Manzano Romulo - Problem Set 5

December 9, 2017

0.1 Problem Set 5

0.2 Experiments and Causality

Romulo Manzano (20Nov2017)

1 1. Online advertising natural experiment.

These are simulated data (closely, although not entirely) based on a real example, adopted from Randall Lewis' dissertation at MIT.

1.1 Problem Setup

Imagine Yahoo! sells homepage ads to advertisers that are quasi-randomly assigned by whether the user loads the Yahoo! homepage (www.yahoo.com) on an even or odd second of the day. More specifically, the setup is as follows. On any given week, Monday through Sunday, two ad campaigns are running on Yahoo!'s homepage. If a user goes to www.yahoo.com during an even second that week (e.g., Monday at 12:30:58pm), the ads for the advertiser are shown. But if the user goes to www.yahoo.com during an odd second during that week (e.g., Monday at 12:30:59), the ads for other products are shown. (If a user logs onto Yahoo! once on an even second and once on an odd second, they are shown the first of the campaigns the first time and the second of the campaigns the second time. Assignment is not persistent within users.)

This natural experiment allows us to use the users who log onto Yahoo! during odd seconds/the ad impressions from odd seconds as a randomized control group for users who log onto Yahoo! during even seconds/the ad impressions from even seconds. (We will assume throughout the problem there is no effect of viewing advertiser 2's ads, from odd seconds, on purchases for advertiser 1, the product advertised on even seconds.)

Imagine you are an advertiser who has purchased advertising from Yahoo! that is subject to this randomization on two occasions. Here is a link to (fake) data on 500,000 randomly selected users who visited Yahoo!'s homepage during each of your two advertising campaigns, one you conducted for product A in March and one you conducted for product B in August (~250,000 users for each of the two experiments). Each row in the dataset corresponds to a user exposed to one of these campaigns.

```
In [1]: library(data.table)
        library(stargazer)

        d1 <- fread('./data/ps5_no1.csv')
        head(d1)
```

product_b	total_ad_exposures_week1	treatment_ad_exposures_week1	week0	week1	week2	wee
1	4	3	5.5	6.2	0.0	0.0
1	1	1	6.2	0.0	8.6	2.4
1	3	1	0.0	5.3	0.0	8.1
0	5	0	0.0	4.1	0.0	8.8
0	1	1	7.6	3.6	4.6	5.5
1	4	4	6.3	5.5	9.8	5.0

The variables in the dataset are described below:

- **product_b**: an indicator for whether the data is from your campaign for product A (in which case it is set to 0), sold beginning on March 1, or for product B, sold beginning on August 1 (in which case it is set to 1). That is, there are two experiments in this dataset, and this variable tells you which experiment the data belong to.
- **treatment_ad_exposures_week1**: number of ad exposures for the product being advertised during the campaign. (One can also think of this variable as "number of times each user visited Yahoo! homepage on an even second during the week of the campaign.")
- **total_ad_exposures_week1**: number of ad exposures on the Yahoo! homepage each user had during the ad campaign, which is the sum of exposures to the "treatment ads" for the product being advertised (delivered on even seconds) and exposures to the "control ads" for unrelated products (delivered on odd seconds). (One can also think of this variable as "total number of times each user visited the Yahoo! homepage during the week of the campaign.")
- **week0**: For the treatment product, the revenues from each user in the week prior to the launch of the advertising campaign.
- **week1**: For the treatment product, the revenues from each user in the week during the advertising campaign. The ad campaign ends on the last day of week 1.
- **week2-week10**: Revenue from each user for the treatment product sold in the weeks subsequent to the campaign. The ad campaign was not active during this time.

Simplifying assumptions you should make when answering this problem:

- The effect of treatment ad exposures on purchases is linear. That is, the first exposure has the same effect as the second exposure.
- There is no effect of being exposed to the odd-second ads on purchases for the product being advertised on the even second.
- Every Yahoo! user visits the Yahoo! home page at most six times a week.
- You can assume that treatment ad exposures do not cause changes in future ad exposures. That is, assume that getting a treatment ad at 9:00am doesn't cause you to be more (or less) likely to visit the Yahoo home pages on an even second that afternoon, or on subsequent days.

1.2 Questions to Answer

a. Run a crosstab of total_ad_exposures_week1 and treatment_ad_exposures_week1 to sanity check that the distribution of impressions looks as it should. Does it seem reasonable? Why does it look like this? (No computation required here, just a brief verbal response.)

Distribution of Total Adds
Formatting the histogram to reduce the total size

```
In [2]: library('repr')
        options(repr.plot.width=4, repr.plot.height=3)
```

```
In [3]: hist(d1$total_ad_exposures_week1,main = 'Total Ad exposures',xlab = 'Number of total t
```
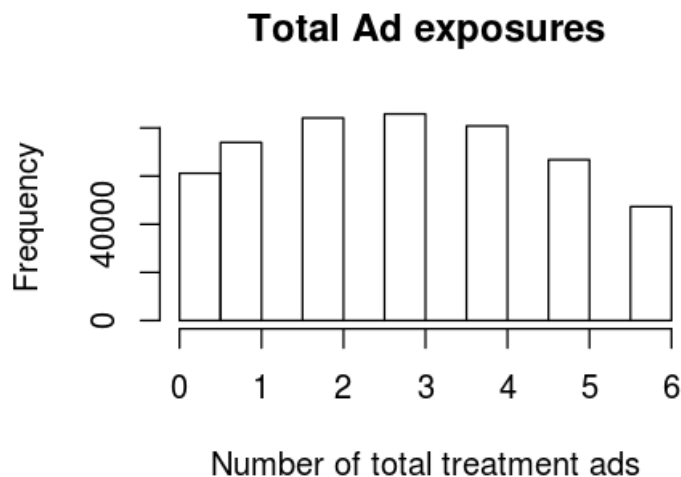

Total Ad exposures

Analysis

The distribution of total ad exposures looks fairly uniform (though not truly so) in the sense that there is wide enough that there were almost as many people exposed to 4 ads, as there were to 2 ads.

This effectively means that the population has a fairly good representation across frequent users of Yahoo, and occassional users from Yahoo, which could help develop arguments around potentially heterogeneous groups (if pre-treatment covariates were availabe around the propensity to visit Yahoo).

One could also argue that the ratio between treatment ads and total ads can effectively be thought of as different treatments (drawing a parallel to medical studies where dosage constitutes a different treatment). However, it seems that based on the assumptions put forth as part of the question we would not be pursuing this perspective.

