Juanjo Carin

W241 (Field Experiments) – Problem Set 5 – MIDS Spring 2015

April 20, 2015

Contents

W241 – Problem Set 5

1. Online advertising natural experiment

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.)

```r
# Read dataset
library(knitr)
ps5_no1 <- read.csv("ps5_no1.csv")
colnames(ps5_no1)[2:3] <- c("total_ad_exposures", "treatment_ad_exposures")
attach(ps5_no1)
# Counts
ps5_no1_table <- table(total_ad_exposures, treatment_ad_exposures)
```

```r
kable(formatC(ps5_no1_table, digits = 0, format = "f", big.mark = ","),
      align = "r", caption = "Number of observations by number of total visits
      and visits on even seconds")
```

Table 1: Number of observations by number of total visits and visits on even seconds

	0	1	2	3	4	5	6
0	61,182	0	0	0	0	0	0
1	36,754	37,215	0	0	0	0	0
2	21,143	42,036	20,965	0	0	0	0
3	10,683	32,073	32,314	10,726	0	0	0
4	5,044	20,003	30,432	20,223	5,115	0	0
5	2,045	10,563	20,970	20,793	10,293	2,131	0
6	729	4,437	10,977	14,771	11,147	4,486	750

```r
# Frequency
(ps5_no1_table_freq <- round(prop.table(ps5_no1_table, 1), 3))
```

```
##                    treatment_ad_exposures
## total_ad_exposures     0     1     2     3     4     5     6
##                  0 1.000 0.000 0.000 0.000 0.000 0.000 0.000
##                  1 0.497 0.503 0.000 0.000 0.000 0.000 0.000
##                  2 0.251 0.500 0.249 0.000 0.000 0.000 0.000
##                  3 0.125 0.374 0.377 0.125 0.000 0.000 0.000
##                  4 0.062 0.248 0.377 0.250 0.063 0.000 0.000
##                  5 0.031 0.158 0.314 0.311 0.154 0.032 0.000
##                  6 0.015 0.094 0.232 0.312 0.236 0.095 0.016
```

```r
# Counts and frequency
library(gmodels)
CrossTable(total_ad_exposures, treatment_ad_exposures, digits = 3,
           prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  500000 
## 
## 
##                     | treatment_ad_exposures 
## total_ad_exposures |         0 |         1 |         2 |         3 |         4 |         5 |         6 | Row Total | 
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##                  0 |     61182 |         0 |         0 |         0 |         0 |         0 |         0 |     61182 | 
##                    |     1.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.122 | 
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##                  1 |     36754 |     37215 |         0 |         0 |         0 |         0 |         0 |     73969 | 
##                    |     0.497 |     0.503 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |     0.148 | 
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##                  2 |     21143 |     42036 |     20965 |         0 |         0 |         0 |         0 |     84144 | 
##                    |     0.251 |     0.500 |     0.249 |     0.000 |     0.000 |     0.000 |     0.000 |     0.168 | 
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##                  3 |     10683 |     32073 |     32314 |     10726 |         0 |         0 |         0 |     85796 | 
##                    |     0.125 |     0.374 |     0.377 |     0.125 |     0.000 |     0.000 |     0.000 |     0.172 |
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##                  4 |      5044 |     20003 |     30432 |     20223 |      5115 |         0 |         0 |     80817 |
##                    |     0.062 |     0.248 |     0.377 |     0.250 |     0.063 |     0.000 |     0.000 |     0.162 |
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##                  5 |      2045 |     10563 |     20970 |     20793 |     10293 |      2131 |         0 |     66795 |
##                    |     0.031 |     0.158 |     0.314 |     0.311 |     0.154 |     0.032 |     0.000 |     0.134 |
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##                  6 |       729 |      4437 |     10977 |     14771 |     11147 |      4486 |       750 |     47297 |
##                    |     0.015 |     0.094 |     0.232 |     0.312 |     0.236 |     0.095 |     0.016 |     0.095 |
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##       Column Total |    137580 |    146327 |    115658 |     66513 |     26555 |      6617 |       750 |    500000 |
## -------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
##
##
```

All the tables are lower triangular matrices, which makes sense—the number of times each user visited the Yahoo! homepage on an even second during the week of the campaign cannot exceed the total number of times each user visited the Yahoo! homepage during the week of the campaign.

Besides, the frequencies also make sense:

- the number of people who did not visit the Yahoo! homepage equals to the number of people who did not visit the Yahoo! home on an even second,
- those who visited the Yahoo! homepage once are almost equally splitted between even and odd seconds,
- of those who visited the Yahoo! homepage twice half, about 25% visited it on either odd or even seconds, and about the remaining 50% visited it once on an even second and once on an odd second,
- and so forth.

Actually, we should notice that visiting the Yahoo! homepage on an even second (and hence being assigned to the treatment group) is a binomial process with probability equal to 0.5 (since there is the same amount of even an odd seconds in a week), so the probability of visiting the Yahoo! homepage x times on an even second, provided that you visited it n times during that period is:

$$P(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} = \binom{n}{x} \cdot 0.5^x \cdot 0.5^{n-x} = \binom{n}{x} \cdot 0.5^n$$

And that's why the distribution of impressions looks like that.

So let's compare the probabilities we obtained before to those of a binomial distribution:

```r
binomial_table_freq <- matrix(0, 7, 7)
colnames(binomial_table_freq) <- rownames(binomial_table_freq) <- c(0:6)
for (i in 0:6) {
    for (j in 0:6) {
        binomial_table_freq[i+1, j+1] <- round(dbinom(j, i, 0.5), 3)
    }
}
difference <- round(100*(ps5_no1_table_freq - binomial_table_freq)/
                            binomial_table_freq, 2)
all(is.na(difference[abs(difference)>2.5]))
```

```
## [1] FALSE
```

None of the frequencies are 2.5 percentage points greater or lower than the ones we expected.

b. Your colleague proposes the code printed below to analyze this experiment. `summary(lm(week1 ~ treatment_ad_exposures, 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?

```r
summary(lm(week1 ~ treatment_ad_exposures))
```

```
##
## Call:
## lm(formula = week1 ~ treatment_ad_exposures)
##
## 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 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
```

It seems that each exposure to the ad increases revenues per user by $0.3.

But if we run a placebo test with the prior week's purchases as the outcome the effect of *future* exposure to the ads is still highly statistically significant:

```r
summary(lm(week0 ~ treatment_ad_exposures))
```

```
##
## Call:
## lm(formula = week0 ~ treatment_ad_exposures)
##
## 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 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
```

I.e., the placebo test failed: we also detected an effect on an outcome that might not be affected by the independent variable *treatment_ad_exposures*.

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.

What we should demonstrate is that **conditional on the total number of visits to the Yahoo! homepage**, being exposed to the ad increases the revenues from those users. Just as Ebonya Washington tried to demonstrate in her paper that, *conditional on total number of children, each daughter increases a congressperson's propensity to vote liberally, particularly on reproductive rights issues.*

Experimenters have not randomized how many times a user visits the Yahoo! homepage—each user decides that—but whether the ad for the advertiser is shown or not when that happens. Hence, *treatment_ad_exposures* are not randomly allocated, and they depend on *total_ad_exposures*, which are not random whatsoever. I.e., the number of times a user visited the Yahoo! homepage on an even second during the week of the campaign is highly correlated with the number times he or she visited it at any given time. E.g.,—as introduced in part (a)—the probability that a user visited the Yahoo! homepage on an even second 4 times during the week of the campaign is 0 if that user visited the Yahoo! homepage 3 times or less during that period, 0.06 if that user visited the Yahoo! homepage 4 times, 0.15 if that user visited the Yahoo! homepage 5 times, and 0.24 if that user visited the Yahoo! homepage 6 times.

So the bias comes from an omitted variable (*total_ad_exposures*). The number of times a user visited the Yahoo! homepage during the week of the campaign is very likely to be correlated with the number of times that user visits the Yahoo! homepage at any given week, which in turn may be somehow correlated with his or her purchases of the advertised product (and hence the interest of the advertiser in using Yahoo!).

```r
kable(merge(aggregate(week0 ~ total_ad_exposures, ps5_no1, mean),
            aggregate(week1 ~ total_ad_exposures, ps5_no1, mean)), digits = 2,
      caption = "Mean revenues per user on weeks 0 and 1 depending on the total
      number of visits to the Yahoo! homepage on week 1")
```

Table 2: Mean revenues per user on weeks 0 and 1 depending on the total number of visits to the Yahoo! homepage on week 1

total_ad_exposures	week0	week1
0	1.31	1.30
1	1.62	1.58
2	1.86	1.84
3	2.08	2.07
4	2.31	2.33
5	2.54	2.57
6	2.84	2.83

```r
week1_table_1 <- aggregate(week1 ~ treatment_ad_exposures + total_ad_exposures,
                           , mean)
week1_table_2 <- xtabs(week1 ~ total_ad_exposures + treatment_ad_exposures,
                       data = week1_table_1)
week1_table_3 <- apply(week1_table_2, 2, formatC, digits = 2, format = "f",
                           drop0trailing = FALSE)
week1_table_3[upper.tri(week1_table_2)] <- ""
```

```r
kable(week1_table_3, caption = "Mean revenues per user on week 1 depending on
        the total number of visits to the Yahoo! homepage on week 1 and the total
        number of visits to the Yahoo! homepage on an even second")
```

Table 3: Mean revenues per user on week 1 depending on the total number of visits to the Yahoo! homepage on week 1 and the total number of visits to the Yahoo! homepage on an even second

	0	1	2	3	4	5	6
0	1.30						
1	1.56	1.61					
2	1.79	1.85	1.88				
3	1.97	2.04	2.09	2.21			
4	2.23	2.24	2.34	2.40	2.43		
5	2.35	2.49	2.53	2.61	2.70	2.65	
6	2.78	2.84	2.79	2.82	2.84	2.90	3.20

As shown in the tables above, the revenues per user are highly correlated with the visits per user to the Yahoo! homepage on that week, or any other. That is the reason why the number of times a user is exposed to the ads during the week of the campaign is also correlated with his or her purchases *before* he or she was exposed (or any other week).

