Juanjo Carin

W241 (Field Experiments) – Problem Set #3 – MIDS Spring 2015

March 2, 2015

Contents

W241 – Problem Set #3

1. Broockman and Green

a. Using regression without clustered standard errors (that is, ignoring the clustered assignment), compute a confidence interval for the effect of the ad on candidate name recognition in Study 1 only.

i. Note: Ignore the blocking the article mentions throughout this problem.

```r
# Read the dataset
BG <- read.csv("broockman_green_anon_pooled_fb_users_only.csv", header = TRUE)

# We focus on Study 1
Study1 <- BG[BG$studyno == 1, ]
# And regress Recall on Treatment
reg_Study1 <- lm(name_recall ~ treat_ad, data = Study1)
summary(reg_Study1)
```

```
##
## Call:
## lm(formula = name_recall ~ treat_ad, data = Study1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1825 -0.1825 -0.1727 -0.1727  0.8273
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.182469   0.016142  11.304   <2e-16 ***
## treat_ad    -0.009798   0.021012  -0.466    0.641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3817 on 1362 degrees of freedom
## Multiple R-squared:  0.0001596,  Adjusted R-squared:  -0.0005745
## F-statistic: 0.2174 on 1 and 1362 DF,  p-value: 0.6411
```

```
# A function to calculate RSEs
RSEs <- function(model){
    require(sandwich, quietly = TRUE)
    require(lmtest, quietly = TRUE)
    newSE <- vcovHC(model)
    coeftest(model, newSE)
    }
# I'll use Robust SEs instead of the standard ones (given by lm)
    # for this Exercise and also Exercise 5 (not for Exercises 2 to 4, because
    # the authors of that Study did not use them, either)

(coef_reg_Study1 <- RSEs(reg_Study1))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1824687  0.0163651 11.1499   <2e-16 ***
## treat_ad    -0.0097979  0.0211120 -0.4641   0.6427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# I'll also use 1.96 rather than 2
    CI_marginerror <- function(alpha) qnorm(1-alpha/2)
    CI_halfwidth <- CI_marginerror(0.05)
(CI_lower_Study1 <- coef_reg_Study1[2, 1] - CI_halfwidth*coef_reg_Study1[2, 2])
```

```
## [1] -0.05117671
```

```
(CI_upper_Study1 <- coef_reg_Study1[2, 1] + CI_halfwidth*coef_reg_Study1[2, 2])
```

```
## [1] 0.03158093
```

So the confidence interval for the effect of the ad on candidate name recognition in Study 1 is **[-0.051, 0.032]**. It crosses zero, which means that it is not statistically significant ($p = 0.64$).

And of course the coefficients of the regression make sense: there were 1364 participants in Study 1; 102 of the 559 participants who were not treated—i.e., who were in the Control group—recalled the participant (i.e., a 18.2%, which is the value of the intercept); 139 of the 805 participants who were treated recalled the participant (i.e., a 17.3%, which is $\beta_0 + \beta_1 = 0.1824687 + -0.0097979$). See the tables below—especially the 2nd column—for more detailed information about the number and percentage of participants in each condition.)

```
BG_table <- table(BG$treat_ad, BG$name_recall, BG$studyno)
dimnames(BG_table)[[1]] <- c("Control", "Treatment")
dimnames(BG_table)[[2]] <- c("Recall = NO", "Recall = YES")
dimnames(BG_table)[[3]] <- c("Study 1", "Study 2")
BG_table
```

```
## , ,  = Study 1
##
##
##              Recall = NO Recall = YES
##   Control            457          102
##   Treatment          666          139
##
## , ,  = Study 2
##
##
##              Recall = NO Recall = YES
##   Control            395          607
##   Treatment          133          202
```

```
(BG_table_freq <- round(100*prop.table(BG_table, c(1,3)), 1))
```

```
## , ,  = Study 1
##
##
##              Recall = NO Recall = YES
##   Control           81.8         18.2
##   Treatment         82.7         17.3
##
## , ,  = Study 2
##
##
##              Recall = NO Recall = YES
##   Control           39.4         60.6
##   Treatment         39.7         60.3
```

b. What are the clusters in Broockman and Green's study? Why might taking clustering into account increase the standard errors?

In Study 1, clusters were formed by unique combinations of age, gender, and location (e.g., 24 year old males in San Francisco). In Study 2—unlike Study 1, in which the towns were widely dispersed—, the Congressional district of the collaborating candidate "included some areas that were more densely populated", so "to minimize misclassification of subjects' treatment status (e.g., a control user logging into Facebook from a treatment location)" clusters were formed by assembling "groups of zip codes that fell within county boundaries and targeted the ads on the basis of these zip code groups. The clusters in Study 2 thus comprised contiguous groups of zip codes."

Clustering might increase the standard errors because it involves more sampling variability than complete random assignment (especially if subjects within the same cluster tend to share similar potential outcomes, and there are a few clusters with large numbers of subjects).

c. Now repeat part (a), but taking clustering into account. That is, compute a confidence interval for the effect of the ad on candidate name recognition in Study 1, but now correctly accounting for the clustered nature of the treatment assignment.

```
# A function to calculate CSEs
cl <- function(fm, cluster){
    require(sandwich, quietly = TRUE)
    require(lmtest, quietly = TRUE)
    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)
    }

# Remove missing data
BG <- BG[complete.cases(BG), ]
    # An alternative way
    # BG <- na.omit(BG)

# We still focus on Study 1
cluster_Study1 <- BG[BG$studyno == 1, ]
# So we need the number of clusters in it
cluster_Study1$cluster <- factor(cluster_Study1$cluster)

reg_cluster_Study1 <- lm(name_recall ~ treat_ad, data = cluster_Study1)

(coef_reg_cluster_Study1 <- cl(reg_cluster_Study1, cluster_Study1$cluster))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1824687  0.0184915  9.8677   <2e-16 ***
## treat_ad    -0.0097979  0.0237536 -0.4125   0.6801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(CI_lower_cluster_Study1 <- coef_reg_cluster_Study1[2, 1] -
    CI_halfwidth*coef_reg_cluster_Study1[2, 2])
```

```
## [1] -0.05635414
```

```
(CI_upper_cluster_Study1 <- coef_reg_cluster_Study1[2, 1] +
    CI_halfwidth*coef_reg_cluster_Study1[2, 2])
```

```
## [1] 0.03675836
```

Considering clustering, the confidence interval for the effect of the ad on candidate name recognition in Study 1 is **[-0.056, 0.037]**. As expected (for the reasons explained in part 1.b), it is even wider than before.

d. Repeat part (c), but now for Study 2 only.

```
# Now we focus on Study 2
cluster_Study2 <- BG[BG$studyno == 2, ]
# So we need the number of clusters in it
cluster_Study2$cluster <- factor(cluster_Study2$cluster)

reg_cluster_Study2 <- lm(name_recall ~ treat_ad, data = cluster_Study2)
(coef_reg_cluster_Study2 <- cl(reg_cluster_Study2, cluster_Study2$cluster))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6057884  0.0181889  33.305   <2e-16 ***
## treat_ad    -0.0028033  0.0355033  -0.079   0.9371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(CI_lower_Study2 <- coef_reg_cluster_Study2[2, 1] -
    CI_halfwidth*coef_reg_cluster_Study2[2, 2])
```

```
## [1] -0.07238862
```

```
(CI_upper_Study2 <- coef_reg_cluster_Study2[2, 1] +
    CI_halfwidth*coef_reg_cluster_Study2[2, 2])
```

```
## [1] 0.06678192
```

Considering clustering, the confidence interval for the effect of the ad on candidate name recognition in Study 2 is [**-0.072, 0.067**].

e. Repeat part (c), but using the entire sample from both studies. Do not take into account which study the data is from (more on this in a moment), but just pool the data and run one omnibus regression. What is the treatment effect estimate and associated *p*-value?

```
BG$cluster <- factor(BG$cluster)
reg_cluster_entire <- lm(name_recall ~ treat_ad, data = BG)
(coef_reg_cluster_entire <- cl(reg_cluster_entire, BG$cluster))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.454196   0.018576 24.4504 < 2.2e-16 ***
## treat_ad    -0.155073   0.026730 -5.8014 7.344e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(CI_lower_cluster_entire <- coef_reg_cluster_entire[2, 1] -
    CI_halfwidth*coef_reg_cluster_entire[2, 2])
```

```
## [1] -0.207464
```

```
(CI_upper_cluster_entire <- coef_reg_cluster_entire[2, 1] +
    CI_halfwidth*coef_reg_cluster_entire[2, 2])
```

```
## [1] -0.1026824
```

Considering clustering, the confidence interval for the effect of the ad on candidate name recognition in both Studies is **[-0.207, -0.103]**.

The treatment effect estimate is **-0.155**, and the associated p-value is **p = 7.34e-09** (highly statistically significant).