Distribution of Treatment Ads
Overall ditribution

```
In [4]: hist(d1$treatment_ad_exposures_week1,main = 'Treatment Ad exposures',xlab = 'Number of
```

3

Treatment Ad exposures
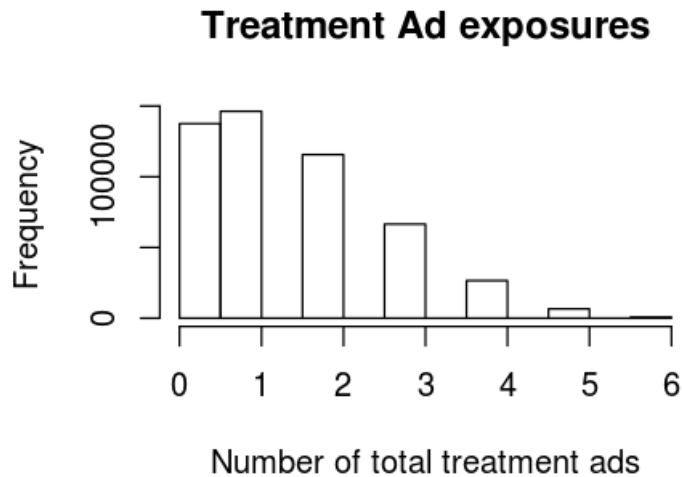
Analysis

This view of the distribution is perhaps not the most informative one, given the random assignment every time a customer visits the Yahoo, we would expect the treatment ad exposure to be normally distributed when controlling for the number of times an individual visited the website.

As such, as better set of plots are shown below, as expected they illustrate the normal distribution across sub populations (based on number of total visits).

Normally distributed when controlling for # of total visits

```
In [5]: hist(d1[d1$total_ad_exposures_week1 == 1]$treatment_ad_exposures_week1,
         main="When exposed to a total of 1 ad(s)",breaks=(c(0,1,2,3,4,5,6,7)-0.5),
         xlab = 'Number of total treatment ads')
```

When exposed to a total of 1 ad(s)
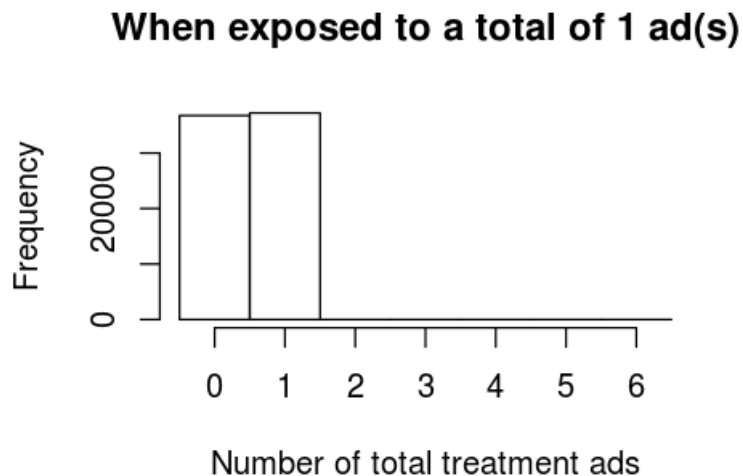
```
In [6]: hist(d1[d1$total_ad_exposures_week1 == 2]$treatment_ad_exposures_week1,
        main="When exposed to a total of 2 ad(s)",breaks=(c(0,1,2,3,4,5,6.7)-0.5),
        xlab = 'Number of total treatment ads')
```

When exposed to a total of 2 ad(s)
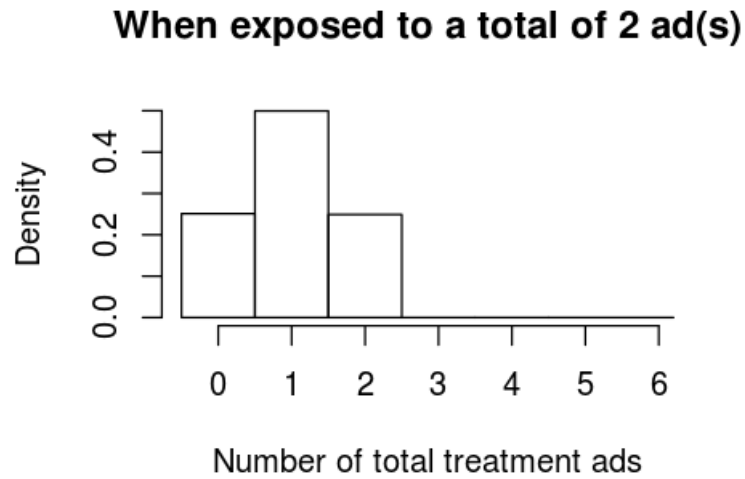

```
In [7]: hist(d1[d1$total_ad_exposures_week1 == 3]$treatment_ad_exposures_week1,
        main="When exposed to a total of 3 ad(s)",breaks=(c(0,1,2,3,4,5,6.7)-0.5),
        xlab = 'Number of total treatment ads')
```
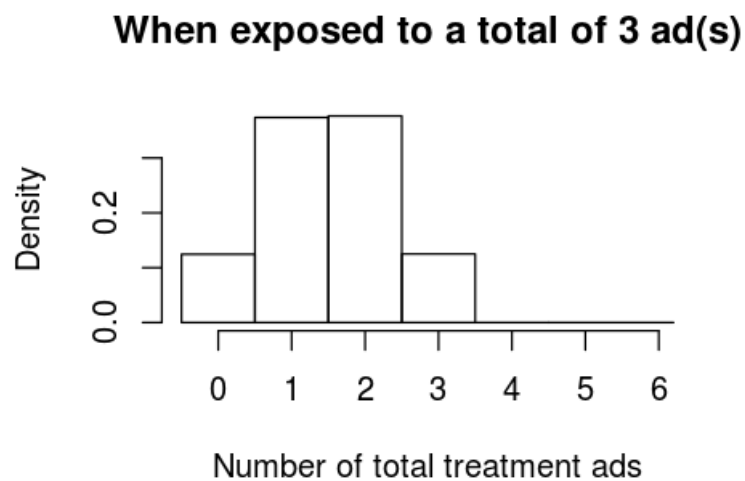
When exposed to a total of 3 ad(s)

```
In [8]: hist(d1[d1$total_ad_exposures_week1 == 4]$treatment_ad_exposures_week1,
        main="When exposed to a total of 4 ad(s)",breaks=(c(0,1,2,3,4,5,6,7)-0.5),
        xlab = 'Number of total treatment ads')
```