In brief, we need to add *total_ad_exposures* as a covariate, to introduce a set of fixed effects for the total number of visits to the Yahoo! homepage. That way, the coefficient associated with *treatment_ad_exposures* would identify the impact of being exposed to an ad for the advertiser, as compared to an ad for other products. Excluding the visit fixed effects causes our estimate to combine both the impact of visiting the Yahoo! homepage and the impact of being exposed to an ad for the advertiser.

d. Implement the procedure you propose from part (c), run the placebo test for the Week 0 data again, and report the results.

```r
summary(lm(week0 ~ total_ad_exposures + treatment_ad_exposures))
```

```
##
## Call:
## lm(formula = week0 ~ total_ad_exposures + treatment_ad_exposures)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.817 -2.079 -1.589  2.455  7.823
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.345375   0.007295 184.436   <2e-16 ***
## total_ad_exposures      0.245348   0.003149  77.922   <2e-16 ***
## treatment_ad_exposures -0.002245   0.004629  -0.485    0.628
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.779 on 499997 degrees of freedom
## Multiple R-squared:  0.02555,    Adjusted R-squared:  0.02555
## F-statistic:  6556 on 2 and 499997 DF,  p-value: < 2.2e-16
```

This new placebo test now succeeds: the effect of ads users might see the *following* week is no where near statistically significant ($p = 0.628$), and thus consistent with the assumption that the potential exposure to ads in the future has no effect on purchases during the current week.

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).

```
summary(lm(week1 ~ total_ad_exposures + treatment_ad_exposures))
```

```
##
## Call:
## lm(formula = week1 ~ total_ad_exposures + treatment_ad_exposures)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.003 -2.104 -1.542  2.447  8.110
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.317960   0.007263  181.47   <2e-16 ***
## total_ad_exposures     0.224478   0.003135   71.61   <2e-16 ***
## treatment_ad_exposures 0.056340   0.004609   12.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.767 on 499997 degrees of freedom
## Multiple R-squared:  0.02782,    Adjusted R-squared:  0.02781
## F-statistic:  7153 on 2 and 499997 DF,  p-value: < 2.2e-16
```

Each additional visit to the Yahoo! homepage increases revenue per user by $0.22, and conditional on those visits, each exposure to the ad (i.e., each additional visit to the Yahoo! homnepage on an even second) increases revenue per user by $0.06 ($0.005), as compared to a visit to the Yahoo! homepage on an odd second (i.e., to an exposure to an ad for other products). The effect of ad exposure is highly statistically significant ($p = 2.3e - 34$).

The coefficients we have obtained are also coherent with the mean revenues by number of visits and number of visits on an even second that we obtained before (see Table 3):

E.g., notice that (if there are no visits to the Yahoo! homepage on an even second; see the 1st column of the table above), each additional visit to the Yahoo! homepage (on an odd second) increases the mean revenue per user by approximately $0.22:

```
round(week1_table_2[2:7, 1] - week1_table_2[1:6, 1], 2)
```

```
##    1    2    3    4    5    6
## 0.26 0.23 0.19 0.26 0.12 0.43
```

And (provided users have visited the Yahoo! homepage at least once), a visit to that page on an even second (see the first 2 columns of the previous table) increases the mean revenue per user by approximately $0.06:

```
round(week1_table_2[2:7, 2] - week1_table_2[2:7, 1], 2)
```

```
##    1    2    3    4    5    6
## 0.05 0.06 0.07 0.01 0.14 0.05
```

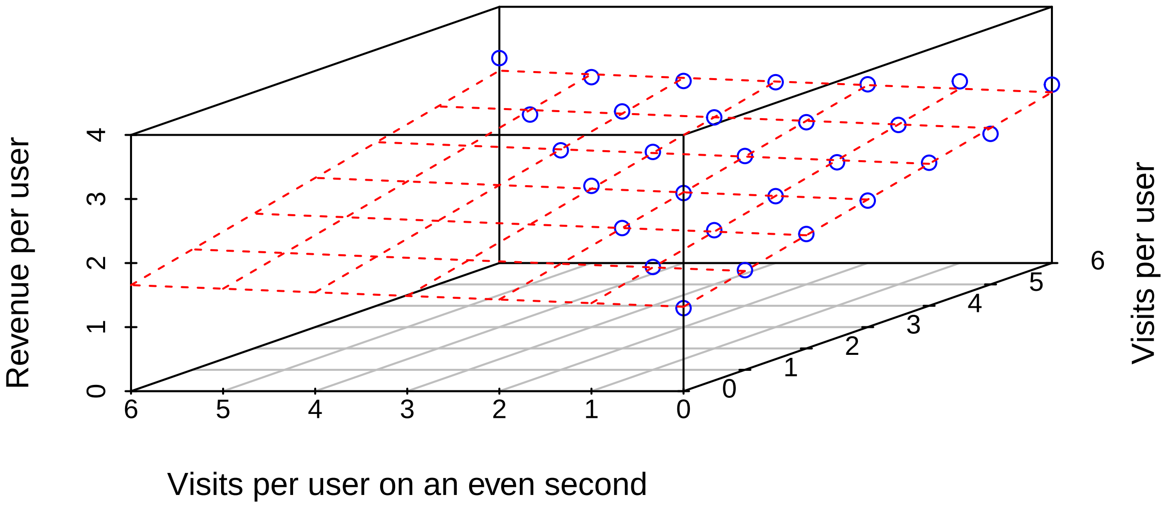

Figure 1: Regression plane and mean revenues per number of total visits and visits on an even second

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.

By considering the total number of times each user visited Yahoo! homepage during the week of the campaign—a variable that is not dependent on the number of visits on an even second,but the other way around—, I have isolated the fixed effects these total visits have on users' purchases (here I don't mean a causal effect, of course, but a correlation). On the other hand, by **omitting** that variable, his/her estimate of the causal effect of ad exposure includes a selection bias term that exists even if the causal effect of ad exposure does not exist, and that reflects the fact that number of visits to the Yahoo! homepage changes the composition of the pool of users that visit it on an even second: the more likely they visit the Yahoo! homepage, the more likely they are to do that on an even second. So that selection bias is like an apple-to-oranges comparison, so to speak—it is a comparison of users who visit the Yahoo! homesite very frequently to users who don't.

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

```
aggregate_sales_1to10 <- apply(ps5_no1[, 5:14], 1, sum)
summary(lm(aggregate_sales_1to10 ~ total_ad_exposures + treatment_ad_exposures))
```

```
##
## Call:
## lm(formula = aggregate_sales_1to10 ~ total_ad_exposures + treatment_ad_exposures)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.597  -7.372  -0.731   6.654  59.782
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            17.15081    0.02771 618.949   <2e-16 ***
## total_ad_exposures      2.22834    0.01196 186.307   <2e-16 ***
## treatment_ad_exposures  0.01274    0.01758   0.724    0.469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.56 on 499997 degrees of freedom
## Multiple R-squared:  0.1321, Adjusted R-squared:  0.1321
## F-statistic: 3.804e+04 on 2 and 499997 DF,  p-value: < 2.2e-16
```

The effect of each treatment ad exposure on total purchases *during and after* the campaign is negligible: it is close to zero (\$0.01), and no where near statistically significant ($p = 0.469$).