See the tables below (2nd column) for more information:

```
BG_table_entire <- table(BG$treat_ad, BG$name_recall)
dimnames(BG_table_entire)[[1]] <- c("Control", "Treatment")
dimnames(BG_table_entire)[[2]] <- c("Recall = NO", "Recall = YES")
BG_table_entire
```

```
##
##             Recall = NO Recall = YES
##   Control           852          709
##   Treatment         799          341
```

```
(BG_table_entire_freq <- round(100*prop.table(BG_table_entire, 1), 1))
```

```
##
##             Recall = NO Recall = YES
##   Control          54.6         45.4
##   Treatment        70.1         29.9
```

The treatment effect estimate is the difference between the percentage of treated participants that recalled the candidate (29.9%) and the percentage of untreated participants that recalled the candidate (45.4%), in the whole sample.

f. Now, repeat part (e) but include a dummy variable (a 0/1 binary variable) for whether the data are from Study 1 or Study 2. What is the treatment effect estimate and associated p-value?

```
BG$study2 <- ifelse(BG$studyno == 2, 1, 0)
reg_cluster_entire2 <- lm(name_recall ~ treat_ad + study2, data = BG)
(coef_reg_cluster_entire2 <- cl(reg_cluster_entire2, BG$cluster))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1806848  0.0169702 10.6472   <2e-16 ***
## treat_ad    -0.0067752  0.0204154 -0.3319     0.74
## study2       0.4260988  0.0206970 20.5875   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(CI_lower_cluster_entire2 <- coef_reg_cluster_entire2[2, 1] -
    CI_halfwidth*coef_reg_cluster_entire2[2, 2])
```

```
## [1] -0.04678874
```

```r
(CI_upper_cluster_entire2 <- coef_reg_cluster_entire2[2, 1] +
    CI_halfwidth*coef_reg_cluster_entire2[2, 2])
```

```
## [1] 0.03323824
```

Considering clustering, and each Study (i.e., Block), the confidence interval for the effect of the ad on candidate name recognition is **[-0.047, 0.033]**.

The treatment effect estimate now is **-0.0068**, and the associated p-value is **p = 0.74**.

I.e., this time the treatment effect estimate is not statistically significant (and the confidence interval crosses zero), as it happened for each Study (Block) separately.

g. Why did the results from parts (e) and (f) differ? Which result is biased, and why? (Hint: see pages 75-76 of Gerber and Green, with more detailed discussion optionally available on pages 116-121.)

The result from part (e) is biased because the probability of being assigned to the treatment varies from one Study to the other (i.e., from block to block).

```r
BG_table_treatment <- table(BG$treat_ad, BG$studyno)
dimnames(BG_table_treatment)[[1]] <- c("Control", "Treatment")
dimnames(BG_table_treatment)[[2]] <- c("Study 1", "Study 2")
BG_table_treatment
```

```
##
##             Study 1 Study 2
##   Control       559    1002
##   Treatment     805     335
```

```r
(BG_table_treatment_freq <- round(100*prop.table(BG_table_treatment, 2), 1))
```

```
##
##             Study 1 Study 2
##   Control      41.0    74.9
##   Treatment    59.0    25.1
```

The probability of being assigned to the treatment, as shown above, is 59% in Study 1 and 25.1% in Study 2.

So participants in Study 2 were less likely to be assigned to the Treatment group. But they (whether assigned to Control or Treatment group) also recalled the candidate much more than participants in Study 1.

```r
BG_table_freq
```

```
## , ,  = Study 1
##
##
##             Recall = NO Recall = YES
##   Control          81.8         18.2
##   Treatment        82.7         17.3
##
## , ,  = Study 2
##
##
##             Recall = NO Recall = YES
##   Control          39.4         60.6
##   Treatment        39.7         60.3
```

More than 60% of the participants in Study 2 recalled the candidate (no matter which group they were assigned to), while about 18% of the participants in Study 1 did recall him.

That's why, even when the treatment effect is small and not statistically significant, when we pool together both Studies, the percentage of participants that recalled the candidate is lower in the Treatment group than in the Control group, so we get—in part (e)—an estimate much lower (-0.1550732) than it really is (-0.0067752).

Another way to calculate the unbiased estimated of the ATE that was calculated in part (f) would be using equation (3.10) of GG (page 73):

```r
t <- table(BG$studyno)
names(t) <- c("Study 1", "Study 2")
t
```

```
## Study 1 Study 2
##    1364    1337
```

```r
cat((t[1]*coef_reg_cluster_Study1[2, 1] +
        t[2]*coef_reg_cluster_Study2[2, 1]) / sum(t))
```

```
## -0.006335577
```

h. Skim this Facebook case study and consider two claims they make reprinted below. Why might their results differ from Broockman and Green's? Please be specific and provide examples.

i. "There was a 19 percent difference in the way people voted in areas where Facebook Ads ran versus areas where the ads did not run."

ii. "In the areas where the ads ran, people with the most online ad exposure were 17 percent more likely to vote against the proposition than those with the least."

My answer is mainly based on the following excerpts from the study:

- "*The agency relied on Facebook's Location Targeting to reach people in two of the most populated counties in Florida, Dade and Broward, which have a combined population of 4.2 million. It chose to focus the impressions here because it wanted to be able to benchmark voter results against the rest of the state.*"
- "*The agency also used Facebook to target people who liked politically oriented Facebook Pages or who listed relevant Likes & Interests or Education & Work. For example, it targeted people who listed terms like "teacher," "pta" "math teacher" to reach educators. Because both the polling and Facebook research indicated that the issue carries special resonance with parents of school children, it even included interests like "I love my son" and "I love my daughter" (and layered them with demographic targeting.)*"

Both paragraphs make clear that the study is not really a randomized field experiment. Neither the clusters were randomly assigned to the Treatment group (they deliberately chose two counties in Florida) nor the Facebook users with the most online ad exposure were, on average, similar to those with the least—the former were "politically active" and had interests related to the campaign, so they might have voted against the proposition even in the absence of the treatment. So the authors were not really comparing apples to apples.

---

2. Recycling in Peru – Intensity

a. In Column 3 of Table 4A, what is the estimated ATE of providing a recycling bin on the average weight of recyclables turned in per household per week, during the six-week treatment period? Provide a 95% confidence interval.

The estimated ATE of providing a recycling bin is **0.187**, and based on its SE (0.032), the 95% confidence would be **[0.124, 0.25]** (Note: using 1.96 rather than 2).

b. In Column 3 of Table 4A, what is the estimated ATE of sending a text message reminder on the average weight of recyclables turned in per household per week? Provide a 95% confidence interval.

The estimated ATE of providing a recycling bin is **−0.024**, and based on its SE (0.039), the 95% confidence would be **[-0.1, 0.052]**.

c. Which outcome measures in Table 4A show statistically significant effects (at the 5% level) of providing a recycling bin?

All of them except the one in the 5th column—i.e.,

- Percentage of visits turned in bag
- Average number of bins turned in per week
- Average weight (in kg) of recyclables turned in per week
- Average market value of recyclables given per week

but not Average percentage of contamination per week.

d. Which outcome measures in Table 4A show statistically significant effects (at the 5% level) of sending text messages?

None of them.

e. Suppose that, during the two weeks before treatment, household A turns in 2kg per week more recyclables than household B does, and suppose that both households are otherwise identical (including being in the same treatment group). From the model, how much more recycling do we predict household A to have than household B, per week, during the four weeks of the experiment? Provide only a point estimate, as the confidence interval would be a bit complicated. This question is designed to test your understanding of slope coefficients in regression.

Based on the equation (3) of the article ($Y_i = \beta_1 B + \beta_2 S + \lambda Ybl_i + P_i + \alpha_j + \varepsilon_1$), and supposed that both households are identical in every way except for Ybl_i, given that $\hat{\alpha} = 0.281$ we have:

$Y_A - Y_B = \hat{\alpha} \times (Ybl_A - Ybl_B) = 0.281 \times 2$. That is, **0.562 kg more**.

f. Suppose that the variable “percentage of visits turned in bag, baseline” had been left out of the regression reported in Column 1. What would you expect to happen to the results on providing a recycling bin? Would you expect an increase or decrease in the estimated ATE? Would you expect an increase or decrease in the standard error? Explain your reasoning.

It would depend on the allocation of subjects to treatment: sometimes the estimated ATE would be higher, sometimes it would be lower—but on average the estimate would be equal to what it is when the covariate “percentage of visits turned in bag, baseline” is included in the regression. Without the data set (which we are still not supposed to have in this Exercise), there is no way we can know.

And the standard error would always increase if we leave the covariate (the pre-test outcome, in this case) out of the regression, because the sampling variability would also increase.