When exposed to a total of 4 ad(s)

Number of total treatment ads

```
In [9]: hist(d1[d1$total_ad_exposures_week1 == 5]$treatment_ad_exposures_week1,
        main="When exposed to a total of 5 ad(s)",breaks=(c(0,1,2,3,4,5,6,7)-0.5),
        xlab = 'Number of total treatment ads')
```
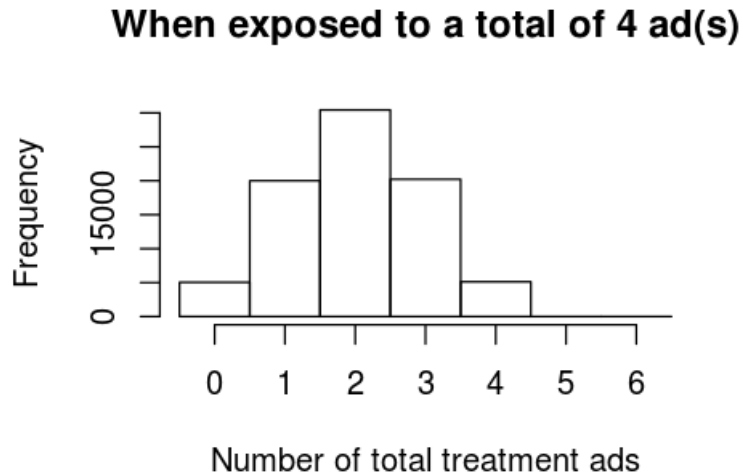
When exposed to a total of 5 ad(s)

Number of total treatment ads

```
In [10]: hist(d1[d1$total_ad_exposures_week1 == 6]$treatment_ad_exposures_week1,
              main="When exposed to a total of 6 ad(s)",breaks=(c(0,1,2,3,4,5,6,7)-0.5),
              xlab = 'Number of total treatment ads')
```
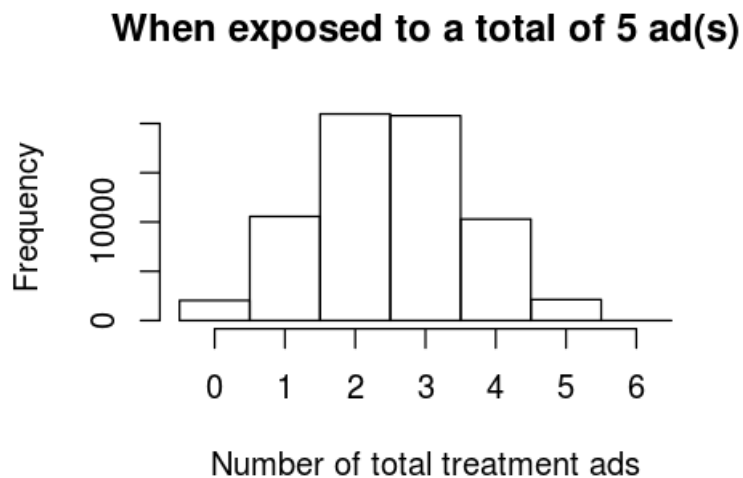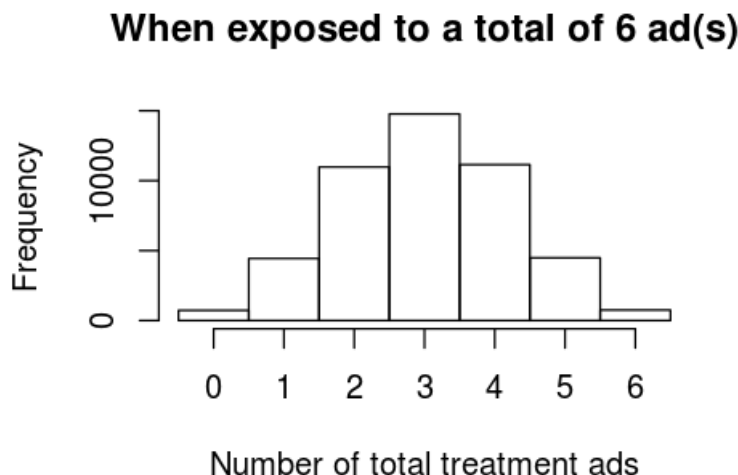
When exposed to a total of 6 ad(s)

Number of total treatment ads

b. Your colleague proposes the code printed below to analyze this experiment: `lm(week1 ~ treatment_ad_exposures_week1, data)` You are suspicious. Run a placebo test with the prior week's purchases as the outcome and report the results. Did the placebo test "succeed" or "fail"? Why do you say so?

Colleague proposal Setting
The regression proposed by my colleague is fundamentally flawed as it doesn't capture the effect of being exposed to the add at all. Given the nature of the data we can expect that the starting point (in terms of expenditure) for customers is not truly zero. Hence, a more sensible approach would be to capture the difference in expenditure (pre-treatment expenditure vs. post-treatment expenditure) to assess any incremental effect in purchases due to the ads.

Furthermore, the proposed framework fails to control for the campaign (a or b) which would be of interest given the data effectively captures two different experiments.
Summarizing the outcome of my colleague's proposal

```
In [11]: summary(lm(week1 ~ treatment_ad_exposures_week1, d1))


Call:
lm(formula = week1 ~ treatment_ad_exposures_week1, data = d1)

Residuals:
   Min      1Q Median     3Q     Max
-3.409 -2.213 -1.615  2.388  8.285
```

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.614684   0.005995  269.34   <2e-16 ***
treatment_ad_exposures_week1 0.299113 0.003138   95.32   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.781 on 499998 degrees of freedom
Multiple R-squared:  0.01785,Adjusted R-squared:  0.01785
F-statistic:  9086 on 1 and 499998 DF,  p-value: < 2.2e-16
```

Analysis

Naively, this suggests that the effect of each impression of the treatment add is an incremental $0.299 in dollar expenditure

Placebo Setting

```
In [12]: summary(lm(week0 ~ treatment_ad_exposures_week1, d1))


Call:
lm(formula = week0 ~ treatment_ad_exposures_week1, data = d1)

