer to the analyses you reported in Problem Set 2)?

> Based on these results, I would conclude that small class size increased average student reading test scores by almost 6 points on average.

*Bonus – Test for covariate imbalance using an F statistic*

Recall how you used Chi-square tests to check covariate balance in Problem Set 2. Now, use the method described by Gerber & Green in *FEDAI*, Section 4.3 to check for covariate imbalance by regressing treatment assignment on the covariates and then creating a sampling distribution of F-statistics for the models estimated with and without covariates.[2]

*R hint:* You will probably want to write a for-loop or a function to estimate regression models for the sampling (randomization) distribution with and without covariates and collect the sum of squared residuals (SSR) for each model. Finding the residuals for a linear model is pretty straightforward. For example, if you estimate and store the results of a linear model like this:

```
my.lm <- lm(y ~ x, data = d)
```

You can access the residuals by calling `my.lm$residuals`.

## Part II – Gerber & Green Chapter 3 exercises

**Excercise 7**

**Exercise 9 (Only parts a-c!)**

[2] Please note that you should always check covariate balance *before* you conduct hypothesis tests, but since the purpose of this Question is to revisit your old tests (and it involves more complicated computation than Question 1), I've put them in the reverse order here.

## PART III – WANTCHEKON REVISITED

For this question you'll use randomization inference to revisit the findings from Leonard Wantchekon's (2003) field experiment on the effect of clientelist vs. public policy framing for campaign messages on voter turnout in Benin.

The dataset is available either a csv or an RData file from the course website.

Recall that the structure of the experiment was *block* randomization. Villages were divided into groups of 3 based on geography and treatment status was randomized within the 8 groups of 3. The outcome variable is the vote share of the candidate participating in the experiment. The only covariate we'll use here is the number of registered voters. In the dataset, `block` indicates block group, `reg.voters` is the registered voters covariate, `vote.pop` is the outcome variable, `treat` is a variable indicating treatment status. The following questions will ask you to analyze the differences between the Clientelist and Public Policy conditions.

**Part a** Estimate the effect the clientelist message compared to the public policy message, using a regression estimator. For the purpose of comparison, do not include the covariate `reg.vote`. Note that the block structure of this experiment will affect how you calculate these quantities. For the regression estimate, you can account for the blocks by including block level dummy variables (R users: include `as.factor(block)`) in your regression equation.[3]

```
# Load the dataset:
d <- read.csv('http://aaronshaw.org/teaching/2015/causal/data/wantchekon.csv')

# Here's the regression model:
ols.model <- lm(vote.pop ~ treat + as.factor(block), data = d)
summary(ols.model)

##
## Call:
## lm(formula = vote.pop ~ treat + as.factor(block), data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.225000 -0.055833 -0.009167  0.052500  0.200833
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

[3] Note that this approach would produce biased estimates if the blocks were different sizes! See *FEDAI* exercise 4.9 for an illustration of this.

```
## (Intercept)         0.70583    0.08004   8.819 4.33e-07 ***
## treatclient         0.10000    0.06200   1.613    0.129
## treatpub.pol       -0.05750    0.06200  -0.927    0.369
## as.factor(block)2  0.13667    0.10124   1.350    0.198
## as.factor(block)3  0.15000    0.10124   1.482    0.161
## as.factor(block)4 -0.17333    0.10124  -1.712    0.109
## as.factor(block)5  0.17000    0.10124   1.679    0.115
## as.factor(block)6  0.02000    0.10124   0.198    0.846
## as.factor(block)7  0.06667    0.10124   0.659    0.521
## as.factor(block)8 -0.06667    0.10124  -0.659    0.521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.124 on 14 degrees of freedom
## Multiple R-squared:  0.6475,Adjusted R-squared:  0.4209
## F-statistic: 2.857 on 9 and 14 DF,  p-value: 0.0383
```

**Part b** Now calculate an estimate of the treatment effect using the difference-in-means as your test statistic. Remember that the block structure of the experiment effects how you calculate this! (see *FEDAI* for details). To help you complete Part c, you'll want to write a function that does this in three steps:

1. Calculate the mean outcome within blocks for each experimental condition. R users: the `tapply()` function will help you do this. For both the Clientelist and Public Policy conditions, you'll want to run `tapply(outcome, block, mean)`.

2. Calculate weights that you will apply to the mean values for each block that you gathered in Step 1. Every weight, $w_b$, for every block, $b$, is equal to the proportion of the sample included in that block. Formally, $w_b = \frac{n_b}{N}$, where $n_b$ is the size of the sample block and $N$ is the total sample size.

3. Calculate the aggregate difference of means (your test statistic) estimate by multiplying the difference of means within each block, $b$ by their respective weights, $w_b$ and then taking the difference of these weighted estimates.