> Let’s see it with a simple example (in which the treatment effect is almost the same for every subject):

```
N <- 200
# Number of participants in an experiment
    # That aims to test if miracle pills help keeping same weight after
    # Christmas. Numbers used are not important.
X <- round(runif(N, 100, 200), 1)
# Weight in pounds before Christmas
    # (Uniformly distributed between 100 and 200 pounds)
Y <- round(X + 10 + rnorm(N)/10, 1)
# Weight in pounds after Christmas in the absence of treatment
    # Everybody gains about 10 pounds (10 + some random noise)
Z <- rep(c(rep(0, (N/2)), rep(1, (N/2))))
Z <- sample(Z)
# Half of the participants are randomly assigned to treatment (miracle pills)
Y_exp <- Y
Y_exp[Z == 1] <- round(Y_exp[Z == 1] - 10 + rnorm(N/2)/10, 1)
# Miracle pills help subjects keep their previous weight
    # (I.e., lose about 10 pounds, what they would have gained otherwise)
head(X, 10)
```

```
##  [1] 138.7 131.2 168.0 164.2 192.5 174.4 145.3 106.2 148.7 144.0
```

```
# First 10 values of pre-test outcomes (weight before Christmas)
head(Y_exp, 10)
```

```
##  [1] 148.9 131.3 178.0 164.2 202.5 184.4 155.4 116.3 158.6 154.1
```

```
# First 10 values of post-test outcomes (weight after Christmas)
    # About the same for treated subjects
    # About 10 pounds higher for untreated subjects
summary(lm(Y_exp ~ Z))$coefficients
```

```
##             Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  157.025   2.751541 57.068024 6.795702e-125
## Z             -7.121   3.891266 -1.829996  6.875422e-02
```

```
# If we regress on the treatment the ATE estimate depends on the randomization
    # Sometimes below -10, sometimes above
summary(lm(Y_exp ~ Z + X))$coefficients
```

```
##              Estimate    Std. Error  t value      Pr(>|t|)
## (Intercept)  9.971221 0.0484155105  205.951 5.628270e-232
## Z          -10.001916 0.0174571803 -572.940 2.432483e-319
## X            1.000318 0.0003184878 3140.837  0.000000e+00
```

```
# If we add the pre-test outcome as a covariate the ATE estimate is about -10
summary(lm((Y_exp - X) ~ Z))$coefficients
```

```
##             Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)   10.018 0.01232698  812.6886  0.000000e+00
## Z            -10.001 0.01743299 -573.6825 7.588848e-321
```

```
# The same happens if we use difference-in-difference
    # the ATE estimate is about -10

num_iter <- 100e3
# Randomization inference
estimate_treatment_effect <- function(Y, X, Z){
    Y_exp <- Y
    Z <- sample(Z)
    # A new randomization
    Y_exp[Z == 1] <- round(Y_exp[Z == 1] - 10 + rnorm(N/2)/10, 1)
    c(wo_pretest = summary(lm(Y_exp ~ Z))$coefficients[2, 1:2],
      w_pretest  = summary(lm(Y_exp ~ Z + X))$coefficients[2, 1:2],
      diff_in_diff = summary(lm((Y_exp - X) ~ Z))$coefficients[2, 1:2])
    # Estimate and SE:
        # withouth the covariate
        # with the covariate
        # using difference-in-differences
    }
ATE <- as.data.frame(replicate(num_iter, estimate_treatment_effect(Y, X, Z),
                                          simplify = FALSE))
ATE <- as.data.frame(t(ATE))
rownames(ATE) <- c(1:num_iter)
colnames(ATE)[c(2, 4, 6)] <- c("wo_pretest.SE", "w_pretest.SE",
                                          "diff_in_diff.SE")
# Calculate each of the 3 estimates in n different allocations
head(ATE, 20)
```

```
##    wo_pretest.Estimate wo_pretest.SE w_pretest.Estimate w_pretest.SE
## 1              -13.988      3.885303         -10.001785   0.01714935
## 2               -8.405      3.894556         -10.001328   0.01700993
## 3              -10.800      3.895462         -10.001904   0.01688849
## 4               -5.694      3.883253          -9.973709   0.01762372
## 5               -7.424      3.890666         -10.031779   0.01720087
## 6              -12.568      3.891980         -10.013475   0.01783704
## 7              -10.368      3.897296          -9.985806   0.01752648
## 8               -8.373      3.893313          -9.988833   0.01736581
## 9               -9.130      3.895379          -9.988099   0.01737829
## 10              -8.857      3.894596          -9.976999   0.01832831
## 11              -4.842      3.878256          -9.994120   0.01784775
## 12              -6.820      3.888269          -9.985483   0.01711005
## 13             -11.935      3.893356         -9.988801   0.01866255
## 14              -3.080      3.865022        -10.019696   0.01731606
## 15              -9.929      3.895052         -9.980995   0.01715341
## 16              -8.334      3.893942        -10.002142   0.01827470
## 17              -6.210      3.885755         -9.993648   0.01766800
## 18             -11.262      3.893890        -10.014164   0.01753660
## 19              -7.236      3.891630         -9.970750   0.01794412
## 20             -13.074      3.888332        -10.004759   0.01731850
##    diff_in_diff.Estimate diff_in_diff.SE
## 1                -10.002      0.01706198
## 2                -10.001      0.01697860
## 3                -10.002      0.01685050
## 4                 -9.974      0.01752747
## 5                -10.032      0.01714113
## 6                -10.014      0.01779059
## 7                 -9.986      0.01759362
## 8                 -9.989      0.01731905
## 9                 -9.988      0.01733799
## 10                -9.977      0.01827815
## 11                -9.994      0.01772403
## 12                -9.986      0.01705013
## 13                -9.989      0.01860786
## 14               -10.018      0.01715938
## 15                -9.981      0.01711370
## 16               -10.002      0.01822308
## 17                -9.994      0.01758500
## 18               -10.014      0.01749517
## 19                -9.970      0.01790830
## 20               -10.004      0.01727438
```

```r
apply(ATE, 2, mean)
```

```
##   wo_pretest.Estimate         wo_pretest.SE    w_pretest.Estimate
##           -9.97743980            3.88592324           -9.99998905
##          w_pretest.SE diff_in_diff.Estimate       diff_in_diff.SE
##            0.01762996           -9.99998982            0.01755965
```

```r
# On average, the ATE without considering the pre-test outcome as covariate
    # is the expected one

# Plot results
df <- data.frame(
    Estimate = c(ATE$wo_pretest.Estimate, ATE$w_pretest.Estimate),
    Analysis = c(rep("Without Pre-test", num_iter),
                 rep("With Pre-test", num_iter)))
require(ggplot2)
ggplot(df, aes(x = Estimate, fill = Analysis)) +
    geom_histogram(binwidth = 0.5, position = 'identity') +
    scale_x_continuous(limits = c(-20, 0))
```

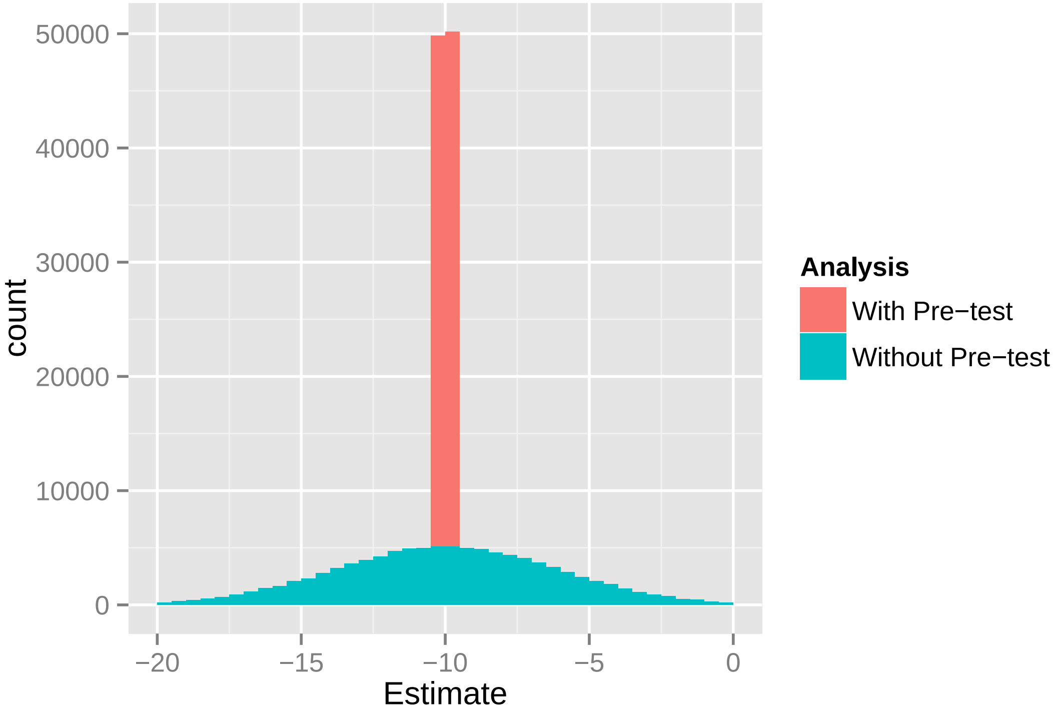

Figure 1: Example: Histogram of 2 ATE distributions, depending on whether a covariate (pre-test outcome) is used in the regression

g. In column 1 of Table 4A, would you say the variable "has cell phone" is a bad control? Explain your reasoning.