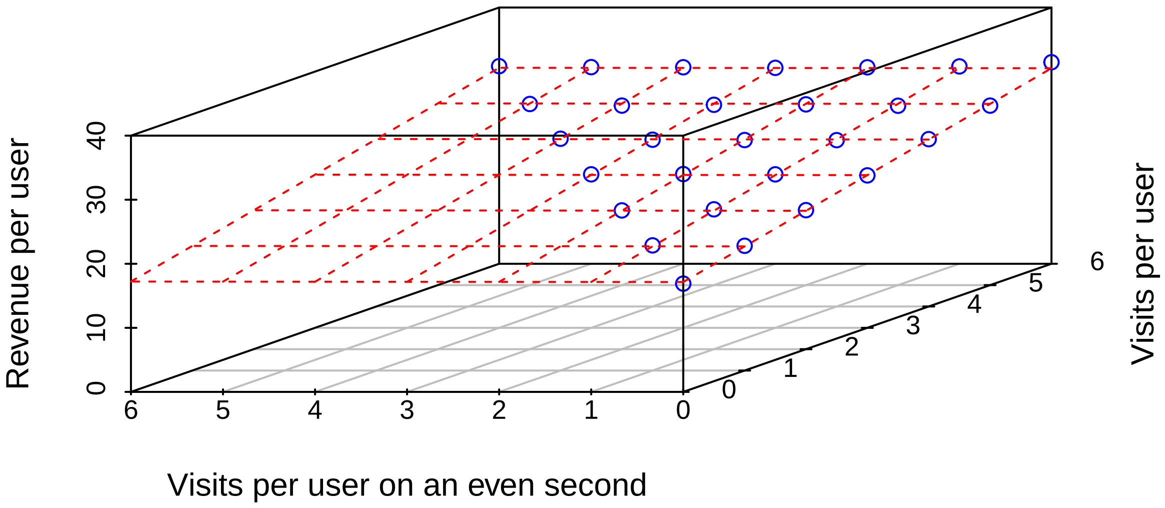

Figure 2: Regression plane and mean aggregate revenues (weeks 1 through 10) per number of total visits and visits on an even second

See the effect of the coefficient of *treatment_ad_exposures* being close to zero: the intersection of the regression plane with any plane where $x = \mathit{total\_ad\_exposures}$ remains constant (e.g., $x = \mathit{total\_ad\_exposures} = \mathit{6}$) is almost parallel to the plane $z = \mathit{aggregate\_revenues} = \mathit{0}$.

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

```
aggregate_sales_2to10 <- apply(ps5_no1[, 6:14], 1, sum)
summary(lm(aggregate_sales_2to10 ~ total_ad_exposures + treatment_ad_exposures))
```

```
##
## Call:
## lm(formula = aggregate_sales_2to10 ~ total_ad_exposures + treatment_ad_exposures)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.856  -7.097  -0.697   6.382  54.079
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            15.83285    0.02654 596.493  < 2e-16 ***
## total_ad_exposures      2.00387    0.01146 174.901  < 2e-16 ***
## treatment_ad_exposures -0.04360    0.01684  -2.588  0.00964 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 499997 degrees of freedom
## Multiple R-squared:  0.1154, Adjusted R-squared:  0.1154
## F-statistic: 3.261e+04 on 2 and 499997 DF,  p-value: < 2.2e-16
```

As expected, the effect of each treatment ad exposure on total purchases *after* the campaign (−\$0.04 (\$0.02)) almost compensates that effect *during* the campaign (\$0.06 (\$0.005)), and that's why the overall effect estimated in part (g) is negligible.

> Another way (maybe less accurate) to detect this situation would have been running one regression per week:

```
Dep_Variables <- lapply(c(1:10), function(x) paste0("week", x, " ~"))
Ad_Exposure_Effects_1 <- Map(function(x)
    summary(lm(as.formula(paste(x, "total_ad_exposures", "+",
                                         "treatment_ad_exposures")),
                  data = ps5_no1))$coefficients[3, c(1:2, 4)], Dep_Variables)
Ad_Exposure_Effects_2 <- matrix(unlist(Ad_Exposure_Effects_1), nrow = 10,
                                          ncol = 3, byrow = TRUE)
colnames(Ad_Exposure_Effects_2) <- names(Ad_Exposure_Effects_1[[1]])
rownames(Ad_Exposure_Effects_2) <- lapply(Dep_Variables, function(x)
    gsub(" ~", "", x))
```

```r
kable(Ad_Exposure_Effects_2, digits = 4, caption = "Ad exposure effect each week")
```

Table 4: Ad exposure effect each week

	Estimate	Std. Error	Pr(>|t|)
week1	0.0563	0.0046	0.0000
week2	-0.0106	0.0054	0.0517
week3	-0.0002	0.0055	0.9689
week4	0.0036	0.0054	0.5077
week5	-0.0086	0.0054	0.1160
week6	-0.0002	0.0054	0.9774
week7	-0.0047	0.0054	0.3874
week8	-0.0039	0.0054	0.4711
week9	-0.0099	0.0054	0.0690
week10	-0.0092	0.0054	0.0926

The effect after the campaign is always negative (so the sum of the effects through weeks 1 to 10 is close to zero: $0.01; actually the value is exactly the same—up to the 5th decimal—than the one we obtained in part (g), but we don't have a confidence interval with this method).

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

This *compensation of the effects* or **intertemporal substitution** (people buy more while they are exposed to the ads, but less afterwards) may be due to:

1. **excess stock** (their needs didn't change so they just purchased *in advance* what they consumed in the succeeding weeks), or
2. **satiation** or **saturation** (they reduced their consumption of this kind of product or even looked for other alternatives).

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.

```r
summary(lm(week1 ~ total_ad_exposures + treatment_ad_exposures * product_b))
```

```
##
## Call:
## lm(formula = week1 ~ total_ad_exposures + treatment_ad_exposures *
##     product_b)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.032 -2.194 -1.500  2.439  8.071
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.289270   0.008010 160.960   <2e-16 ***
## total_ad_exposures                0.210868   0.003230  65.277   <2e-16 ***
## treatment_ad_exposures            0.061251   0.005637  10.867   <2e-16 ***
## product_b                         0.170320   0.013186  12.917   <2e-16 ***
## treatment_ad_exposures:product_b -0.010100   0.006490  -1.556     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.766 on 499995 degrees of freedom
## Multiple R-squared:  0.02848,    Adjusted R-squared:  0.02847
## F-statistic:  3664 on 4 and 499995 DF,  p-value: < 2.2e-16
```

There is an effect of product B (see the corresponding coefficient and its *p*-value) but just because the advertiser makes more money from it than from product A (not necessarily because their ads are more effective): the advertiser may sell more units of product B per user, or it may have a higher price or margin.

```
t.test(week0[product_b == 0], week0[product_b == 1])
```

```
##
##  Welch Two Sample t-test
##
## data:  week0[product_b == 0] and week0[product_b == 1]
## t = -53.0278, df = 377241.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4610082 -0.4281441
## sample estimates:
## mean of x mean of y
##  1.871271  2.315848
```

But (as the interaction term—the coefficient of *treatment_ad_exposures* * *product_b* suggests) the effect of being exposed to the ads of product B is not statistically significant from the effect of being exposed to the ads of product A.

> Even if the effect of the ads for each separate product may seem different ($0.04 for product B vs. $0.07 for product A).

```
summary(lm(week1 ~ total_ad_exposures + treatment_ad_exposures,
              ps5_no1[product_b == 1, ]))
```

```
##
## Call:
## lm(formula = week1 ~ total_ad_exposures + treatment_ad_exposures,
##     data = ps5_no1[product_b == 1, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.021 -2.407 -1.882  2.856  8.068
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.444893   0.015651   92.32  < 2e-16 ***
## total_ad_exposures     0.218411   0.005313   41.11  < 2e-16 ***
## treatment_ad_exposures 0.044232   0.007192    6.15 7.75e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.052 on 199959 degrees of freedom
## Multiple R-squared:  0.01879,    Adjusted R-squared:  0.01878
## F-statistic:  1915 on 2 and 199959 DF,  p-value: < 2.2e-16
```

```r
summary(lm(week1 ~ total_ad_exposures + treatment_ad_exposures,
              ps5_no1[product_b == 0, ]))
```

```
##
## Call:
## lm(formula = week1 ~ total_ad_exposures + treatment_ad_exposures,
##     data = ps5_no1[product_b == 0, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.933 -1.978 -1.296  2.190  7.186
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.295811   0.007939  163.22   <2e-16 ***
## total_ad_exposures     0.204690   0.004028   50.82   <2e-16 ***
## treatment_ad_exposures 0.068154   0.006023   11.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 300035 degrees of freedom
## Multiple R-squared:  0.02623,    Adjusted R-squared:  0.02622
## F-statistic:  4041 on 2 and 300035 DF,  p-value: < 2.2e-16
```

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?

Not confident at all, until I conducted an experiment comparing ads for product A that include celebrity endorsements to ads for product A that do not (and even then, I wouldn't be entirely confident—celebrity endorsements may perform better than the other ad used in the future experiment, but not than any other ad, and there may be another—omitted—difference between the ads that is the real cause for the increase of effectiveness; maybe ads that include celebrity endorsements are more creative than ads that do not, and it's the creativity, not the celebrity endorsement, which increases the effectiveness of the ad).