```
# I'll write a function to handle this:

gen.diff.in.means <- function(y, treat, blocks){
    # First, I calculate a vector of means within blocks:
    avg.t <- tapply(y[treat == "client"],
                    blocks[treat == "client"],
                    mean)
    avg.c <- tapply(y[treat == "pub.pol"],
                    blocks[treat == "pub.pol"],
                    mean)

    # Now the weights for each block
    weight <- tapply(y, blocks, length) / length(y)

    # and the weighted difference in means
    test.stat <- sum(weight * (avg.t-avg.c))
    return(test.stat)
}

diff.means.est <- gen.diff.in.means(d$vote.pop, d$treat, d$block)
diff.means.est

## [1] 0.1575
```

**Part c** Use randomization infrerence to test your difference of means estimate from Part b under the sharp null hypothesis of no effects. You can do this using the function described in Part b above and incorporating it into the randomization inference code from the previous problem set. You can also complete this using the "RI" package (I'll try to generate solution code for both). Note that you don't need to enumerate every possible treatment assignment to create the randomization distribution, but can sample a large number of draws (e.g., 5000) instead.

```
# Since we're going to use random numbers, set a seed:
set.seed(20150428)

# Now I'll create a treatment assignment function:
# Note that the 'if/else' statement allows me to handle data with or
# without blocking.

treat.assign <- function(treat,blocks=NA){
  if(length(unique(blocks))==1){
    treat.vector <- sample(treat)
  }
  else{
    # randomize within blocks using tapply
    treat.vector <- tapply(treat,blocks,sample)
  # tapply returns a list, to turn into a vector, use "unlist"
  treat.vector <- unlist(treat.vector)
  }
  return(treat.vector)
}

# Now use replicate to generate a randomization distribution:
rand.dist <- replicate(5000,treat.assign(d$treat, d$block))

# Like last time, make it a data frame:
rand.dist <- data.frame(rand.dist)

# Drop duplicates:
rand.dist <- unique(rand.dist,MARGIN=2)

# Remember, you already have the true test statistic from the part b:
diff.means.est

## [1] 0.1575

# I'll use lapply again (like in pset 3):

ate.dist <- unlist(lapply(rand.dist, gen.diff.in.means, y=d$vote.pop,
blocks=d$block))


# Now, calculate the p-value. Since prior theory suggested that
# clientelist messages would bring out more voters, I'll conduct
# a one-tailed test:

prop.table(table(ate.dist > diff.means.est))