I think "has cell phone" is **not** a bad control. It is certainly correlated with the treatment "any SMS message", but it is measured before it is applied (and not affected by it, but the other way around).

h. If we were to remove the "has cell phone" variable from the regression, what would you expect to happen to the coefficient on "Any SMS message"? Would it go up or down? Explain your reasoning.

Since both variables are positively correlated, I would say **the coefficient of "Any SMS message" would go up if we removed the "has cell phone" variable from the regression**, because the effect of that variable on the outcome would now be included in "Any SMS message".

I replicated a simpler version of this experiment and this is what I found:

In the following page a table with the possible combinations of experimental conditions is shown:

Has cell phone	Any SMS message	Any bin
1	0	0
1	0	1
1	1	0
1	1	1
0	0	0
0	0	1
~~0~~	~~1~~	~~0~~
~~0~~	~~1~~	~~1~~

Only the first 6 cases are possible—the last 2 ones (not having a phone but receiving any SMS message) are not.

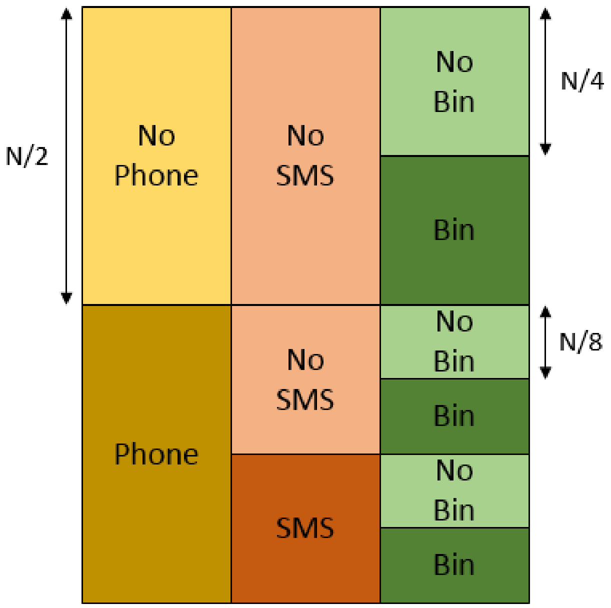

Figure 2: Split of experimental conditions

```r
N <- 200
# 200 participants
recycling <- data.frame(pretest = round(runif(N, 5, 10) + rnorm(N)/10, 1),
                        phone = c(rep(0, N/2), rep(1, N/2)),
                        sms = c(rep(0, 3*N/4), rep(1, N/4)),
                        bin = c(rep(0, N/4), rep(1, N/4),
                                rep(0, N/8), rep(1, N/8),
                                rep(0, N/8), rep(1, N/8)))
# Suppose the average weight (in kg) of recyclables turned in per week are
    # uniformly distributed between 5 and 10 (plus some noise)
# Half of the participants (N/2) have a phone
# Half of those who have a phone (N/4) receive a SMS message
# Half of the participants (N/2: equally splitted between those who have a
    # phone and those who do not) get a bin
recycling[recycling$phone == 1, ]$pretest <-
    recycling[recycling$phone == 1, ]$pretest + 2
# Suppose the average weight of recyclables turned in per week by those who
    # have a phone is 2 kg higher
mean(recycling$pretest[recycling$phone==0])
```

```
## [1] 7.761
```

```
mean(recycling$pretest[recycling$phone==1])
```

```
## [1] 9.282
```

```
recycling$outcome_notreatment <- recycling$pretest + round(rnorm(N)/10, 1)
# Add some extra-noise to outcomes (in the absence of treatment)

randomization <- function(){
    c(c(rep(0, N/2), sample(c(rep(0, N/4), rep(1, N/4)))),
      sample(c(rep(0, N/4), rep(1, N/4))),
      sample(c(rep(0, N/8), rep(1, N/8))),
      sample(c(rep(0, N/8), rep(1, N/8))))
    }
# A function to randomize blocks

temp <- randomization()
recycling$sms <- temp[1:N]
recycling$bin[recycling$phone == 0] <- temp[(N+1):(N+N/2)]
recycling$bin[recycling$phone == 1 & recycling$sms == 0] <-
    temp[(N+N/2+1):(N+3*N/4)]
recycling$bin[recycling$phone == 1 & recycling$sms == 1] <-
    temp[(N+3*N/4+1):(2*N)]

sum(recycling$bin[recycling$sms == 1])
```

```
## [1] 25
```

```
sum(recycling$bin[recycling$sms == 0 & recycling$phone == 1])
```

```
## [1] 25
```

```
sum(recycling$bin[recycling$sms == 0 & recycling$phone == 0])
```

```
## [1] 50
```

```
# The N/2 subjects who are given a bin are distributed among the other
    # experimental condition

recycling$outcome <- recycling$outcome_notreatment
recycling[recycling$bin == 1, ]$outcome <-
    recycling[recycling$bin == 1, ]$outcome + 4
recycling[recycling$sms == 1, ]$outcome <-
    recycling[recycling$sms == 1, ]$outcome + 2
# Suppose being given a bin increases average weight of recyclabes by 4
# Suppose receiving an SMS message increases average weight of recyclabes by 2

(wo_pretest_wo_phone <- summary(lm(outcome ~ bin + sms,
                                    data = recycling))$coefficients)
```

```
##             Estimate Std. Error  t value      Pr(>|t|)
## (Intercept)    8.294  0.1740640 47.64914 4.371748e-110
## bin            3.892  0.2279033 17.07742  1.010227e-40
## sms            3.084  0.2631600 11.71911  2.074811e-24
```

```
(wo_pretest_w_phone <- summary(lm(outcome ~ bin + sms + phone,
                                             data = recycling))$coefficients)
```

```
##             Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)    7.804  0.1834153 42.548249 6.060487e-101
## bin            3.892  0.2117897 18.376716  1.655969e-44
## sms            2.104  0.2995159  7.024668  3.451169e-11
## phone          1.470  0.2593884  5.667177  5.130296e-08
```

```
# (For this randomization) It seems the effect of SMS is splitted between the
    # SMS itself and the phone when the latter is included in the regression

(w_pretest_wo_phone <- summary(lm(outcome ~ pretest + bin + sms,
                                             data = recycling))$coefficients)
```

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) -0.09388273 0.04168883  -2.251988  2.543214e-02
## pretest      1.00987853 0.00481981 209.526606 1.757273e-232
## bin          3.99803725 0.01524111 262.319218 1.531816e-251
## sms          2.00073697 0.01833330 109.131284 1.849574e-177
```

```
(w_pretest_w_phone <- summary(lm(outcome ~ pretest + bin + sms + phone,
                                            data = recycling))$coefficients)
```

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) -0.10615049 0.04276246  -2.482329  1.389759e-02
## pretest      1.01236968 0.00520676 194.433703 3.217674e-225
## bin          3.99829882 0.01522035 262.694311 1.327343e-250
## sms          2.01491147 0.02151581  93.647942 4.635649e-164
## phone       -0.02527002 0.02015396  -1.253849  2.113981e-01
```

```
# (For this randomization) It seems the effect of phone disappears when
    # pre-test values are considered (because treatments do not affect the
    # phone variable)
# Including pre-test values also has the effect of reducing the SE of the rest
    # of the estimates

# Let's check what happens on average (not only for a particular realization)
num_iter <- 100e3

estimate_treatment_effect <- function(df){
    colnames(df) <- c("pretest", "phone", "sms", "bin", "outcome_notreatment",
                        "outcome")
    temp <- randomization()
    df$sms <- temp[1:N]
    df$bin[df$phone == 0] <- temp[(N+1):(N+N/2)]
    df$bin[df$phone == 1 & df$sms == 0] <- temp[(N+N/2+1):(N+3*N/4)]
    df$bin[df$phone == 1 & df$sms == 1] <- temp[(N+3*N/4+1):(2*N)]
    # A new randomization
    df[, 6] <- df[, 5]
    df[df[, 4] == 1, 6] <- df[df[, 4] == 1, 6] + 4
    df[df[, 3] == 1, 6] <- df[df[, 3] == 1, 6] + 2
    df$outcome <- df$outcome_notreatment
    df[df$bin == 1, ]$outcome <- df[df$bin == 1, ]$outcome + 4
    df[df$sms == 1, ]$outcome <- df[df$sms == 1, ]$outcome + 2
```

```r
    wo_pretest_wo_phone <- summary(lm(outcome ~ bin + sms,
                                                 data = df))$coefficients
    wo_pretest_w_phone = summary(lm(outcome ~ bin + sms + phone,
                                              data = df))$coefficients
    w_pretest_wo_phone = summary(lm(outcome ~ pretest + bin + sms,
                                              data = df))$coefficients
    w_pretest_w_phone = summary(lm(outcome ~ pretest + bin + sms + phone,
                                             data = df))$coefficients
    c(sms_wo_pretest_wo_phone = wo_pretest_wo_phone[3, 1],
      sms_wo_pretest_w_phone = wo_pretest_w_phone[3, 1],
      phone_wo_pretest_w_phone = wo_pretest_w_phone[4, 1],
      sms_w_pretest_wo_phone = w_pretest_wo_phone[4, 1],
      sms_w_pretest_w_phone = w_pretest_w_phone[4, 1],
      phone_w_pretest_w_phone = w_pretest_w_phone[5, 1])
    }

ATE <- as.data.frame(replicate(num_iter, estimate_treatment_effect(recycling),
                                         simplify = FALSE))
ATE <- as.data.frame(t(ATE))
rownames(ATE) <- c(1:num_iter)
apply(ATE, 2, mean)
```

```
##  sms_wo_pretest_wo_phone   sms_wo_pretest_w_phone phone_wo_pretest_w_phone
##               3.01478733               2.00018100               1.52190950
##   sms_w_pretest_wo_phone    sms_w_pretest_w_phone  phone_w_pretest_w_phone
##               1.98980873               1.99997991              -0.01797485
```