Residuals:
   Min    1Q Median    3Q    Max
-3.248 -2.196 -1.670  2.430  8.330

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.669685   0.006027  277.0    <2e-16 ***
treatment_ad_exposures_week1 0.263099 0.003155   83.4    <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.796 on 499998 degrees of freedom
Multiple R-squared:  0.01372,Adjusted R-squared:  0.01372
F-statistic:  6955 on 1 and 499998 DF,  p-value: < 2.2e-16
```

Analysis

This placebo setting fails as it suggests there is still an effect on being exposed to treatment ads int he future ($0.26 incremental expenditure per treatment ad).

This conditions shouldn't exist given customers have no prior knowledge on future treatments in this setting. Also, this suggests that the outcome variable should account for the difference in expenditure pre and post treatment. This approach should be address in the next question.

c. The placebo test suggests that there is something wrong with our experiment or our data analysis. We suggest looking for a problem with the data analysis. Do you see something that might be spoiling the randomness of the treatment variable? How can you improve your analysis to get rid of this problem? Why does the placebo test turn out the way it does? What one thing needs to be done to analyze the data correctly? Please provide a brief explanation of why, not just what needs to be done. (*Note: This question, and verifying that you answered it correctly in part d below, may require some thinking. If we find many people can't figure it out, we will post another hint in a few days.*)

Analysis

One element that the prior setting fails to account is the fact that frequent users of Yahoo.com are more likely to be exposed to the treatment, given there is a 50% change they'll see a treatment add everytime the visit the page.

In order to analyze the data correctly we should take a different approach to (1) assess treatment effect based on a differences-in-differences approach and (2) capture the linear effect of showing more than one add.

d. Implement the procedure you propose from part (c), run the placebo test for the Week 0 data again, and report the results. (This placebo test should pass; if it does not, re-evaluate your strategy before wasting time proceeding.)

Implementation

We now add a dummy variable to capture whether the customer received the treatment or not, and we calculate the differences between pre-treatment expenditure and week_1 (differences in differences) and week_0 (to test placebo)

```
In [13]: d1$treatment_ad_exposure_received <- ((d1$treatment_ad_exposures_week1 >=1)*1)
         d1$week1_did <- (d1$week1 - d1$week0)
         d1$week0_did <- (d1$week0 - d1$week0)
         head(d1)
```

product_b	total_ad_exposures_week1	treatment_ad_exposures_week1	week0	week1	week2	wee
1	4	3	5.5	6.2	0.0	0.0
1	1	1	6.2	0.0	8.6	2.4
1	3	1	0.0	5.3	0.0	8.1
0	5	0	0.0	4.1	0.0	8.8
0	1	1	7.6	3.6	4.6	5.5
1	4	4	6.3	5.5	9.8	5.0

Testing the placebo effect

```
In [14]: summary(lm(week0_did ~ treatment_ad_exposures_week1, d1))


Call:
lm(formula = week0_did ~ treatment_ad_exposures_week1, data = d1)

Residuals:
   Min     1Q Median     3Q    Max
     0      0      0      0      0
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                          0          0      NA       NA
treatment_ad_exposures_week1         0          0      NA       NA

Residual standard error: 0 on 499998 degrees of freedom
Multiple R-squared:    NaN,Adjusted R-squared:    NaN
F-statistic:   NaN on 1 and 499998 DF,  p-value: NA
```

As observed, the placebo test is successful as there is no observable difference in the outcome due to the future treatment assignment.

   e. Now estimate the causal effect of each ad exposure on purchases during the week of the campaign itself using the same technique that passed the placebo test in part (d).

Implementation for product a

```
In [15]: summary(lm(week1_did ~ treatment_ad_exposures_week1, d1[d1$product_b ==0]))


Call:
lm(formula = week1_did ~ treatment_ad_exposures_week1, data = d1[d1$product_b ==
    0])

Residuals:
    Min      1Q  Median      3Q     Max
-9.5212 -1.9327  0.0115  1.9231  9.4904

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -0.055720   0.009443  -5.901 3.63e-09 ***
treatment_ad_exposures_week1   0.044219   0.005645   7.834 4.76e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.606 on 300036 degrees of freedom
Multiple R-squared:  0.0002045,Adjusted R-squared:  0.0002012
F-statistic: 61.37 on 1 and 300036 DF,  p-value: 4.756e-15
```

Analysis
Based on these results we can observe that exposure to each treatment ad resulted in a \$0.044 increase in expenditure for product a
Implementation for product b

```
In [16]: summary(lm(week1_did ~ treatment_ad_exposures_week1, d1[d1$product_b ==1]))


Call:
lm(formula = week1_did ~ treatment_ad_exposures_week1, data = d1[d1$product_b ==
    1])

Residuals:
    Min      1Q   Median      3Q      Max
-10.1261  -2.4714  -0.0024   2.5286  10.9358

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -0.059538   0.016585  -3.590 0.000331 ***
treatment_ad_exposures_week1   0.030946   0.007472   4.142 3.45e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.312 on 199960 degrees of freedom
Multiple R-squared:  8.578e-05,Adjusted R-squared:  8.078e-05
F-statistic: 17.15 on 1 and 199960 DF,  p-value: 3.449e-05
```

Analysis

Based on these results we can observe that exposure to each treatment ad resulted in a $0.03 increase in expenditure for product b

f. The colleague who proposed the specification in part (b) challenges your results -- they make the campaign look less successful. Write a paragraph that a layperson would understand about why your estimation strategy is superior and his/hers is biased.

Analysis

The estimation strategy proposed in step(c) is superior to my colleague's given the later fails to account for the fact that the treatment was applied to customers who already showed expending behaviour for the products, and that failing to use the existing expenditure behaviour as a baseline results in overestimating the effect of the treatment.

My proposal effectively controls for this factor effectively shedding light into the true effect of being exposed to the treatment ads on incremental customer expenditure.

g. Estimate the causal effect of each treatment ad exposure on purchases during and after the campaign, up until week 10 (so, total purchases during weeks 1 through 10).

Implementation

We first need to capture the difference-in-differences metrics or each one of the 10 weeks in order to capture the incremental weekly expenditure. Then we need to aggregate the incremental expenditure across weeks - which will then be used as the target variable in our regression