---

2. Vietnam draft lottery

a. 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?

```r
# Read dataset
ps5_no2 <- read.csv("ps5_no2.csv")
attach(ps5_no2)
summary(lm(income ~ years_education))
```

```
##
## Call:
```

```
## lm(formula = income ~ years_education)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91655 -17459   -837  16346 141587
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -23354.64    1252.74  -18.64   <2e-16 ***
## years_education   5750.48      83.34   69.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26590 on 19565 degrees of freedom
## Multiple R-squared:  0.1957, Adjusted R-squared:  0.1957
## F-statistic:  4761 on 1 and 19565 DF,  p-value: < 2.2e-16
```

It suggests that each additional year of education increases income by $5, 750.48.

> None of the subjects in the dataset has less than 9 years of education so we don't have to worry about the negative intercept (average income with 0 years of education).

b. Tell a concrete story, not having to do with the natural experiment, about why the observational regression in part (a) may be biased.

Earnings (the outcome or dependent variable) depend on a series of factors (e.g., ability, family background) apart from years of education (the independent variable or regressor we are considering). If those other factors are correlated with the only one we have considered, there is an **omitted variable bias** (which may be negative or positive depending on how years of education is correlated with those other factors).

For instance, being raised in a wealthy family increases your chances to enroll in college. But it may also influence your interests and socialnetwork, and hence your future job and income.

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.

```r
high_ranked <- ifelse(draft_number <= 80, 1, 0)
model_education_ranked <- lm(years_education ~ high_ranked)

# Function to calculate CSEs
cl <- function(fm, cluster){
    require(sandwich, quietly = FALSE)
    require(lmtest, quietly = FALSE)
    M <- length(unique(cluster))
    N <- length(cluster)
    K <- fm$rank
    dfc <- (M/(M-1))*((N-1)/(N-K))
    uj <- apply(estfun(fm),2, function(x) tapply(x, cluster, sum));
    vcovCL <- dfc*sandwich(fm, meat=crossprod(uj)/N)
    coeftest(fm, vcovCL)
}

cl(model_education_ranked, draft_number)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 14.434305   0.017703 815.345 < 2.2e-16 ***
## high_ranked  2.125756   0.038188  55.666 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Having a high-ranked draft number increases years of education by 2.13 (0.04)—from an average of 14.43 years to 16.56 years.

(We have to use clustered SEs because randomization—assignment of draft number—was made at the birthday level: all individuals who were born the same day are assigned the same draft number.)

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

```
model_income_ranked <- lm(income ~ high_ranked)
cl(model_income_ranked, draft_number)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 60761.89     244.36 248.656 < 2.2e-16 ***
## high_ranked  6637.55     511.90  12.966 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Having a high-ranked draft number increases income by \$6,637.55 (\$511.90)—from an average of \$60,761.89 to \$67,399.44.

e. Divide the estimate from part (d) by the estimate in part (c) to estimate the effect of education on income. What do the results suggest?

```
cl(model_income_ranked, draft_number)[2, 1] /
    cl(model_education_ranked, draft_number)[2, 1]
```

```
## [1] 3122.444
```

The implementation of regression with instrumental variables is similar to what we did in the presence of one-sided non-compliance: we use Two-Stage Least Squares (2SLS) instead of OLS (when there is one-sided non-compliance we are interested in the effect among Compliers; now we are interested in the effect among those whose years of education are affected by having a high-ranked draft number).

```r
library(AER)
model_income_education <- ivreg(income ~ years_education, ~ high_ranked)
cl(model_income_education, draft_number)
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      15691.58    3371.50  4.6542 3.274e-06 ***
## years_education   3122.44     225.89 13.8228 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An additional year of education increases income by $3,122.44 ($225.89).

The effect of education on income is not as high as we had naively estimated in part (a). The problem with that estimate—as previously mentioned in part (b)—was that years of education depend on multiple factors that remain unobserved. The instrument is correlated with years of education, but uncorrelated with any other factors that may affect it.

Now we are really comparing apples to apples: those who enrolled in college against those who did not, just due to the day they were born, and not by any other reason that might also influence their future incomee.

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.

Having a high-ranked draft number caused individuals (who chose not to enroll in college) to be more likely to serve in the army, which might have affected their future incomes. Depending on the source, veterans have higher or lower incomes than their nonveteran counterparts (e.g., due to specific programs or lack of opportunities, respectively).

According to other studies (like this one), it also changed those individuals' political views—those with a high-ranked draft number "became more anti-war, more liberal, and more Democratic in their voting." That may not necessarily be linked to higher or lower incomes, but their activism might have influenced the kind of jobs they were interested in, or how much time they wanted to devote to their jobs and how much to political activities.

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.

```r
# library(plyr)
# count(ps5_no2, "draft_number")
ps5_no2_bis <- aggregate(income ~ draft_number, data = ps5_no2, FUN = length)
colnames(ps5_no2_bis)[2] <- "observations"
ps5_no2_bis$mean_income <- aggregate(income ~ draft_number, data = ps5_no2,
                                     FUN = mean)[, 2]
```

```
ps5_no2_bis$high_ranked <- ifelse(ps5_no2_bis$draft_number <= 80, 1, 0)
t.test(ps5_no2_bis$observations[ps5_no2_bis$high_ranked == 1],
         ps5_no2_bis$observations[ps5_no2_bis$high_ranked == 0])
```

```
##
##  Welch Two Sample t-test
##
## data:  ps5_no2_bis$observations[ps5_no2_bis$high_ranked == 1] and ps5_no2_bis$observations[ps5_no2_bis$high_rank
## t = -6.6358, df = 123.311, p-value = 9.121e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.160996 -4.410934
## sample estimates:
## mean of x mean of y
##  48.70000  54.98596
```

```
summary(lm(observations ~ high_ranked, ps5_no2_bis))
```

```
##
## Call:
## lm(formula = observations ~ high_ranked, data = ps5_no2_bis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.986  -4.986   0.014   5.014  24.300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.9860     0.4346 126.532  < 2e-16 ***
## high_ranked  -6.2860     0.9282  -6.772 5.13e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.336 on 363 degrees of freedom
## Multiple R-squared:  0.1122, Adjusted R-squared:  0.1097
## F-statistic: 45.86 on 1 and 363 DF,  p-value: 5.128e-11
```

The number of individuals with a high-ranked draft number is statiscally significantly different (lower) than the number of individuals with a low-ranked draft number, which indicates the presence of differential attrition—since draft numbers were randomly assigned, there should be approximately the same number of observations per draft number (i.e., birthday) in each group, but it seems like treatment caused individuals to attrit.

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

As mentioned in part (g), "it seems like treatment caused individuals to attrit." That's actually the truth because treatment is closely related to being drafted going to war. Those who were drafted were much likelier to die during the war, and hence became unobservable by the time incomes were measured.

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

Treatment causes some observations that would be observed if assigned to the control group to be missing. We are not able to know those missing values (i.e., there is no way we could know what the effect of education would have been in those who died in Vietnam), so our estimate is not really the ATE, but just a part of it… and we can only guess what's the direction of that bias.

The ATE may be expressed as:

$$E[Y_i(1) - Y_i(0)] =$$
$$E[R_i(1)]E[Y_i(1) \mid R_i(1) = 1] + \{1 - E[R_i(1)]\}E[Y_i(1) \mid R_i(1) = 0] -$$
$$E[Y_i(0)] = E[R_i(0)]E[Y_i(0) \mid R_i(0) = 1] - \{1 - E[R_i(0)]\}E[Y_i(0) \mid R_i(0) = 0]$$

But what we're able to estimate is:

$$E[Y_i(1) \mid R_i(1) = 1] - E[Y_i(0) \mid R_i(0) = 1]$$

To build a simple story, let's assume that the effect of education is a dummy variable (enrolling in college) and that there is no need for an instrument (i.e., we were able to randomly assign enrollment in college).
Let's suppose $E[Y_i(1) \mid R_i(1) = 1] = \$50,000$ and $E[Y_i(0) \mid R_i(0) = 1] = \$40,000$. I.e., we observe a mean income of \$50,000 among those who enrolled in college, and a mean income of \$40,000 among those who did not, so the estimate a positive effect of \$10,000.
Now let's suppose $E[R_i(1)] = 1$ (all the individuals who enrolled in college still lived by the time we collected data, so all of them were observable), but $E[R_i(0)] = 0.8$ (20% of those who did not enroll in college went to war *and* died in Vietnam).
Therefore, the ATE would be:
$ATE = \$50,000 - 0.8 \times \$40,000 - 0.2 \times E[Y_i(0) \mid R_i(0) = 0] = \$50,000 - \$32,000 - 0.2 \times E[Y_i(0) \mid R_i(0) = 0] = \$18,000 - 0.2 \times E[Y_i(0) \mid R_i(0) = 0]$

What is the value of $E[Y_i(0) \mid R_i(0) = 0]$ (what would the income of those who did not enroll in college—because they were drafted—have been, had they not died in Vietnam)? We can't know. Had it been also \$40,000, the ATE would be equal to our estimate of \$10,000. But if those individuals who died in Vietnam had survived and earned more than their counterparts (say \$45,000, because the government developed special programs for veterans), the ATE would be lower (\$9,000\$) and we would have over-estimated the effect, and vice versa.

---

3. FE 11.6

a. Estimate the ATE for each study.