This simple example confirms that the coefficient of "Any SMS message" goes up when we don't consider the "has cell phone variable". The effect is obvious when we don't include the baseline (pre-test outcomes) in the regression, and not so much when we do: because "has cell phone" is measured before the experiment, and it is not affected by the treatment, its effect is included in the baseline, so when both variables are used in the regression the coefficient estimate of "has cell phone" is close to zero... but still removing that variable causes the coefficient of "Any SMS message" to go up (though just a little bit: actually it goes up about the value of the coefficient of "has cell phone"). Though the baseline is included in the paper, the coefficient of "has cell phone" is not so close to zero in Table 4A, but that may be due to some facts that I've not considered in my example (which is a rough simplification of the real experiment)[1].

---

3. Recycling in Peru – Multifactor

a. What is the full experimental design for this experiment? Tell us the dimensions, such as 2x2x3. (Hint: the full results appear in Panel 4B.)

This is a **3×(3+1)** design. There are 2 treatment conditions:

1. recycling bins, which has 3 categories:

   1.1. No bin

[1] For example, I did not consider what is mentioned in the paper: "*We estimate this equation separately among households that did not have a cell phone and those that had a cell phone but did not receive a SMS message, in which case P drops out of the equation and interaction effects are not a concern. We also estimate the equation on the full sample.*""

  1.2. Bin without sticker
  1.3. Bin with sticker
2. SMS message, with 4 categories:
  2.1. No phone
  2.2. Phone – No SMS
  2.3. Phone – Generic SMS
  2.4. Phone – Personal SMS

Though "no phone" is actually a covariate, I consider it in the experimental design because it imposes the participants whom the SMS treatment can be applied.

b. In the results of Table 4B, describe the baseline category. That is, in English, how would you describe the attributes of the group of people for whom all dummy variables are equal to zero?

The baseline category consists of people who were not treated in any way, i.e., they didn't receive a recycling bin (with or without sticker), and (because they didn't have a phone—the baseline value for the variable "has cell phone" is also zero) they didn't receive an SMS message of any kind.

c. In column (1) of Table 4B, interpret the magnitude of the coefficient on "bin without sticker." What does it mean?

Column 1 corresponds to "Percentage of visits turned in bag" (which indicates the proportion of weeks in which the household had an opportunity to turn in a bag or bin of recyclables and in which they actually did so). Since "bin without sticker" is a binary/dummy variable, its coefficient has the same magnitude as the outcome ("Percentage of visits turned in bag"), and it indicates the average percentage increase (+3.5%) in that outcome for those who were given a "bin without sticker."

d. In column (1) of Table 4B, which seems to have a stronger treatment effect, the recycling bin with message sticker, or the recycling bin without sticker? How large is the magnitude of the estimated difference?

The recycling bin with message sticker. Its treatment effect estimate is bigger: 5.5% compared to 3.5%, so the difference is **2%**—all other dummy variables equal to zero, the proportion of weeks those who were given a bin with message sticker turned in a bag or bin of recyclables increases by 2 percentage points (from the baseline, 37.4%, to 39.4%).

e. Is this difference you just described statistically significant? Explain which piece of information in the table allows you to answer this question.

Each treatment effect is statiscally significant ("bin with sticker" at the 1% level, "bin without sticker" at the 5% level)... but the difference between them is not: **F-test *p*-value (1) = (2) = 0.31** (see 1st column, 12th row of the table).

f. Notice that Table 4C is described as results from “fully saturated” models. What does this mean? Looking at the list of variables in the table, explain in what sense the model is “saturated.”

A “fully saturated model” means that we consider the possible treatment-by-treatment interactions, i.e., the fact that the effect of one treatment (e.g., begin given a recyclinb bin) is not fixed but depends on whether another treatment (e.g., having received an SMS message) has taken place.

The model is saturated in the sense that it includes $3 \times (3 + 1) = 12$ coefficients, one for each possible combination of the experimental conditions (the intercept or baseline category—all dummy variables equal to zero—appear in the 2nd page of Table 4C). Another way to build a fully saturated model would have been by using the 3 variables, and the 2 possible interaction terms (for a $3 \times (3 + 1)$ design this alternative model would usually include 2 variables and 1 interaction term, but keep in mind the SMS effect implies 2 variables—“has cell phone” and “type of SMS”—rather than just 1).

4. Recycling in Peru – Bins turned

a. For simplicity, let’s start by measuring the effect of providing a recycling bin, ignoring the SMS message treatment (and ignoring whether there was a sticker on the bin or not). Run a regression of Y on only the bin treatment dummy, so you estimate a simple difference in means. Provide a 95% confidence interval for the treatment effect.

```r
library(foreign)
Recycling <- read.dta("karlan_data_subset_for_class.dta")
head(Recycling)
```

```
##   street havecell avg_bins_treat base_avg_bins_treat bin sms bin_s bin_g
## 1      7        1      1.0416666               0.750   1   1     1     0
## 2      7        1      0.0000000               0.000   0   1     0     0
## 3      7        1      0.7500000               0.500   0   0     0     0
## 4      7        1      0.5416667               0.500   0   0     0     0
## 5      6        1      0.9583333               0.375   1   0     0     1
## 6      8        0      0.2083333               0.000   1   0     0     1
##   sms_p sms_g
## 1     0     1
## 2     1     0
## 3     0     0
## 4     0     0
## 5     0     0
## 6     0     0
```

```r
library(pastecs)
round(stat.desc(Recycling, desc = F), digits = 2)
```

```
##              street havecell avg_bins_treat base_avg_bins_treat  bin  sms
## nbr.val        1782     1784        1785.00             1785.00 1785 1785
## nbr.null          0      730          52.00              182.00 1182 1234
## nbr.na            3        1           0.00                0.00    0    0
## min            -999        0           0.00                0.00    0    0
## max             263        1           4.17                6.38    1    1
## range      1262        1           4.17                6.38    1    1
## sum      122612     1054        1215.73             1314.38  603  551
##          bin_s bin_g sms_p sms_g
## nbr.val   1785  1785  1785  1785
## nbr.null  1485  1482  1507  1512
## nbr.na       0     0     0     0
## min          0     0     0     0
## max          1     1     1     1
## range        1     1     1     1
## sum        300   303   278   273
```

```
# "street" variable has 3 NAs (see above)
# "havecell" has 1 NA (see above)
sum(na.omit(Recycling$street == -999))
```

```
## [1] 120
```

```
# 120 "street" cells have -999 value

model_4a <- lm(avg_bins_treat ~ bin, data = Recycling)
# Based on the last 2 parts (g & h) of this exercise, it seems that the authors
    # did not use RSEs, so I will not use them either.
summary(model_4a)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.6353497 0.01179325 53.873993 0.000000e+00
## bin         0.1353800 0.02029056  6.672067 3.355794e-11
```

```
t.test(Recycling[Recycling$bin == 1, ]$avg_bins_treat,
         Recycling[Recycling$bin == 0, ]$avg_bins_treat)
```

```
##
##  Welch Two Sample t-test
##
## data:  Recycling[Recycling$bin == 1, ]$avg_bins_treat and Recycling[Recycling$bin == 0, ]$avg_bins_treat
## t = 6.5061, df = 1132.467, p-value = 1.156e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.09455292 0.17620707
## sample estimates:
## mean of x mean of y
## 0.7707297 0.6353497
```

```
# Of course, almost same results

(CI_lower_4a <- summary(model_4a)$coefficients[2, 1] -
    CI_halfwidth*summary(model_4a)$coefficients[2, 2])
```

```
## [1] 0.09561122
```

```
(CI_upper_4a <- summary(model_4a)$coefficients[2, 1] +
    CI_halfwidth*summary(model_4a)$coefficients[2, 2])
```

```
## [1] 0.1751488
```

The estimate for the effect of providing a recycling bin is **0.135** (with a p-value $p = 3.36e - 11$. The 95% confidence interval is **[0.096, 0.175]**.

b. Now add the pre-treatment value of Y as a covariate. Provide a 95% confidence interval for the treatment effect. Explain how and why this confidence interval differs from the previous one.

```
model_4b <- update(model_4a, . ~ . + base_avg_bins_treat)
summary(model_4b)$coefficients
```

```
##                      Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)         0.3496026 0.01373121 25.460434 3.229065e-122
## bin                 0.1246930 0.01666714  7.481365  1.148786e-13
## base_avg_bins_treat 0.3929647 0.01338618 29.356002 7.517988e-155
```

```
(CI_lower_4b <- summary(model_4b)$coefficients[2, 1] -
    CI_halfwidth*summary(model_4b)$coefficients[2, 2])
```

```
## [1] 0.09202598
```

```
(CI_upper_4b <- summary(model_4b)$coefficients[2, 1] +
    CI_halfwidth*summary(model_4b)$coefficients[2, 2])
```

```
## [1] 0.15736
```

The estimate for the effect of providing a recycling bin is now **0.125** (lower than it previously was, but with an even lower p-value: $p = 1.15e - 13$. The 95% confidence interval is now **[0.092, 0.157]**.