```
In [17]: d1$week2_did <- (d1$week2 - d1$week0)
         d1$week3_did <- (d1$week3 - d1$week0)
         d1$week4_did <- (d1$week4 - d1$week0)
         d1$week5_did <- (d1$week5 - d1$week0)
         d1$week6_did <- (d1$week6 - d1$week0)
         d1$week7_did <- (d1$week7 - d1$week0)
         d1$week8_did <- (d1$week8 - d1$week0)
         d1$week9_did <- (d1$week9 - d1$week0)
         d1$week10_did <- (d1$week10 - d1$week0)
         d1$incremental_exp_all_weeks <- d1$week1_did+d1$week2+d1$week3+d1$week4_did+d1$week5_
         head(d1)
```

product_b	total_ad_exposures_week1	treatment_ad_exposures_week1	week0	week1	week2	wee
1	4	3	5.5	6.2	0.0	0.0
1	1	1	6.2	0.0	8.6	2.4
1	3	1	0.0	5.3	0.0	8.1
0	5	0	0.0	4.1	0.0	8.8
0	1	1	7.6	3.6	4.6	5.5
1	4	4	6.3	5.5	9.8	5.0

Implementation for product a

```
In [18]: summary(lm(incremental_exp_all_weeks ~ treatment_ad_exposures_week1, d1[d1$product_b =
```

```
Call:
lm(formula = incremental_exp_all_weeks ~ treatment_ad_exposures_week1,
    data = d1[d1$product_b == 0])

Residuals:
   Min     1Q Median     3Q    Max
-78.48 -16.93   5.62  16.27  75.37

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.57974    0.05910 111.329   <2e-16 ***
treatment_ad_exposures_week1  0.35053    0.03533   9.922   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22.57 on 300036 degrees of freedom
Multiple R-squared:  0.000328,Adjusted R-squared:  0.0003247
F-statistic: 98.45 on 1 and 300036 DF,  p-value: < 2.2e-16
```

Analysis

Based on these results we can observe that exposure to each treatment ad resulted in a \$0.35 increase in total expenditure for product a in all weeks during which or after the treatment was applied

Implementation for product b

```
In [19]: summary(lm(incremental_exp_all_weeks ~ treatment_ad_exposures_week1, d1[d1$product_b =

Call:
lm(formula = incremental_exp_all_weeks ~ treatment_ad_exposures_week1,
    data = d1[d1$product_b == 1])

Residuals:
    Min      1Q  Median      3Q     Max
-87.959 -22.034   7.466  20.078  82.466

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    7.10937    0.10294   69.06  < 2e-16 ***
treatment_ad_exposures_week1   0.21241    0.04638    4.58 4.65e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 26.77 on 199960 degrees of freedom
Multiple R-squared:  0.0001049,Adjusted R-squared:  9.99e-05
F-statistic: 20.98 on 1 and 199960 DF,  p-value: 4.649e-06
```

Analysis

Based on these results we can observe that exposure to each treatment ad resulted in a \$0.21 increase in total expenditure for product b in all weeks during which or after the treatment was applied

h. Estimate the causal effect of each treatment ad exposure on purchases only after the campaign. That is, look at total purchases only during week 2 through week 10, inclusive.

Implementation

As a first step, we'd need to calculate the incremental expenditure from weeks 2 to 10

```
In [20]: d1$incremental_exp_weeks_2_to_10 <- d1$week2+d1$week3+d1$week4_did+d1$week5_did+d1$wee
```

Implementation for product a

```
In [21]: summary(lm(incremental_exp_weeks_2_to_10 ~ treatment_ad_exposures_week1, d1[d1$produc

Call:
lm(formula = incremental_exp_weeks_2_to_10 ~ treatment_ad_exposures_week1,
    data = d1[d1$product_b == 0])

Residuals:
    Min      1Q  Median      3Q     Max
-71.254 -14.654   4.365  14.365  71.158
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.63546    0.05269 125.922   <2e-16 ***
treatment_ad_exposures_week1   0.30631    0.03150   9.725   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 20.12 on 300036 degrees of freedom
Multiple R-squared:  0.0003151,Adjusted R-squared:  0.0003118
F-statistic: 94.57 on 1 and 300036 DF,  p-value: < 2.2e-16
```

Analysis

Based on these results we can observe that exposure to each treatment ad resulted in a \$0.31 increase in total expenditure for product a in the 9 weeks following the treatment alone

Implementation for product b

```
In [22]: summary(lm(incremental_exp_weeks_2_to_10 ~ treatment_ad_exposures_week1, d1[d1$product
```

```
Call:
lm(formula = incremental_exp_weeks_2_to_10 ~ treatment_ad_exposures_week1,
    data = d1[d1$product_b == 1])

Residuals:
   Min     1Q Median     3Q    Max
-77.89 -19.13   5.85  17.73  73.17

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    7.16890    0.09159  78.268  < 2e-16 ***
treatment_ad_exposures_week1   0.18147    0.04127   4.398  1.1e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 23.82 on 199960 degrees of freedom
Multiple R-squared:  9.67e-05,Adjusted R-squared:  9.17e-05
F-statistic: 19.34 on 1 and 199960 DF,  p-value: 1.096e-05
```

Analysis

Based on these results we can observe that exposure to each treatment ad resulted in a \$0.18 increase in total expenditure for product b in the 9 weeks following the treatment alone

i. Tell a story that could plausibly explain the result from part (h).

Analysis

There are two potential perspectives that we can think:

1. The treatment (i.e. being exposed to the treatment ad) does have lasting effect (past one week) on the consumer's expenditure.

2. The treatment has no lasting effect on expenditure, but prior expenditure (as measure by dollar amount) influences how much a customer will spend the following week. In other words, this claims that the customer might spend more money in weeks 2-10 just because the higher expenditure in week 1 serves as a higher anchor, rather than the customer remembering the ad and spending more as a result.

j. Test the hypothesis that the ads for product B are more effective, in terms of producing additional revenue in week 1 only, than are the ads for product A. (*Hint: The easiest way to do this is to throw all of the observations into one big regression and specify that regression in such a way that it tests this hypothesis.*) (*Hint 2: There are a couple defensible ways to answer this question that lead to different answers. Don't stress if you think you have an approach you can defend.*)

Implementation

```
In [23]: summary(lm(week1_did ~ treatment_ad_exposures_week1 + product_b, d1))


Call:
lm(formula = week1_did ~ treatment_ad_exposures_week1 + product_b,
    data = d1)

Residuals:
     Min        1Q    Median        3Q       Max
-10.1568   -2.0897    0.0103    2.1338   10.9197