```
# Replicate the table in page 380
ps5_no3_table <- data.frame(Study = c(rep("Michigan2006", 4),
                                      rep("Michigan2007", 4),
                                      rep("Illinois2009", 4)),
                            Treated = rep(c(0, 0, 1, 1), 3),
                            Voted = rep(c(0, 1, 0, 1), 3),
                            N = c(26481 - 8755, 8755, 5310 - 2123, 2123,
                                  348277 - 88960, 88960, 12391 - 3791, 3791,
                                  15676 - 2600, 2600, 9326 - 1936, 1936),
                            stringsAsFactors = F)
# Create the dataframe based on the table in page 380
ps5_no3 <- data.frame(Study = unlist(mapply(rep, ps5_no3_table$Study,
                                                            ps5_no3_table$N)),
                                      Treated = unlist(mapply(rep, ps5_no3_table$Treated,
                                                            ps5_no3_table$N)),
                                      Voted = unlist(mapply(rep, ps5_no3_table$Voted,
                                                         ps5_no3_table$N)))
ps5_no3$Study <- factor(ps5_no3$Study,
                                levels = levels(ps5_no3$Study)[c(2, 3, 1)])
# Check that we've correctly replicated the table in page 380
table(ps5_no3[, c(2,3,1)])
```

```
## , , Study = Michigan2006
##
##        Voted
## Treated      0      1
##       0  17726   8755
##       1   3187   2123
##
## , , Study = Michigan2007
##
##        Voted
## Treated      0      1
##       0 259317  88960
##       1   8600   3791
##
## , , Study = Illinois2009
##
##        Voted
## Treated      0      1
##       0  13076   2600
##       1   7390   1936
```

```
library(pander)
ftable(addmargins(table(ps5_no3), FUN = list(Total = sum), quiet = TRUE))
```

```
##                        Voted      0      1  Total
## Study        Treated
## Michigan2006 0                17726   8755  26481
##              1                 3187   2123   5310
##              Total            20913  10878  31791
## Michigan2007 0               259317  88960 348277
##              1                 8600   3791  12391
##              Total           267917  92751 360668
## Illinois2009 0                13076   2600  15676
##              1                 7390   1936   9326
##              Total            20466   4536  25002
## Total        0               290119 100315 390434
##              1                19177   7850  27027
##              Total           309296 108165 417461
```

```
# A function to calculate RSEs
RSEs <- function(model){
    require(sandwich, quietly = TRUE)
    require(lmtest, quietly = TRUE)
    newSE <- vcovHC(model)
    coeftest(model, newSE)
    }
```

```
Study <- unique(ps5_no3_table$Study)
Encouragement_Effect_1 <- Map(function(x)
    RSEs(lm(Voted ~ Treated, data = subset(ps5_no3, Study == x))),
    x = Study)
Encouragement_Effect_2 <- lapply(c(1:3), function(x)
    Encouragement_Effect_1[[x]][2, c(1:2, 4)])
Encouragement_Effect_2 <- matrix(unlist(Encouragement_Effect_2), nrow = 3,
                                 ncol = 3, byrow = TRUE)
colnames(Encouragement_Effect_2) <-
    colnames(Encouragement_Effect_1[[1]][, c(1:2, 4)])
rownames(Encouragement_Effect_2) <- Study
options(digits=3)
print(Encouragement_Effect_2)
```

```
##              Estimate Std. Error Pr(>|t|)
## Michigan2006   0.0692    0.00732 3.46e-21
## Michigan2007   0.0505    0.00421 3.09e-33
## Illinois2009   0.0417    0.00514 5.22e-16
```

As shown above, the ATE for the study in Michigan, 2006, was 6.92 (0.73) percentage points, the ATE for the study in Michigan, 2007, was 5.05 (0.42), and the ATE for the study in Illinois, 2009, was 4.17 (0.51). All these ATEs are highly statistically significant.

b. Estimate the standard error for each study. Use the standard errors (squared) to calculate the precision of each study.

The *robust* standard errors are already shown in part (a) (see the 2nd column of the last table), but I've also calculated them as chapter 11 in the book says:

```
# ps5_no3_table_freq <- prop.table(table(ps5_no3[, c(2,3,1)]), c(1, 3))
# ps5_no3_table_freq <- matrix(ps5_no3_table_freq, ncol = 2, byrow = TRUE)
# ps5_no3_table_freq <- ps5_no3_table_freq[seq(2, 6, by = 2), ]
# rownames(ps5_no3_table_freq) <- Study
# colnames(ps5_no3_table_freq) <- c("Control", "Treatment")

ps5_no3_table_freq <- data.frame(Control = rep(NA, 3), Treatment = rep(NA, 3))
rownames(ps5_no3_table_freq) <- Study
for (study in unique(ps5_no3$Study)) {
    for (treated in c(0, 1)) {
        ps5_no3_table_freq[study, treated+1] <-
            round(mean(subset(ps5_no3, study == Study &
                                  treated == Treated)$Voted == 1), 3)
    }
    }
group_size <- as.numeric(tapply(ps5_no3_table$N,
                                (seq_along(ps5_no3_table$N) - 1) %/% 2, sum))
ps5_no3_table_freq$N_Control <- group_size[seq(1, 6, by = 2)]
ps5_no3_table_freq$N_Treatment <- group_size[seq(2, 6, by = 2)]
ps5_no3_table_freq$SE <- apply(ps5_no3_table_freq, 1, function(x)
    sqrt(x[2] * (1 - x[2]) / x[4] + x[1] * (1 - x[1]) / x[3]))
ps5_no3_table_freq$Precision <- ps5_no3_table_freq$SE^2
ps5_no3_table_freq
```

```
##              Control Treatment N_Control N_Treatment      SE Precision
## Michigan2006   0.331     0.400     26481        5310 0.00732  5.36e-05
## Michigan2007   0.255     0.306    348277       12391 0.00421  1.77e-05
## Illinois2009   0.166     0.208     15676        9326 0.00515  2.65e-05
```

We obtain the same results (compare the 2nd column of the last table in part (a) to the 5th column of the last table). The precision of each study is 0.0000536, 0.0000177, and 0.0000265, respectively.

c. Assuming that these three samples are random draws from the same population, calculate a precision-weighted average of the three studies.

```
ps5_no3_table_freq$Weight <- 1 / ps5_no3_table_freq$Precision
ps5_no3_table_freq$ATE <- Encouragement_Effect_2[, "Estimate"]
ps5_no3_table_freq$ATE_2 <- ps5_no3_table_freq$Treatment -
    ps5_no3_table_freq$Control
(Encouragement_Effect_Total <- weighted.mean(ps5_no3_table_freq$ATE,
                                                            ps5_no3_table_freq$Weight))
```

```
## [1] 0.0507
```

```
(SE_Encouragement_Effect_Total <- sqrt(1/sum(1/ps5_no3_table_freq$Precision)))
```

```
## [1] 0.00298
```

The weighted average of the three studies is 5.07 (0.30) percentage points. As expected, the SE of the pooled estimate is smaller than the SE of each of the studies whose results are pooled.

e. Use equation (~~11.3~~11.10) to estimate the variance of the precision-weighted average. Take the square root of the variance in order to obtain the standard error. In order to estimate the 95% confidence interval, use the following procedure, which is based on a large-sample approximation. Obtain the lower bound of the interval by subtracting 1.96 times the standard error from the precision-weighted average; obtain the upper bound of the interval by adding 1.96 times the standard error to the precision-weighted average.

Equation (11.10) was already used in part (c) to estimate the standard error of the precision-weighted average. Therefore, the variance is 0.0000089 (lower than the previous three ones, so the precision-weighted average is more precise) .

```
SE_Encouragement_Effect_Total^2
```

```
## [1] 8.85e-06
```

```
CI <- print(paste0("[", formatC(100*(Encouragement_Effect_Total - qnorm(.975) *
                                                    SE_Encouragement_Effect_Total),
                                          digits = 2, drop0trailing = FALSE,
                                          format = "f"), ", ",
                         formatC(100*(Encouragement_Effect_Total + qnorm(.975) *
                                            SE_Encouragement_Effect_Total), digits = 2,
                                  drop0trailing = FALSE, format = "f"), "]"))
```

```
## [1] "[4.48, 5.65]"
```

The estimated confidence interval, as shown above, is [4.48, 5.65].

f. Explain why the confidence interval in part (e) is likely to understate the true amount of uncertainty associated with the estimate of the population ATE.

Because **the three samples are unlikely to be random draws from the same population**: the three of them correspond to different moments (and locations: primary election in Michigan, in 2006; municipal elections in small cities and towns in Michigan, in 2007; and special congressional election in Illinois, in 2009). Our overall population is the population of both states for the period covering the three studies (2006 to 2009), and we are assuming that the underlying parameters that govern cause and effect do not change over time (and within locations).

---

4. FE 11.10

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