Including a covariate so informative (see how R^2 increases from the previous model to this one: from 0.024 to 0.342) as the pre-test values reduces uncertainty and hence tightens the confidence interval.

c. Now add the street fixed effects. (You'll need to use the R command factor().) Provide a 95% confidence interval for the treatment effect.

```
# First we omit the NA values of "street"
Recycling <- Recycling[!is.na(Recycling$street), ]

model_4c <- update(model_4b, . ~ . + factor(street))
summary(model_4c)$coefficients[1:3, ]
```

```
##                      Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)         0.3677440 0.03161561 11.631723  4.442673e-30
## bin                 0.1138868 0.01705784  6.676508  3.364562e-11
## base_avg_bins_treat 0.3737068 0.01432809 26.082099 2.936935e-125
```

```
(CI_lower_4c <- summary(model_4c)$coefficients[2, 1] -
    CI_halfwidth*summary(model_4c)$coefficients[2, 2])
```

```
## [1] 0.08045404
```

```r
(CI_upper_4c <- summary(model_4c)$coefficients[2, 1] +
    CI_halfwidth*summary(model_4c)$coefficients[2, 2])
```

```
## [1] 0.1473195
```

The estimate for the effect of providing a recycling bin is now **0.114** (once again lower than it previously was, and this time with a slightly bigger p-value: $p = 3.36e - 11$. The 95% confidence interval is now **[0.08, 0.147]**. It is just slightly wider, because of the increase of (the p-value and) the standard error (as well as the decrease of the estimate).

d. Recall that the authors described their experiment as “stratified at the street level,” which is a synonym for blocking by street. Explain why the confidence interval with fixed effects does not differ much from the previous one.

Because the outcome do not differ too much between blocks (strata). Blocking only improves precision when it involves grouping observations with similar potential outcomes—if the variability within blocks is not much lower than it is between blocks (as it seems to happen in this case), blocking does not help us reduce the uncertainty.

e. Perhaps having a cell phone helps explain the level of recycling behavior. Instead of “has cell phone,” we find it easier to interpret the coefficient if we define the variable " no cell phone." Give the R command to define this new variable, which equals one minus the “has cell phone” variable in the authors’ data set. Use “no cell phone” instead of “has cell phone” in subsequent regressions with this dataset.

Supposed that the dataframe where we have loaded the data is called “Recycling”, the command could be, for instance:

`Recycling$no_cell_phone <- 1 - Recycling$havecell`

```r
# First we omit the NA value of "havecell"
Recycling <- Recycling[!is.na(Recycling$havecell), ]

Recycling$no_cell_phone <- 1 - Recycling$havecell
# Recycling$no_cell_phone <- as.numeric(Recycling$havecell == 0)
    # Alternative way
```

f. Now add “no cell phone” as a covariate to the previous regression. Provide a 95% confidence interval for the treatment effect. Explain why this confidence interval does not differ much from the previous one.

```r
model_4f <- update(model_4c, . ~ . - factor(street) + no_cell_phone +
                             factor(street))
# I take "street" out of the model and then include it again, but after
    # "no cell phone", so its coefficients appear at the end
summary(model_4f)$coefficients[1:4, ]
```

```
##                         Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)           0.38749355 0.03225539 12.013296  6.836102e-32
## bin                   0.11510074 0.01704496  6.752774  2.023950e-11
## base_avg_bins_treat   0.37338230 0.01429941 26.111720 1.805396e-125
## no_cell_phone        -0.04950989 0.01686377 -2.935873  3.373650e-03
```

```
(CI_lower_4f <- summary(model_4f)$coefficients[2, 1] -
    CI_halfwidth*summary(model_4f)$coefficients[2, 2])
```

```
## [1] 0.08169324
```

```
(CI_upper_4f <- summary(model_4f)$coefficients[2, 1] +
    CI_halfwidth*summary(model_4f)$coefficients[2, 2])
```

```
## [1] 0.1485082
```

The estimate for the effect of providing a recycling bin is now **0.115**, almost equal to the previous one (as well as its p-value, which is now $p = 2.02e - 11$. The 95% confidence interval is also pretty similar: [**0.082, 0.149**].

The covariate "no cell phone" helps explain the level of recycling behavior (its coefficient is statistically significant)…

```
t.test(Recycling[Recycling$no_cell_phone == 0, ]$avg_bins_treat,
         Recycling[Recycling$no_cell_phone == 1, ]$avg_bins_treat)
```

```
##
##  Welch Two Sample t-test
##
## data:  Recycling[Recycling$no_cell_phone == 0, ]$avg_bins_treat and Recycling[Recycling$no_cell_phone == 1, ]$av
## t = 2.6772, df = 1625.834, p-value = 0.007498
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01390935 0.09013759
## sample estimates:
## mean of x mean of y
## 0.7020120 0.6499886
```

…but the effect of providing a recycling bin is the same (i.e., it is homogeneous) among both groups (those who have a cell phone and those who don't), that is why adding the "no cell phone" covariate helps explain better the level of recycling behavior (R^2 increases from 0.436 to 0.439), but does not reduce the uncertainty about the effect of providing a recycling bin.

```
by(data = Recycling$avg_bins_treat,
    INDICES = Recycling[, c("no_cell_phone", "bin")],  FUN = mean)
```

```
## no_cell_phone: 0
## bin: 0
## [1] 0.66
## ---------------------------------------------------------
## no_cell_phone: 1
## bin: 0
## [1] 0.6001751
## ---------------------------------------------------------
## no_cell_phone: 0
## bin: 1
## [1] 0.7873679
## --------------------------------------------------------
## no_cell_phone: 1
## bin: 1
## [1] 0.7437088
```

```
t.test(Recycling[Recycling$no_cell_phone == 0 &
                        Recycling$bin == 1, ]$avg_bins_treat,
         Recycling[Recycling$no_cell_phone == 1 &
                        Recycling$bin == 1, ]$avg_bins_treat)
```

```
##
##  Welch Two Sample t-test
##
## data:  Recycling[Recycling$no_cell_phone == 0 & Recycling$bin == 1,  and Recycling[Recycling$no_cell_phone == 1
## t = 1.2458, df = 533.263, p-value = 0.2134
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02518626  0.11250444
## sample estimates:
## mean of x mean of y
## 0.7873679 0.7437088
```

g. Now let's add in the SMS treatment. Re-run the previous regression with "any SMS" included. You should get the same results as in Table 4A. Provide a 95% confidence interval for the treatment effect of the recycling bin. Explain why this confidence interval does not differ much from the previous one.

```
model_4g <- lm(avg_bins_treat ~ bin + sms + no_cell_phone +
                        base_avg_bins_treat + factor(street), data = Recycling)
round(summary(model_4g)$coefficients[2:5, 1:2], 3)
```

```
##                     Estimate Std. Error
## bin                    0.115      0.017
## sms                    0.005      0.021
## no_cell_phone         -0.047      0.020
## base_avg_bins_treat    0.373      0.014
```

```
(CI_lower_4g <- summary(model_4g)$coefficients[2, 1] -
    CI_halfwidth*summary(model_4g)$coefficients[2, 2])
```

```
## [1] 0.08163421
```

```
(CI_upper_4g <- summary(model_4g)$coefficients[2, 1] +
    CI_halfwidth*summary(model_4g)$coefficients[2, 2])
```

```
## [1] 0.1484731
```

The estimate for the effect of providing a recycling bin is now **0.115**, almost equal to the previous one (as well as its p-value, which is now $p = 2.1e - 11$. The 95% confidence interval is also pretty similar: **[0.082, 0.148]**.

The estimates of the coefficients (and their associated standard errors) are equal to the ones that appear in Table 4A, with the only exception of the "no cell phone variable": because this binary variable is the opposite of "has cell phone" (the former is true when the latter is false, and vice versa), the estimate we got is the same than the one that appears in Table 4A, but with the opposite sign.