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.048565   0.008965  -5.417 6.06e-08 ***
treatment_ad_exposures_week1    0.038253   0.004534   8.436  < 2e-16 ***
product_b                      -0.024169   0.011600  -2.084   0.0372 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.904 on 499997 degrees of freedom
Multiple R-squared:  0.0001423,Adjusted R-squared:  0.0001383
F-statistic: 35.59 on 2 and 499997 DF,  p-value: 3.511e-16
```

Analysis

The data suggest that ads for product B are actually less effective than product A, however this hypothesis is not supported at a statistically significant level.

k. You notice that the ads for product A included celebrity endorsements. How confident would you be in concluding that celebrity endorsements increase the effectiveness of advertising at stimulating immediate purchases?

Analysis

I would not be confident in concluding that. In order to make that assessment we would need to run an experimentw with multiple treatments, where ads would be shown for product A with and without celebrity endorsements (everything else being equal) plus a control group.

It is easy to imagine why the current data does not provide the necessary foundations to make such a claim. One element that challenges the validity of the claim is the fact that experiments for products A and B were conducted in different seasons, and we can not disregard that the differences observed are due to some external factors having seasonality effects in either product.

2   2. Vietnam Draft Lottery

A famous paper by Angrist exploits the randomized lottery for the Vietnam draft to estimate the effect of education on wages. (*Don't worry about reading this article, it is just provided to satisfy your curiosity; you can answer the question below without referring to it. In fact, it may be easier for you not to, since he has some complications to deal with that the simple data we're giving you do not.*)

2.1   Problem Setup

Angrist's idea is this: During the Vietnam era, draft numbers were determined randomly by birth date -- the army would literally randomly draw birthdays out of a hat, and those whose birthdays came up sooner were higher up on the list to be drafted first. For example, all young American men born on May 2 of a given year might have draft number 1 and be the first to be called up for service, followed by November 13 who would get draft number 2 and be second, etc. The higher-ranked (closer to 1) your draft number, the likelier it was you would be drafted.

We have generated a fake version of this data for your use in this project. You can find real information (here)[https://www.sss.gov/About/History-And-Records/lotter1]. While we're defining having a high draft number as falling at 80, in reality in 1970 any number lower than 195 would have been a "high" draft number, in 1971 anything lower than 125 would have been "high".

High draft rank induced many Americans to go to college, because being a college student was an excuse to avoid the draft -- so those with higher-ranked draft numbers attempted to enroll in college for fear of being drafted, whereas those with lower-ranked draft numbers felt less pressure to enroll in college just to avoid the draft (some still attended college regardless, of course). Draft numbers therefore cause a natural experiment in education, as we now have two randomly assigned groups, with one group having higher mean levels of education, those with higher draft numbers, than another, those with lower draft numbers. (In the language of econometricians, we say the draft number is "an instrument for education," or that draft number is an "instrumental variable.")

Some simplifying assumptions:

- Suppose that these data are a true random sample of IRS records and that these records measure every living American's income without error.
- Assume that the true effect of education on income is linear in the number of years of education obtained.

16

- Assume all the data points are from Americans born in a single year and we do not need to worry about cohort effects of any kind.

2.2 Questions to Answer

a. Suppose that you had not run an experiment. Estimate the "effect" of each year of education on income as an observational researcher might, by just running a regression of years of education on income (in R-ish, `income ~ years_education`). What does this naive regression suggest?

```
In [24]: d2 <- fread("./data/ps5_no2.csv")
         head(d2)
```

draft_number	years_education	income
267	16	44573.90
357	13	10611.75
351	19	165467.80
205	16	71278.40
42	19	54445.09
240	11	32059.12

```
In [37]: d2 %>% group_by(draft_number)
```

```
       Error in d2 %>% group_by(draft_number): could not find function "%>%"
    Traceback:
```

Implementation

```
In [25]: library("lmtest")
         coeftest(lm(income~years_education,d2))
```

```
Loading required package: zoo

Attaching package: zoo

The following objects are masked from package:base:

    as.Date, as.Date.numeric



t test of coefficients:

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)   -23354.64    1252.74 -18.643 < 2.2e-16 ***
```

```
years_education   5750.48      83.34  69.000 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Analysis
This regresion suggests that each additional year of education results in an additional $5,750 of income.

b. Continue to suppose that we did not run the experiment, but that we saw the result that you noted in part (a). Tell a concrete story about why you don't believe that observational result tells you anything causal.

Analysis
Both income and years of education can be affected by a common externality. For example, if wealthier individuals were more likely to pursue higher level of education that would explain both the higher income and the higher years of education for that population. In such example, the externality would be pre-existing wealth prior education.

c. Now, let's get to using the natural experiment. We will define "having a high-ranked draft number" as having a draft number of 80 or below (1-80; numbers 81-365, for the remaining 285 days of the year, can be considered "low-ranked"). Create a variable in your dataset indicating whether each person has a high-ranked draft number or not. Using regression, estimate the effect of having a high-ranked draft number, the dummy variable you've just created, on years of education obtained. Report the estimate and a correctly computed standard error. (*Hint: Pay special attention to calculating the correct standard errors here. They should match how the draft is conducted.)

Implementation
We first define a variable stating whether the subject has a high-ranked draft number.

```
In [26]: d2$high_ranked <- ((d2$draft_number<=80)*1)
         head(d2)
```

draft_number	years_education	income	high_ranked
267	16	44573.90	0
357	13	10611.75	0
351	19	165467.80	0
205	16	71278.40	0
42	19	54445.09	1
240	11	32059.12	0

Regression

```
In [27]: library("sandwich")
         coeftest(lm(years_education~high_ranked,d2), vcov = sandwich)        # robust
```

```
t test of coefficients:
```

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 14.434305   0.016964 850.880 < 2.2e-16 ***
high_ranked  2.125756   0.037555  56.604 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Analysis

The effect of being high-ranked is an additional 2.13 years in education (std. error of 0.038)

d. Using linear regression, estimate the effect of having a high-ranked draft number on income. Report the estimate and the correct standard error.