```r
# Replicate the table in page 381
ps5_no4_table <- data.frame(Duration = c(rep("Sec01_10", 2), rep("Sec11_20", 2),
                                                      rep("Sec21_30", 2), rep("Sec31_40", 2)),
                                     Treated = rep(c(0, 1), 4),
                                 N = c(143, 187, 619, 784, 1132, 983, 2012, 2032),
                                 Turnout = c(17.5, 16.6, 18.3, 17.4, 18.9, 19.7, 19.8,
                                                24.3), stringsAsFactors = F)
# Create the dataframe based on the table in page 381
ps5_no4 <- data.frame(Duration = unlist(mapply(rep, ps5_no4_table$Duration,
                                                                    ps5_no4_table$N)),
                                      Treated = unlist(mapply(rep, ps5_no4_table$Treated,
                                                                    ps5_no4_table$N)))
ps5_no4$Voted = unlist(mapply(function(n, turnout) {
    n1 = round(n * turnout/100)
    return(c(rep(0, n-n1), rep(1, n1)))
    }, ps5_no4_table$N, ps5_no4_table$Turnout))
# Check that we've correctly replicated the table in page 381
table(ps5_no4[, c(2,3,1)])
```

```
## , , Duration = Sec01_10
##
##        Voted
## Treated    0    1
##       0  118   25
##       1  156   31
##
## , , Duration = Sec11_20
##
##        Voted
## Treated    0    1
##       0  506  113
##       1  648  136
##
## , , Duration = Sec21_30
##
```

```
##        Voted
## Treated    0    1
##       0  918  214
##       1  789  194
##
## , , Duration = Sec31_40
##
##        Voted
## Treated    0    1
##       0 1614  398
##       1 1538  494
```

```
pander(ftable(addmargins(table(ps5_no4), FUN = list(Total = sum),
                                 quiet = TRUE)))
```

		"Voted"	"0"	"1"	"Total"
"Duration"	"Treated"				
"Sec01_10"	"0"		118	25	143
	"1"		156	31	187
	"Total"		274	56	330
"Sec11_20"	"0"		506	113	619
	"1"		648	136	784
	"Total"		1154	249	1403
"Sec21_30"	"0"		918	214	1132
	"1"		789	194	983
	"Total"		1707	408	2115
"Sec31_40"	"0"		1614	398	2012
	"1"		1538	494	2032
	"Total"		3152	892	4044
"Total"	"0"		3156	750	3906
	"1"		3131	855	3986
	"Total"		6287	1605	7892

```
ps5_no4_table_freq <- prop.table(table(ps5_no4[, c(2,3,1)]), c(1, 3))
ps5_no4_table_freq <- matrix(ps5_no4_table_freq, ncol = 2, byrow = TRUE)
ps5_no4_table_freq <- ps5_no4_table_freq[seq(2, 8, by = 2), ]
rownames(ps5_no4_table_freq) <- c("Sec01_10", "Sec11_20", "Sec21_30", "Sec31_40")
colnames(ps5_no4_table_freq) <- c("Control", "Treatment")
round(100*t(ps5_no4_table_freq)[c(2, 1), ], 1)
```

```
##           Sec01_10 Sec11_20 Sec21_30 Sec31_40
## Treatment     16.6     17.3     19.7     24.3
## Control       17.5     18.3     18.9     19.8
```

```
kable(100*t(ps5_no4_table_freq)[c(2, 1), ], digits = 1)
```

	Sec01_10	Sec11_20	Sec21_30	Sec31_40
Treatment	16.6	17.3	19.7	24.3
Control	17.5	18.3	18.9	19.8

```
RSEs(lm(Voted ~ Treated, ps5_no4))
```

```
##
```

```
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.19201    0.00630   30.46   <2e-16 ***
## Treated      0.02249    0.00906    2.48    0.013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Without worrying about the duration of the call, the apparent average effect of the script that encouraged voting was a 2.2(0.9) percentage point increase in voting turnout.

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 voter turnout"? Why or why not?

We begin by plotting turnout rate by experimental condition and call duration:

```
ps5_no4_Turnout <- data.frame(Turnout = c(ps5_no4_table_freq[, 1],
                                          ps5_no4_table_freq[, 2]),
                              Mean_Duration = rep(c(5, 15, 25, 35), 2),
                              Group = c(rep("Control", 4), rep("Treatment", 4)))
library(ggplot2)
scatter <- ggplot(ps5_no4_Turnout, aes(Mean_Duration, Turnout, colour = Group))
scatter + geom_point() + geom_smooth(method = "lm", aes(fill = Group),
                                     alpha = 0.3) +
    xlab("Call duration") + ylab("Turnout rate")
```

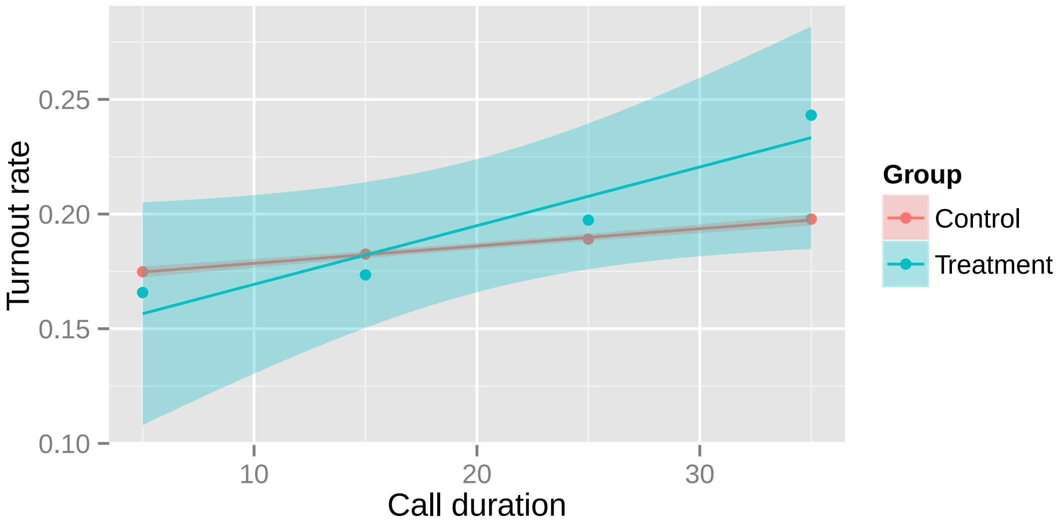

Figure 3: Scatterplot of Turnout Rate against Call Duration by experimental condition

As the plot suggests, the longer a person listens to a recorded message that encourages voting, the likelier is that he or she votes (the slope of the corresponding line is more pronounced than the one for the placebo group)… but only a correlation exists—there is no evidence of a causal effect, since subjects decide how long they want to listen to the recorded message (self-selection bias); it might be the case than those who have already decided to vote are more willing to listen to that kind of message.

5. FE 12.3

a. One part of this experiment focuses on whether the treatment influences which projects villages select.

i. Describe the experimental subjects. What units are assigned to treatment versus control? What is the treatment?

The experimental subjects were inhabitants of 49 Indonesian villages from 3 subdistricts located in different parts of rural Indonesia (which represent the wide variety of conditions in rural Indonesia). Villages are made up of 2 to 7 hamlets, which are naturally occurring clusters of between 25 and 250 households.
Subjects were clustered on a village level (i.e., the experimental units were villages).

ii. What units are assigned to treatment versus control?

The experiment was conducted in 2 phases. First, Phase I was conducted in 10 villages in East Java Province and 19 villages in North Sumatra Province. Based on qualitative reports from Phase I areas, the experimental protocol was changed slightly, and then run again in Phase II in an additional 20 villages in Southeast Sulawesi Province.
Randomization was conducted on the province level—In Phase I, 25% of villages (5 out of the 19 villages in North Sumatra Province, and 3 out of the 10 villages in East Java Province) were allocated to treatment, whereas in Phase II, 45% of villages (9 out of the 20 villages in Southeast Sulawesi Province) were allocated to the treatment.
The sample appears balanced across 26 covariates.

ii. What is the treatment?

A change in the decision-making mechanism: instead of following a meeting-based process[1] described previously, some villages were randomly allocated to choose their projects via a direct election-based plebiscite. The idea behind the plebiscite was that it would move the political process from a potentially elite-dominated meeting to a more participatory process that might be less subject to elite capture. The method for selecting the list of projects to be chosen (i.e., the agenda-setting procedure) was the same in both cases—the list of projects to be decided on at the meeting or the list of projects on the ballot was determined from the results of hamlet-level meetings, where each hamlet was allowed to nominate one general project and one women's project.
The plebiscite was conducted as follows. Two paper ballots were prepared—one for the general project and one for the women's project. The ballots had a picture of each project, along with a description of the project. Village officials distributed voting cards to all adults in the village who had been eligible to vote in national

[1] For a period of several months, a village facilitator organizes small meetings at the hamlet level; for large hamlets, multiple meetings might be held in different neighborhoods within each hamlet. These meetings aim to create a list of ideas for what projects the village should propose. These ideas are then divided into two groups—those that originated from women's-only meetings and those suggested by mixed meetings or men's meetings. The village facilitator presents the women's list to a women-only village meeting and the men's and joint ideas to a village meeting open to both genders. Although these meetings are open to the public, those that attend represent a highly selected sample. In particular, government officials (e.g., the village head, village secretary, and other members of the village executive), neighborhood heads, and those selected to represent village groups compose the majority of attendees. A typical meeting would have between 9 and 15 people representing the various hamlets, as well as various formal and informal village leaders, with on average about 48 people attending in total out of an average village population of 2,200. In the general meeting, the representatives are usually (but not always) men, whereas in the women's meeting, all representatives are women. At each meeting, the representatives in attendance discuss the proposals, with substantial help from an external facilitator, deciding ultimately on a single proposal from each meeting.

parliamentary elections held approximately six-months previously. The voting cards also indicated the date of the election and the voting place. Voting places were set up in each hamlet in the village. When arriving at the voting place to vote, men received one ballot (for the general project) and women received two ballots (one for the general project, one for the women's project). The selected project (for both the general and women's projects) was the proposal that received a plurality of the votes in the respective vote.