The covariate "sms" does not even help explain the level of recycling behavior…

```r
t.test(Recycling[Recycling$sms == 0, ]$avg_bins_treat,
         Recycling[Recycling$sms == 1, ]$avg_bins_treat)
```

```
##
##  Welch Two Sample t-test
##
## data:  Recycling[Recycling$sms == 0, ]$avg_bins_treat and Recycling[Recycling$sms == 1, ]$avg_bins_treat
## t = 0.8085, df = 1169.411, p-value = 0.419
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02315587  0.05561801
## sample estimates:
## mean of x mean of y
## 0.6857211 0.6694900
```

…so the effect of providing a recycling bin is not affected by that covariate, and hence it is homogeneous among both groups: those who recevie an SMS message and those who don't.

h. Now reproduce the results of column 2 in Table 4B, estimating separate treatment effects for the two types of SMS treatments and the two types of recycling-bin treatments. Provide a 95% confidence interval for the effect of the unadorned recycling bin. Explain how your answer differs from that in part (g), and explain why you think it differs.

```r
model_4h <- lm(avg_bins_treat ~ bin_s + bin_g + sms_p + sms_g + no_cell_phone +
                        base_avg_bins_treat + factor(street), data = Recycling)
round(summary(model_4h)$coefficients[2:7, 1:2], 3)
```

```
##                     Estimate Std. Error
## bin_s                  0.128      0.022
## bin_g                  0.103      0.022
## sms_p                 -0.008      0.025
## sms_g                  0.020      0.025
## no_cell_phone         -0.046      0.020
## base_avg_bins_treat    0.374      0.014
```

```r
(CI_lower_4h <- summary(model_4h)$coefficients[3, 1] -
    CI_halfwidth*summary(model_4h)$coefficients[3, 2])
```

```
## [1] 0.06028885
```

```r
(CI_upper_4h <- summary(model_4h)$coefficients[3, 1] +
    CI_halfwidth*summary(model_4h)$coefficients[3, 2])
```

```
## [1] 0.1460916
```

The estimate for the effect of providing an **unadorned** recycling bin is **0.103** (and the associated p-value is $p = 2.64e - 06$. The 95% confidence interval is [**0.06, 0.146**].

The estimates of the coefficients (and their associated standard errors) are equal to the ones that appear in Table 4B, with the only exception of the "no cell phone variable": because this binary variable is the opposite of "has cell phone" (the former is true when the latter is false, and vice versa), the estimate we got is the same than the one that appears in Table 4B, but with the opposite sign.

The effect of providing a recycling bin that we estimated in part (g) is somewhere between the effect of providing a generic one and the effect of providing one with a sticker.
The standard errrors of the estimates of these two "refined" effects are larger (and hence the confidence intervals are wider) than the SE of the estimate of the effect of "any bin". Because we are now making an additional comparison / contrast (bin vs. no bin plus bin with sticker vs. bin without sticker), we are less certain about the results of each comparison (and that is why the standard error is larger).
This increase in the standard error can also be explained as the need for a lower significance level: because now we have 3 rather than 2 experimental conditions, we're performing $\binom{3}{2} = 3$ t-tests, and the familywise error rate would increase from 0.05 to $1 - (0.95)^3 = 0$. In other words, to compensate for that increase in the standard error, we whould decrease the significance level from 0.05 to $\frac{0.05}{3} = 0.0167$.

---

5. Fictional scenario – ZMapp and Ebola

a. Without using any covariates, answer this question with regression: What is the estimated effect of ZMapp (with standard error in parentheses) on whether someone was vomiting on day 14? What is the p-value associated with this estimate?

```
ebola <- read.csv("ebola_rct2.csv")

model_5a <- lm(vomiting_day14 ~ treat_zmapp, data = ebola)
(coef_model_5a <- RSEs(model_5a))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.847458   0.047616  17.798  < 2e-16 ***
## treat_zmapp -0.237702   0.091459  -2.599  0.01079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("The estimated effect of Zmapp is ", RSEs(model_5a)[2,1],
     " (", coef_model_5a[2,2], ")", sep = "")
```

```
## The estimated effect of Zmapp is -0.2377015 (0.09145949)
```

```
cat("The associated p-value is ", coef_model_5a[2, 4], sep = "")
```

```
## The associated p-value is 0.0107933
```

So **the estimated effect of Zmapp is** **-0.238 (0.091)**, i.e., on average, it reduces the likelihood of vomiting by 23.8 ± 18.2%.
The associated *p*-value is **0.011** (statistically significant at the 5 percent level).

b. Add covariates for vomiting on day 0 and patient temperature on day 0 to the regression from part (a) and report the ATE (with standard error). Also report the p-value.

```
model_5b <- update(model_5a, . ~ . + vomiting_day0 + temperature_day0)
(coef_model_5b <- RSEs(model_5b))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -19.469655   7.607812 -2.5592 0.012054 *
## treat_zmapp       -0.165537   0.081976 -2.0193 0.046242 *
## vomiting_day0      0.064557   0.178032  0.3626 0.717689
## temperature_day0   0.205548   0.078060  2.6332 0.009859 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("The estimated effect of Zmapp is ", RSEs(model_5b)[2,1],
     " (", coef_model_5b[2,2], ")", sep = "")
```

```
## The estimated effect of Zmapp is -0.1655367 (0.0819765)
```

```
cat("The associated p-value is ", coef_model_5b[2, 4], sep = "")
```

```
## The associated p-value is 0.04624205
```

Now **the estimated effect of Zmapp is -0.166 (0.082)**, i.e., on average, it reduces the likelihood of vomiting by 16.6 ± 16.4%.
The associated *p*-value is 0.046 (statistically significant at the 5 percent level).

c. Do you prefer the estimate of the ATE reported in part (a) or part (b)? Why?

I prefer the one reported in **part (b)**—even though it is lower in absolute value (so ZMapp seems to be less effective), the additional information provided by the pre-test values have also lowered the standard error, so we are less uncertain about the results.

d. The regression from part (b) suggests that temperature is highly predictive of vomiting. Also include a control for temperature on day 14 in the regression from part (b) and report the ATE, the standard error, and the p-value.

```
model_5d <- update(model_5b, . ~ . + temperature_day14)
(coef_model_5d <- RSEs(model_5d))
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -22.591585   7.746036 -2.9165 0.004416 **
## treat_zmapp        -0.120101   0.085798 -1.3998 0.164829
## vomiting_day0       0.046038   0.173177  0.2658 0.790934
## temperature_day0    0.176642   0.077024  2.2933 0.024034 *
## temperature_day14   0.060148   0.025831  2.3286 0.022002 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("The estimated effect of Zmapp is ", RSEs(model_5d)[2,1],
     " (", coef_model_5d[2,2], ")", sep = "")
```

```
## The estimated effect of Zmapp is -0.1201006 (0.08579823)
```

```
cat("The associated p-value is ", coef_model_5d[2, 4], sep = "")
```

```
## The associated p-value is 0.1648294
```

Now **the estimated effect of Zmapp is** **-0.12 (0.086)**, i.e., on average, it reduces the likelihood of vomiting by 12 ± 17.2%.
The associated *p*-value is **0.165** (not statistically significant).

e. Do you prefer the estimate of the ATE reported in part (b) or part (d)? Why?

I prefer the one reported in **part (b)**—the estimate reported in part (d) is biased because of the use of a bad control (temperature on day 14, which is itself an outcome variable, affected by the treatment).

f. Now let's switch from the outcome of vomiting to the outcome of temperature, and use the same regression covariates as in part (b). Test the hypothesis that ZMapp is especially likely to reduce men's temperatures, as compared to women's, and describe how you did so. What do the results suggest?

```
model_5f <- lm(temperature_day14 ~ treat_zmapp*male + temperature_day0 +
                        vomiting_day0, data = ebola)
(coef_model_5f <- RSEs(model_5f))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)       48.712690  10.194000   4.7786 6.499e-06 ***
## treat_zmapp       -0.230866   0.118272  -1.9520   0.05391 .
## male               3.085486   0.121773  25.3379 < 2.2e-16 ***
## temperature_day0   0.504797   0.104511   4.8301 5.287e-06 ***
## vomiting_day0      0.041131   0.194539   0.2114   0.83301
## treat_zmapp:male  -2.076686   0.198386 -10.4679 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We'd like to test the hypothesis that ZMapp is especially likely to reduce men's temperatures, as compared to women's. Hence, the null hypothesis would be that the CATEs in both groups are equal to the estimated

ATE.
Testing that hypothesis is equivalent to test whether the interaction effect is null.

The estimate of the interaction term (considering not only "treat_zmapp" and "male", but also "temperature_day0" and "vomiting_day") is -2.077 (0.198), and it is highly statistically significant ($p = 1.9e - 17$).

So the results suggest that **Zmapp is indeed especially likely to reduce men's temperatures, as compared to women's**. Considering the baseline…