Regression

```
In [28]: library("sandwich")
         coeftest(lm(income~high_ranked,d2), vcov = sandwich)              # robust; sandwic


t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 60761.89     233.37 260.365 < 2.2e-16 ***
high_ranked  6637.55     545.47  12.168 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Analysis

The effect of being high-ranked is an additional $6,673.55 in income (std. error of 545.47)

e. Divide the estimate from part (d) by the estimate in part (c) to estimate the effect of education on income. This is an instrumental-variables estimate, in which we are looking at the "clean" variation in both education and income that is due to the draft status, and computing the slope of the income-education line as "clean change in Y" divided by "clean change in X". What do the results suggest?

Implementation

```
In [29]: 6673.55/2.125756
```

3139.37723802732

Analysis

This results suggests that after controlling for the draft status, the clean effect on income of an additional year of education is $3,139.38.

f. Natural experiments rely crucially on the "exclusion restriction" assumption that the instrument (here, having a high draft rank) cannot affect the outcome (here, income) in any other way except through its effect on the "endogenous variable" (here, education). Give one reason this assumption may be violated -- that is, why having a high draft rank could affect individuals' income other than because it nudges them to attend school for longer.

Analysis

The one other obvious reason is that having a high draft rank increases the probability of military service. This in turn can have an effect on other attributes that drive wages, such as relevant labor market experience.

In this case the fact that a high-ranked individual did military service has an effect in income because (a) didn't acquire relevant skills through education and (b) didn't gain relevant experience in the labor market.

g. Conduct a test for the presence of differential attrition by treatment condition. That is, conduct a formal test of the hypothesis that the "high-ranked draft number" treatment has no effect on whether we observe a person's income. (Note, that an earning of $0 *actually* means they didn't earn any money.)

Implementation

As seen below, the dataset does not have records for which income was not observed

```
In [30]: head(d2[d2$income ==NaN])
```

draft_number	years_education	income	high_ranked

Everyone seems to have an observed income, so we'll have to use a different framework to arrive at our results. If build upon the assumption that every draft number should have the same the expected number of observations, the there should be no difference between the average number of subjects observed in the 'high ranked' vs 'low ranked' groups.

In order to implement this, we'll first create an aggregate variable at the draft number level and then use it as the dependent variable in a regression framework.

```
In [50]: library('plyr')
         agg<-count(d2, "draft_number")
         d2Enriched <- merge(d2,agg,by="draft_number")
         head(d2Enriched)
```

draft_number	years_education	income	high_ranked	income_observed	freq
1	15	103887.10	1	1	44
1	13	45406.28	1	1	44
1	16	38930.93	1	1	44
1	15	48913.50	1	1	44
1	17	32418.87	1	1	44
1	15	52627.48	1	1	44

Regression and analysis