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 findings?

In that case, the effect of novelty would increase the satisfaction that participating and being part of the decision-making process bring to villagers, hence increasing the estimated effect of the treatment.

c. The level of satisfaction is measured by survey responses.

i. Suppose that the researcher calling the subjects knows the subjects' treatment assignment. How might this bias the results?

By subtly trying to obtain a response from the subjects that confirms the researcher's prior hypothesis about the treatment effect.
E.g., if the researcher expected the treatment to have the positive effect it seems it had (villagers who were assigned to the plebiscite 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 wishes), he or she might slightly change the questions asked to the two kinds of villagers, so the ones in the plebiscite group view the selection even fairer, whereas the ones in the control (meeetings') group view the selection even less fair. In this case the estimated effect would be larger than it really is.

ii. Suppose that the subjects consider the externally subsidized development project to have been a gift, and believe the interviewer to be associated with this gift/subsidy. How might this bias the results?

By giving the interviewer the answer that they think he or she expects. E.g., villagers in the control group would report the selection as being fairer than they think it is, so the interviewer does not think they are being ungrateful. In this case the estimated effect would be smaller than it really is.

iii. Differential attrition is an acute concern with survey data. Give an example of why differential attrition might occur here and how it might bias the results.

Differential attrition might occur in this example because those in the control group may be dissatisfied with the usual decision-making mechanism, and hence not willing to talk about the topic. If attrition rate decreases with satisfaction level (i.e., if those more sastisfied are more likely to complete the survey—and hence less likely to attrit) within the control group, missingness would be dependent on potential outcomes, and hence the results would be biased (under this example scenario, the responses from those less satisfied with the usual decision-making mechanism would be missing, so the estimated effect would be smaller than it really is).

6. FE 12.5

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

The evidence regarding the effect of plate size on what people is not very convincing, mainly because the short-term effect may compensated over time (*intertemporal substitution*). We would need to study subjects over a longer period, to check whether those given regular-sized plates ate more in their next meal and vice versa.

The evidence regarding the effect on what people weigh is even less convincing, for the reasons mentioned above, and because the outcome that was measured is not a perfect proxy for weight—or it's not the only possible mediator. Maybe people that ate more because of a larger plate also exercised more afterwards.

b. What design and measurement improvements do you suggest?

I would suggest tracking outcomes for every participant over time, and even performing within-subjects experiments for each of them (i.e., introducing a stimulus on outcomes in subsequent time periods; but the *no-persistence* assumption might be easily broken in this scenario).

An additional improvement would be analyzing participants in a more natural setting (like their own homes)—people may act differently when they are invited to a dinner than they usually do at home (e.g., participants may find rude either asking for a second portion or not finishing what's on their plate).

And, of course, participants' weight should be measured if we want to draw conclusions about possible effects on that outcome.

---

7. Natural experiment in medicine

a. What are the benefits of this study relative to a randomized controlled trial?

As mentioned in the synopsis, that this study "*was 20 times larger and had a much longer follow-up than any prior comparative-effectiveness RCT [. . .]. This permitted the investigators to obtain outcomes on low-frequency events like mortality; prior RCTs principally only had power to examine blood glucose control. The study [. . .] examined more than* 80, 000 *patients for up to 10 years.*"

RCTs frequently use small samples, over relatively short periods, which makes some infrequent outcomes unmeasurable, having to use mediators instead.

b. What are the disadvantages of this study relative to a randomized controlled trial?

The main disadvantage is that the assignment to each experimental condition (TZDs or SUs) is based on a not entirely random process. Thus, there may be idyosincratic differences between the two groups that are being compared, and if those are related to the outcome of interest (risk of avoidable hospitalization and risk of death), we may not be making an apples-to-apples comparison. The best we can do—actually, what the authors did—is checking whether the instrumental variable seems to act as a good randomizer, by analyzing if it's uncorrelated with the outcomes (i.e., if both the outcomes and the instrumental variable are not caused by an unobserved effect); but that does not provide full evidence, just a high chance that the hypothesis of no correlation is correct.

c. This is a natural experiment rather than a deliberate research experiment. Therefore, practice telling a story, consistent with the reported data, about how there might be no causal difference at all between the drugs.

According to the study, relative to TZDs, SUs cause a 68% increase in risk of avoidable hospitalization and a 50% increase in risk of death—i.e., TZDs are a less risky drug to treat type 2 diabetes when initial treatment fails to control blood-sugar level, and should be favored by clinicians.

But the assignment to one treatment or the other was based on *physician prescribing patterns*, which are thought to be a near-random process. What if prescribing patterns were correlated with the outcomes, or any unobservable factor that affects them?

Say physicians tend to prescribe SUs to patients with a poorer condition—because they think TZDs are less effective, or may have less side effects—, or simply that those physicians who prescribe SUs more frequently have, just by chance, patients in worse conditions. In that case, those subjects would be more likely to be hospitalized or even die—as it actually happened—, but just because both those outcomes and the prescribing pattern were caused[2] or at least correlated with [3] patients' condition. We might obtain results like those reported in the study, but we could not claim that SUs were the cause of the increase in risks of avoidable hospitalization or death.

d. Describe the placebo test mentioned in the article. Does this test help to rule out the story you just told? Why or why not?

In observational studies, researchers usually check if there is a balance of observable factors across values of the independent variable, in order to confirm that groups assigned to each treatment are comparable. In other words, that they're making an apple-to-apples comparison. But comparing every possible factor is not plausible—yes, both groups may be round, equal-sized and red... but there may be unobserved differences, so they are actually comparing Galas to Fujis.

I like the way the author of the synopsis expressed the findings: he employs phrases such as "*If we observe such balance, that should* ***increase our confidence*** *that there is also balance among unobservable factors too.*" But that sentence in bold letters is crucial: that would help us gain confidence, but would not provide 100% certainty that describing patterns (the instrument) are uncorrelated with any potential cause of the outcomes under study.

In order to confirm that the instrumental variable was a "good randomizer," "*Prentice's team did more.*" "*The authors examined 2 populations that bracketed the study population in disease severity and did not receive the treatment under study [...]. The same potential bias related to causal factors that are unobservable to the researcher (if there are any) is as likely to apply to these 2 populations as to the primary sample. [...] For neither of these "bracketing" populations [...] was prescribing pattern related to outcomes. This provides strong support for prescribing pattern as a randomizer. It's consistent with the assumption that prescribing pattern only affects outcomes through its effect on treatment with SU or TZD, just as one would desire of an RCT's coin flip.*"

According to the synopsis, the 2 populations used in the placebo tests "*bracketed the study population in disease severity*", and Prentice's team "*showed balance in demographic, diagnoses, and provider quality variables*" (apart from balance in the outcomes). But were those 2 populations 100% comparable to the study population? No, they were broader (and hence bracketed it)—overall, there were no significant differences in the outcomes, for each value of the instrument, but maybe there was some uncaptured heterogeneity. Let's suppose that the study population is a sample of the 2 populations used in the placebo tests, based on an unobserved factor (e.g., the study population has different levels of an unobserved factor related to the

[2] If physicians think SUs are more effective than TZDs.

[3] If physicians that prescribe SUs more frequently have patients in worse conditions.

outcomes than the mean levels in the 2 bracketing populations). In that case, the story I wrote in part (c) would be less likely, but we could not rule it out.

To make the counterexample very simple, let's suppose that outcomes are equally balanced across the instrument in the 2 populations used in the placeo tests. E.g., death rate is 10% among those prescribed—but not treated with—SUs, and also 10% among those prescribed—but not treated with—TZDs). Say both populations are equally splitted into patients with high levels of X (whose death rate is 12%) and patients with low levels of X (whose death rate is 8%). Both groups are comparable from any other point of view. If the study population is similar to a sample from both populations, but one where prescribing patterns are correlated with the level of that unobserved factor, X, an outcome difference would appear, caused by the different levels of X rather than the treatment.

e. What do you think about the prospects for such observational research in medicine? Is this kind of research a complement to, or a substitute for, deliberate field experiments?