```
(baseline_temperature_day0 <-
    mean(ebola[ebola$male == 0 & ebola$treat_zmapp == 0, ]$temperature_day0))
```

```
## [1] 98.54667
```

```
(baseline_vomiting_day0 <-
    mean(ebola[ebola$male == 0 & ebola$treat_zmapp == 0, ]$vomiting_day0))
```

```
## [1] 0.675
```

```
(baseline <- coef_model_5f[1, 1] +
    baseline_temperature_day0 * coef_model_5f[4, 1] +
    baseline_vomiting_day0 * coef_model_5f[5, 1])
```

```
## [1] 98.48654
```

…(i.e., untreated women whose mean likelihood of vomiting on day 0 is 67.5% and whose mean temperature also on day 0 is 98.55 degrees):

1. ZMapp reduces temperature (on day 14) 0.23 degrees.
2. Being a male implies a temperature (on day 14) 3.09 degrees higher.
3. The combined (interaction) effect of treating with ZMapp a male is an extra reduction of temperature (on day 14) of 2.08 degrees.

So here is the average temperature on day 14, by gender and use of ZMapp:

	Female	Male
Untreated	98.49	101.57
Treated	98.26	99.26

> To briefly explain other methods to test the hypothesis, let's use a simplified model which does not consider the covariates on day 0:

```
model_5f_simplified <- update(model_5f, . ~ . - temperature_day0 -
                                          vomiting_day0)
(coef_model_5f_simplified <- RSEs(model_5f_simplified))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)      98.48654    0.10178 967.6308 < 2.2e-16 ***
## treat_zmapp      -0.32346    0.17441  -1.8547   0.06671 .
## male              3.20513    0.18645  17.1902 < 2.2e-16 ***
## treat_zmapp:male -2.26750    0.28577  -7.9346 3.906e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One way (much more complex) to test the null hypothesis is by means of the F-statistic:

```
model_5f_simplified_noIT <- lm(temperature_day14 ~ treat_zmapp + male,
                                         data = ebola)
(coef_model_5f_simplified_noIT <- RSEs(model_5f_simplified_noIT))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 98.80742    0.11120 888.5819 < 2.2e-16 ***
## treat_zmapp -1.20238    0.18320  -6.5631 2.595e-09 ***
## male         2.20873    0.19264  11.4656 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
N <- dim(ebola)[1]
Ftest <- ((sum(model_5f_simplified_noIT$residuals^2) -
                sum(model_5f_simplified$residuals^2)) /
                (length(model_5f_simplified$coefficients) -
                    length(model_5f_simplified_noIT$coefficients))) /
    (sum(model_5f_simplified$residuals^2) /
        (N - length(model_5f_simplified$coefficients)))
Ftest
```

```
## [1] 67.92918
```

```
# What is the p-value of this F-statistic?
num_iter <- 100e3
ATE <- mean(ebola[ebola$treat_zmapp == 1, ]$temperature_day14) -
    mean(ebola[ebola$treat_zmapp == 0, ]$temperature_day14)
Y1 <- Y0 <- ebola$temperature_day14
Y0 <- Y0 - ATE*ebola$treat_zmapp
Y1 <- Y1 + ATE*(1 - ebola$treat_zmapp)
Fdist <- numeric(num_iter)
for (i in 1:num_iter) {
    Z <- sample(ebola$treat_zmapp)
    Y <- Y0*(1 - Z) + Y1*Z
    lm1 <- lm(Y ~ Z * ebola$male)
    lm2 <- lm(Y ~ Z + ebola$male)
    Fdist[i] <- ((sum(lm2$residuals^2) - sum(lm1$residuals^2)) /
                        (length(lm1$coefficients) - length(lm2$coefficients))) /
        (sum(lm1$residuals^2)/(N-length(lm1$coefficients)))
    }
mean(abs(Fdist) >= abs(Ftest))
```

```
## [1] 0
```

Assuming $CATE_{male} = CATE_{female} = ATE$, we don't get an F-statistic as high (in absolute value) as the one we previously calculated (67.9291804) in any of the 100000 replications of the experiment. That's consistent with the p-value of the interaction term (but, as mentioned, this method is far more complex).

Another way to test the hypothesis would be to calculate the p-value of the difference-in-differences:

```
t.test(ebola[ebola$treat_zmapp == 1 & ebola$male == 1, ]$temperature_day14 -
            ebola[ebola$treat_zmapp == 0 & ebola$male == 1, ]$temperature_day14,
         ebola[ebola$treat_zmapp == 1 & ebola$male == 0, ]$temperature_day14 -
            ebola[ebola$treat_zmapp == 0 & ebola$male == 0, ]$temperature_day14)
```

```
##
##  Welch Two Sample t-test
##
## data:  ebola[ebola$treat_zmapp == 1 & ebola$male == 1, ]$temperature_day14 -  and ebola[ebola$treat_zmapp == 1 &
## t = -10.3141, df = 51.383, p-value = 4.07e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.833856 -1.910550
## sample estimates:
##  mean of x  mean of y
## -2.6524032 -0.2802005
```

If we think our sample is not big enough, we could use randomization inference instead of relying on the Central Limit Theorem:

```
CATE_male <-
    mean(ebola[ebola$male == 1 & ebola$treat_zmapp == 1, ]$temperature_day14) -
    mean(ebola[ebola$male == 1 & ebola$treat_zmapp == 0, ]$temperature_day14)
CATE_female <-
    mean(ebola[ebola$male == 0 & ebola$treat_zmapp == 1, ]$temperature_day14) -
    mean(ebola[ebola$male == 0 & ebola$treat_zmapp == 0, ]$temperature_day14)
CATEdif <- numeric(num_iter)
for (i in 1:num_iter){
    Z <- sample(ebola$treat_zmapp)
    CATEdif[i] <- abs(mean(Y1[ebola$male == 1 & Z == 1]) -
                            mean(Y0[ebola$male == 1 & Z == 0]) -
                            (mean(Y1[ebola$male == 0 & Z == 1]) -
                                mean(Y0[ebola$male == 0 & Z==0])))
    }
mean(sum(CATEdif >= abs(CATE_male - CATE_female)))
```

```
## [1] 0
```

The results are the same. Again, the CATEs do not differ as much as the value we obtained in any of the 100000 iterations we ran.

g. Suppose that you had not run the regression in part (f). Instead, you speak with a colleague to learn about heterogenous treatment effects. This colleague has access to a non-anonymized version of the same dataset and reports that he had looked at heterogenous effects of the ZMapp treatment by each of 10,000 different covariates to examine whether each predicted the effectiveness of ZMapp on each of 2,000 different indicators of health, for 20,000,000 different regressions in total. Across these 20,000,000 regressions your colleague ran, the treatment's interaction with gender on the outcome of temperature is the only heterogenous treatment effect that he found to be statistically significant. He reasons that this shows the importance of gender for understanding the effectiveness of the drug, because nothing else seemed to indicate why it worked. Bolstering his confidence, after looking at the data, he also returned to his medical textbooks and built a theory about why ZMapp interacts with processes only present in men to cure. Another doctor, unfamiliar with the data, hears his theory and finds it plausible. How likely do you think it is ZMapp works especially well for curing Ebola in men, and why? (This question is mainly conceptual can be answered without performing any computation.)

I would say it is very unlikely that ZMapp works especially well for curing Ebola in men—this is an example of "**fishing expedition**" or **multiple comparisons**, where comparing among many possible conditions (covariates) overstate the statistical significance.

If results seem to indicate the importance of gender for understanding the effectivenes of ZMapp, it should be tested again in another experiment that only considers gender as a covariate.

h. Now, imagine that what described in part (g) did not happen, but that you had tested this heterogeneous treatment effect, and only this heterogeneous treatment effect, of your own accord. Would you be more or less inclined to believe that the heterogeneous treatment effect really exists? Why?

"Even when the statistical results from hypothesis tests strongly suggest heterogeneity, a more fundamental limitation of subgroup analysis is that it is essentially nonexperimental in character". I.e., even if there is statistical evidence that ZMapp effect is larger in men, that does not imply a cause-and-effect relationship (in other words, that if we were able to change the gender of women, ZMapp would be more effective in them): it could be that gender is just a marker for other factors that increase the effectiveness of ZMapp.

i. Another colleague proposes that being of African descent causes one to be more likely to get Ebola. He asks you what ideal experiment would answer this question. What would you tell him? (Hint: refer to Chapter 1 of Mostly Harmless Econometrics.)

An ideal (but unfeasible) experiment would randomly change the genetic information of participants exposed to the Ebola virus.
A more realistic—and hence less ideal, but equally unethical—experiment would be one in which all participants are still exposed to the virus. Since changing the descent of an individual is not feasible, the sample would consist of people who are of African descent, and people who are not, and the covariates that are supposed to be related with getting Ebola (such as health, age, etc.) should be equally splitted among both groups —i.e., the descent should be the only difference, on average, within the sample.