```
In [54]: coeftest(lm(freq~high_ranked,data=d2Enriched))
```

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 55.945696   0.058736 952.490 < 2.2e-16 ***
high_ranked -6.091486   0.131631 -46.277 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The above suggests that on average, the number of subjects reporting income in the 'high ranked' population is ~6 (11%) lower as compared to the 'low ranked' population. This difference is statistically significant.

h. Tell a concrete story about what could be leading to the result in part (g).

Given that a subset of the subjects with 'high ranked' draft numbers were actually drafted for military service, it is possible that they were unable to resume what would have otherwise been a productive work life after service. This could have been due to various reasons such as post-war traumatic disorders, physical disabilities or simply because the great number of casualties.

Given the assumption provided as part of the question *'Suppose that these data are a true random sample of IRS records and that these records measure every living American's income without error.'* the scenario of not being able to work and reporting an income of $0 is not viable, so casualties would seem to be the more likely scenario.

i. Tell a concrete story about how this differential attrition might bias our estimates.

The fundamental point when it comes to attrition is that it challenges the assumption that the control and treatment populations are on average comparable and no different at any level. Whereas this was true at the pre-treatment due to randomization, it may no longer be true after a subset of the population is no longer represented.

This can render any interpretation of the treatment effects invalid given how heterogeneous the groups could be.

One hypothetical scenario that comes to mind that could result in biased estimates is that more risk-taker individuals could have been more likely to die in the war, but also more likely to have earned more income if they were to participate in the labor market, as a result the average income of the treatment group (high ranked) would be lower due to the attrition, resulting in a biased estimated treatment effect.

3   3. Green and Gerber Practice Problems

Note, none of these require you to program anything. Instead, I'm aiming to have you think critically about these forms of designs. Have fun!

3.1   a. Field Experiments 11.10

The table below considers a case where subjects are randomly assigned treatments but then choose to take different dosages. In this experiment subjects were randomly assigned to receive one of

two pre-recorded phone messages. the treatment script encouraged people to vote and revealed whether members of the household voted int he past two elections. The control script encouraged people to recycle. Both calls were made the day before the election. Bothc calls were answered at similar rates. The table below presents voting rates for households that answered the phone call. Voting rates are broken down by how long the person answering the phone took to hang up after the recorded message began.

Treatment Group	1-10 seconds	11-20 seconds	21-30 seconds	31-40 seconds	Total
Call encouraged voting	16.6% (187)	17.4% (784)	19.7% (983)	24.3% (2,032)	21.4% (3,986)
Call encouraged recycling	17.5% (143)	18% (619)	18.9% (1,132)	19.8% (2,012)	19.2% (3,906)

Entries are percent voting, with Ns in parentheses. The sample is restricted to one-voter households. Both scripts were ~35 seconds long.

a. Focusing only on households that answered the phone, estimate the apparent average effect of assignment to the script that encouraged voting

Answer

The apparent average effect of assignment to the script that encouraged voting is 2.2% increase in voting rates.

b. Does this table provide convincing evidence that "the longer a person listens to a recorded message that encourages voting, the more effective that message will be in terms of boosting turnout? Why or Why not?"

Answer

It does not. The only thing it shows is that if a person is **willing** to listen to a recorded message that encourages voting for longer then he/she is more likely to vote. It **does not** prove however that if we force people to listen to longer messages on voting then they will be more likely to vote.

3.2   b. Field Experiments 12.3

(The paper is available at `./readings/Olken.2010.pdf`.)

In an experiment designed to evaluate the effects of political institutions, Olken randomly assigned 49 villages in Indonesia to alternative political processes for selecting development projects. Some villages were assigned to the status quo selection procedure (village meetings with low attendance), while others were assigned to use an innovative method of direct elections (a village-wide plebiscite). Consistent with expectations, participation in the plesbicite was 20 times greater than attendance at the village meetings. Olken examines the new procedure's effect on which projects are selected and how the villages feel about the selection process. He finds that there are minimal changes in which projects are selected. However a survey after the project selection found that the villagers who were assigned to the plebiscite reported much greater satisfaction with the project selection process, and were significantly more likely to view the selection as fair, and the project as useful and in accordance with their own and the people's wished.

22

a. One part of this experiment focuses on whether the treatment influences which projects villages select. These results are reported in Figure 1, and the study is described on pp. 244-247. Describe the experimental subjects, what units are assigned to treatment versus control? What is the treatment?

Experimental Subjects, Treatment and Control groups

49 villages across three subdistrics, two of which are very homogeneous - one is predominantly muslim, another predominantly christian - and a third one fairly heterogeneous with larger size villages. Each village is comprised of between 2 and 7 hamlets.

Overall the goal of the participation mechanisms was to choose a list of projects that the village was nominating for funding. There were two groups of projects to be nominated, one set of general projects to be chosen by the entire village, and a set of women's projects, nominated exclusively by women.

Control group is comprised of mainly the political elite (e.g. goverment officials, village head). For the general village meetings mostly men attended, whereas to women's village meetings, only women did.

Treatment groups were comprised of a more heterogeneous group as it was intended not to be dominated by political elite nor gender (women were given two ballots to participate in women only and general projects)

b. Suppose that in Indonesia, the plebiscite method is rare, but the village meeting is very common. How would this affect your interpretation of the finding?

This would suggest that researchers should be careful when attributing the effect to the right factors on the experiment design. In this setting, there could be a general stigma against village meetings perhaps as an association to the preceedeng government's regular village parliament, where village heads were seen as an extension of the central government.

This could have an anchoring effect that would discourage villagers to place trust in the village meetings and generate an overall excitement about a different method. In this case the experiment design could be capturing to co-existing factors, in one hand, the effectiveness of having a less elite-dominated method to voice opinions, in addition to having an increase in participation due to general excitement of trying something new.

c. The level of satisfaction is measured by survey responses. From the description on p.250, can you tell who conducted the surveys and whether the interviewers were blinded as to the respondents' assignment to treatment or control? Why might survey measures of satisfaction be susceptible to bias?

There is no information elaborating on who conducted the surveys, and whether there were differences in the methods applied. Survey measures of satisfaction can certainly be susceptible to bias. For example, interviewees could answer differently if the interviewe asks them to fill in the survey in front of him, as opposed to filling them in privacy and hand them over in a sealed, unnamed envelop.

d. There is no indication that the treatment and control villages had contact with each other. Imagine, however that people regularly communicated across village lines. What assumption might be violated by this interaction? Discuss how cross-village communcation might affect treatment effect estimates. What design or measurement strategy might address possible concerns?

The non-interference assumption will be violated. Villagers could express different levels of satisfaction when learning that their neighboring village had a different treatment than theirs and may not constitute a true treatment or control group.

One way to account for potential spillovers is to develop a framework here a probability of interference is assigned to each unit. That way the subject potential outcomes could be related to the probability that they learned about the neighboring village's treatment (spillover).

e. Olken concludes that, consistent with the views of many democratic theorist, participation in political decision making can substantially increase satisfaction with the political process and political legitimacy. Does the experiment provide convincing evidence for this general proposition? What are some of the limitations noted by Olkent (see pp.265-266)? What additional limitations does the experiment have? How might you address these concerns in a future experiment?

In my view the experiment does not provide convincing evidnce for this proposition. Some of the key limitations are vaguely highlighted by Olken:

- Only a subset of villages in the entirety of Indonesia, can be hard to generalize
- Only short-term effects were measured

However, there are quite a few details that have been left out that could totally change the interpretation of the results. For instance, (1) how likely were villages not to learn about the treatment assigned to other villages (spillovers), (2) how exactly were survey results captured? where these done in a consistent way that minimizes the likelihood of introducing bias? (3) Were there treatment effects different across the three districts?

f. It is often claimed that sort-term effects may diminish over time, but the short-run outcome measurements nevertheless reliably indicate the direction, if not the magniture, of the long-term effects. However, inf an institutional change is thought to be a durable feature of the political world, leaders and voters may change their behavior and the way they compete for power. Speculate on why the long-term effects of the plebiscite on satisfaction with the decision process might be negative despite the initial positive response.

One scenario where satisfaction with the decision process might diminish is if the lobbying activity increases in a way that exerts an oversized influence in the projects that ultimately get chosen. In this case, villagers could go back to the believes that their participation is of little importance given the decision making process is still heavily influenced by a powerful minority.

3.3   c. Field Experiments 12.5

The Simester et al. study showed how incomplete outcome measurement can lead to erroneous conclusions. On that note, suppose researchers are concerned with the health consequences of what people eat and how much they weigh. Cocnsider an experiment designed to measure the effect of a proposal to help people diet. Subjects are invited to a dinner and are randomly given regular-sized or slightly larger than regular-sized plates. Hidden cameras record how much peope eat, and the researchers find that those given larger plates eat substantially more food than those assigned small plates. A statistical test shows that the apparent treatment effect is far greater than one would expect by chance. The authors conclude that a minor adjustment, reducing plate size, will help people lose weight.

a. How convincing is the evidence regarding the effect of plate size on what people eat and how much they weigh?

Not very convincing. The researchers are extrapolating their observations on a single meal to overall effects in weight. This is completely ignoring fundamental factors such as what's net intake of food through the day for these individuals. What if subjects that were given smaller plates are hungry after an hour and have the urge to eat again, whereas the subjects with a larger plate are not? This can mean that on a net basis the group given the smaller plates eat more throughout the day, totally contradicting the researchers' conclusions.

b. What design and measurement improvements do you suggest?

First, the experiment should be designed to capture the overall effect of smaller meal portions through an extended period of time. Second, metrics on the net intake of daily food should be captured, as well as other factors such as the amount of exercise or physical activity the subjects do through the experiment.

3.4 5. Think about Treatment Effects

Throughout this course we have focused on the average treatment effect. Think back to *why* we are concerned about the average treatment effect. What is the relationship between an ATE, and some individuals' potential outcomes? Make the strongest case you can for why this is *good* measure.

ATE is a good measure as it helps us generalize the expeced effect of a given treatment on any given individual. This high-level data point helps drive conversations about the efficacy of treatments, policy or any changes that could affect a broad population as a whole.

However, one must be careful when making inferences about individual's potential outcomes given the ATE is intended to account for the heterogeneity of the population as a whole, and as such it is not intented to be an exact prescription of what each individual effect would be. It is prudent to look at the standard deviation and the potential range (for a given statistically significant level) that would bound the true treatment effect when analyzing individual's potential impact at any level.