I think that such observational research in medicine is really helpful, and should be conducted more… as a **complement** to field experiments. These studies cannot substitute them, but can be very helpful to narrow the hypotheses that subsequent field experiments aim to confirm.

---

8. RD question

a. Summarize the study and its conclusion. Which study did you pick? What is the outcome? What is the "treatment"? What is the discontinuity? What is the conclusion?

The study I picked was from Clemens & Tiongson (2012). They "*use a language test in the Philippines that permitted workers to travel to Korea, in order to see impacts on household investments.*"

The treatment is an overseas work, i.e., "*migration from the Philippines to temporary jobs in Korea*" (more especifically, the study defines treatment as "*a household in which any member ever worked in Korea*[4]").

The outcomes are numerous: e.g., expenditures in food, quality of life, education & health, and durables, savings, investings, borrowing (for business or non-business reasons, and from family or others).

The policy discontinuity that resulted in quasi-random assignment of treatment (temporary, partial-household migration to high-wage jobs in Korea) was that "*workers in the Philippines who applied to high-wage temporary jobs in Korea between 2005 and 2007*" were "*required to pass a basic test in the Korean language, and all those who failed were unable to take the job.*" Hence, the researchers "*surveyed the households of those who barely passed for comparison to the households of those who barely failed.*"

The main conclusion is that "*migration from the Philippines to temporary jobs in Korea has important effects on migrants' households*[5].

[4] "*Defining treatment in this way, rather than as current presence of a household member in Korea, prevents self-selected return migration from being a source of endogenous treatment.*"

[5] Other conclusions are that:

- Those effects are "*different from the effects of remittances, as remittances are just one portion of the bundled treatment that is migration."Migration has no significant effect on household entrepreneurial activity, since migration is a different treatment: It both puts remittances into the household and takes potential entrepreneurs out of the household.*"
- "*The model predicts that in unitary households, migration affects investment behavior solely by raising earnings—increasing self-finance and alleviating any borrowing constraints. In collective households, there are two additional channels: migration can alter the balance of power in household decision-making and can alter the technology of home*

b. A throwback to Week 2. Assume the RD was not available, and someone did a simple observational study comparing individuals who happen to get treatment versus those who do not for non-random reasons. What's an alternative story for a simple association between the outcome and this non-random version of the treatment like this? Try to be as concrete as possible in your storytelling.

Let's suppose the Korean language test was not required, and hence the RD was not available. Filipino workers who got the treatment—i.e., who worked in Korea—may differ from those who did not for several reasons besides the treatment, that might affect their households.

For example, they spent more in education (and health; $p = 0.00772$), and their children were more likely to study in a private school ($p = 0.0328$) and also to be awarded ($p < 0.0370$), because—despite sharing background attributes with their counterparts, such as education level[6]—those Filipino workers were more aware of the possible effects of education in their children, and also instilled in them that idea. If Y is the treatment (working in Korea), Z is any of the three outcomes above, and X is that belief in the advantages of education, it is not Y but X which causes Z (Y may be another effect of X, or not).

c. What makes this RD evidence so convincing, relative to an observational study like in part (b)?

In short, that it creates a situation quite similar to a field experiment, making plausible to be able to infer causality, not just a mere correlation.

A Regression Discontinuity makes the treatment to be assigned discontinuously at a point, on some variable (the test score, in this case) that is not random, and thus treatmentassignment becomes a near-random process—the main requirement for an experiment. Individuals usually decide whether they work abroad or not—i.e., they self-select—, but the situation in the Philippines made it possible that assignment was almost as decided by a flip coin: a huge difference in the test score is certainly not due to chance, but a small difference (near the cutoff, or any other point) may be attributed to chance (in general, we would not consider a small difference being caused by intelligence or preparation, but just "noise").

production." Authors found that "*the most important of these are far and away the financial effects, suggesting that the simplicity of the unitary household model is adequate to explain the most important economic impacts of migration in this setting. While migration causes large changes in how household decisions are made, these changes explain almost none of the important impacts of migration on spending, borrowing, or investment. There is suggestive but statistically imprecise evidence that the collective household is relevant: migration does appear to alter the technology of home production, perhaps by drawing breadwinners out of farming, though this evidence is statistically imprecise.*"

- "*No evidence*" was found "*to support any effect of migration on labor force participation by the spouses or other family members of migrants. The model provides potential explanations for this result: the predicted effect of migration on others' labor force participation is smaller when borrowing constraints are smaller. Households in the sample borrow extensively and the increase in income accompanying migration causes them to borrow less, not more. They may therefore face small borrowing constraints, though we do not have direct evidence of this.*"
- "*The above findings are compatible with the households surveyed being credit-constrained human capital investors: Migration causes greater private schooling for children, more awards at school, and greater household expenditure on health and education.*"
- *The findings are not broadly compatible with these households being credit-constrained physical capital investors: Migration has no significant effects on entrepreneurial activity-except perhaps drawing some families' breadwinners out of farming. It does not raise savings, but causes borrowing to markedly decrease.*"
- What the research cannot answer, for instance, is "*how the effect depends purely on the gender of the migrant,*" or what are the effects "*on other households—households from which no member applied to an EPS-Korea job—,*" or "*on return migrants,*" and what is "*the effect of strategic decisions made prior to migration caused by foresight of the future option to migrate.*"

[6] Background attributes are usually checked in observational studies, as if they ensured an apple-to-apples comparison.

d. What's an alternative story for why this RD pattern might exist even if the causal effect did not exist? This story might be hard to tell because these are good studies, but do your best.

When Filipino workers took the Korean language test, they surely were aware of the cut point—i.e., the minimum score to pass; for simplicity, let's say 100 correct answers out of 200 questions. Supposed that almost half of the questions (say 96) were relatively easy, while the remaining (104) were hard, anyone having barely prepared for the test might end up with a score close to the cut point, but only those who were best-prepared would obtain a score above that cut point. It could be said that the latter were most interested in a high-wage job in Korea, for reasons that might also affect their households; for example, they were more responsible or hard-workers. Even if an overseas work had no effect on the outcomes being measured, those who passed the test would have ended up obtaining a better job or earning more more money (and/or they would have invested more, and borrowed less) for the same reasons that made them study more for the test.

In other words, were test-takers able to self-select across the RD?

Authors' arguments against that can be found in Subsection 4.3[7]. Basically, they tested for *sorting*, by checking covariate balance on either side of the cut point, as well as smoothness, and confirmed both (see Figure 4). That does not entirely discard the story I wrote (e.g., it's not possible to check all possible covariates), but makes it much less plausible.

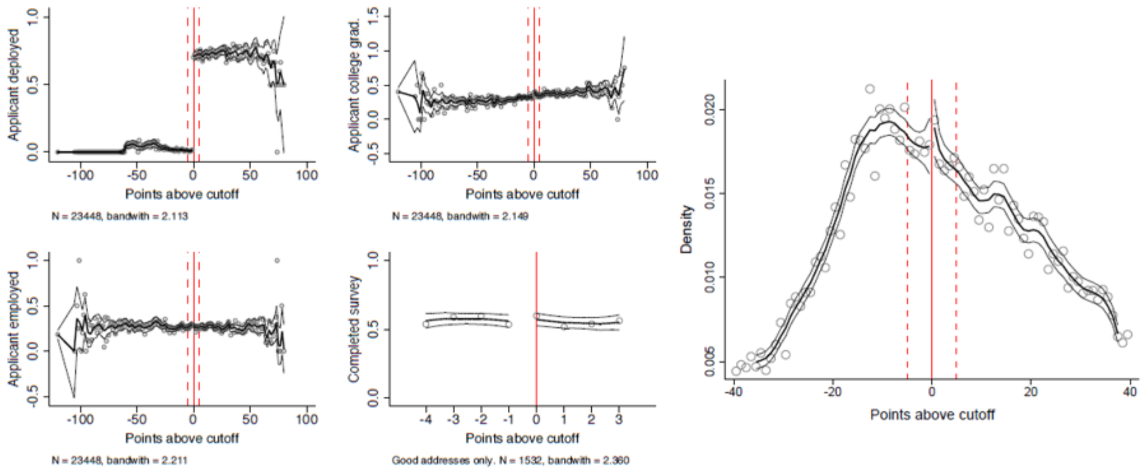

Figure 4: Discontinuities in sampling universe, and McCrary nonparametric test for score manipulation

[7] Excerpt from the paper:

1. "*There is no statistically significant jump in test-score density at the passing threshold. [. . .] While there is an increase in the density at the passing threshold, it is small in magnitude, not statistically significant, and well within the observed variance in score density at nearby levels. This is reassuring but does not per se rule out self-selection.*"
2. "*In all of the analysis to follow, the test score we use is exclusively the test score from each worker's first attempt to pass the KLT. A small number of failers re-took the test in later rounds, and if we were to use scores from subsequent attempts, this would raise the possibility of workers self-selecting across the passing-score cutoff.*"
3. "*The test was administered and scored by a Korean institution andwe are not aware of any substantial reports of corruption or other irregularities in scoring or record-keeping.*"