##
## FALSE  TRUE
## 0.991 0.009
```

**Part d** What do you conclude about the effect of clientelist-oriented campaign messages versus those oriented towards public policy on voter turnout in Benin based on these results? How do your findings compare with Wantchekon's published findings?

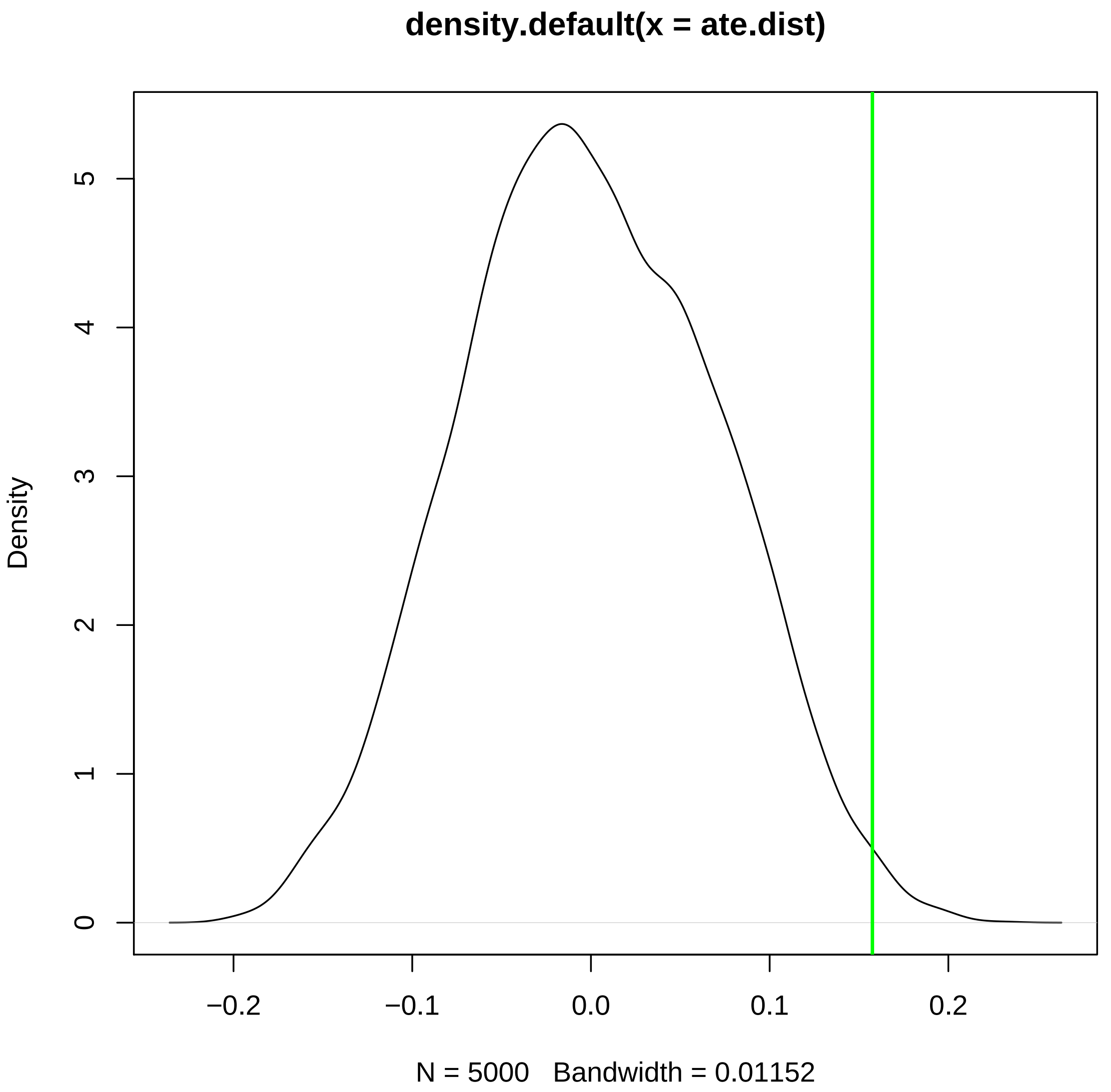
The results indicate that the clientelist message increased voter support by about 16% over the policy-oriented message. The p-value (0.009) suggests we can reject the sharp null hypothesis of no treatment effects.

Using randomization inference to compare directly the impact of clientelist and public policy oriented messages produces more striking results than Wantchekon's reported findings. Whereas randomization inference indicates that clientelism has a significant, positive effect, Wantchekon's regression-based estimators (on p. 414) suggest that he could only reject the null of no effect among Northern villages!

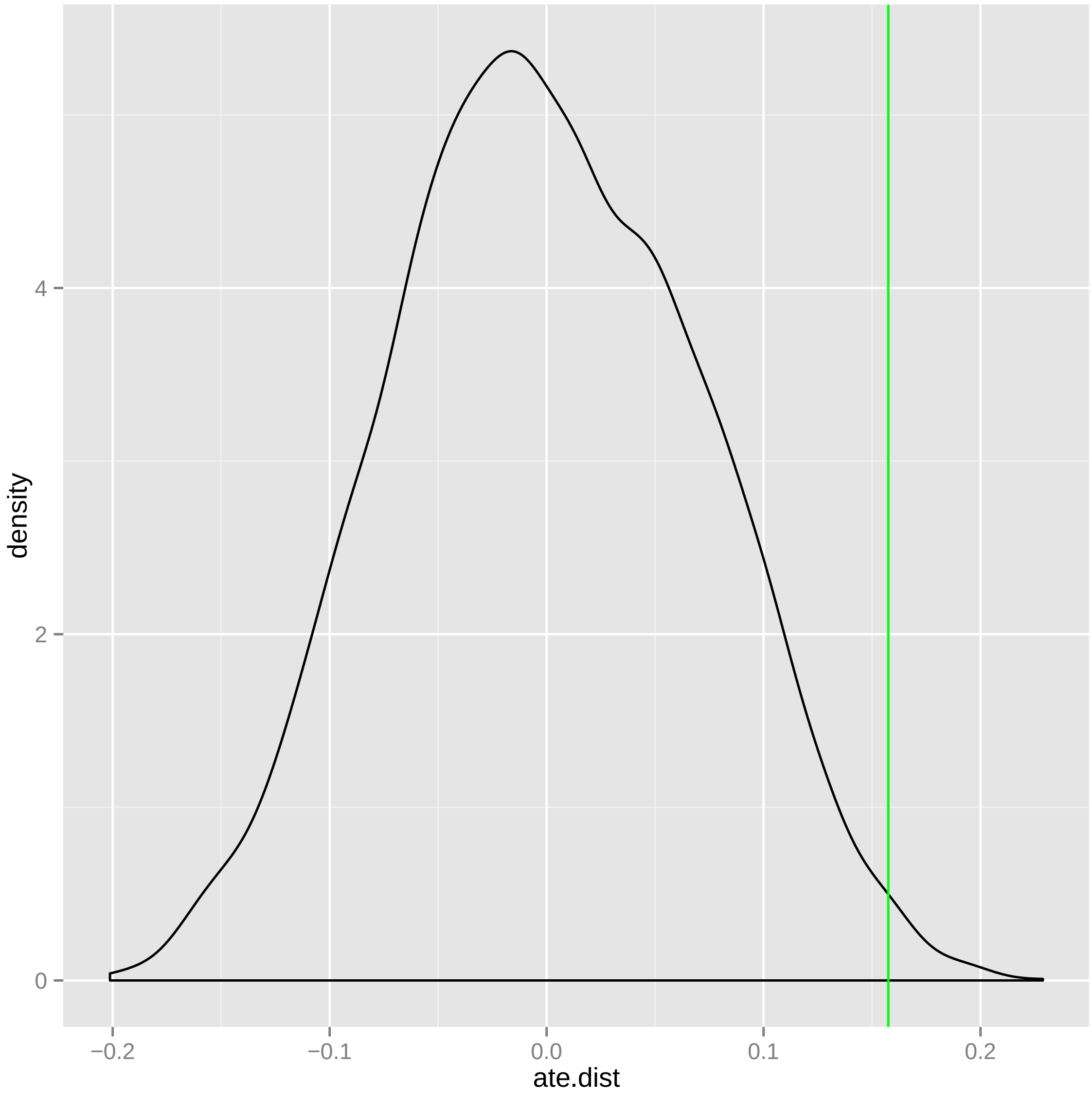
**Bonus** Plot the randomization distribution using a kernel density plot and the true test statistic as an overlaid a vertical line. (R users can overaly the vertical line using something like: `abline(v=test.statistic, col=''red'', lwd=2)`).

```
# First, a version using R's default graphics engine:
plot(density(ate.dist))
abline(v=diff.means.est, col="green", lwd=2)

# Since the code and graphics generated by R's default engine are sort
# of ugly, here's a prettier version using ggplot2
library(ggplot2)
```



```
qplot(ate.dist, geom="density") +  geom_vline(xintercept = diff.means.est,
colour = "green")
```



## Part IV – Key concepts for next class

- The differences between a Gerber & Green's ideas of a research proposal, preanalysis plan (planning document), report, and article (as well as the corresponding checklists)
- Researcher degrees of freedom in a "garden of forking paths."
- Why multiple comparisons might be a problem.
- P-hacking and "fishing" (and why Gelman & Loken regret both terms).
- Internal validity vs. external validity (generalizability).

- Pre-registration of study protocols.
- Replication.