Experiments and Causality: Problem Set 3

Alex, Micah and Scott

12/7/2020

```
options(digits = 3)
```

```
library(data.table)

library(sandwich)
library(lmtest)

library(ggplot2)
library(patchwork)

library(foreign)
library(knitr)

library(stargazer)
```

1. Replicate Results

Skim Broockman and Green's paper on the effects of Facebook ads and download an anonymized version of the data for Facebook users only.

```
d <- fread("../data/broockman_green_anon_pooled_fb_users_only.csv")
```

```
# To calculate linear model
calculate_lm <- function(d, study){
    return(d[ studyno == study, lm(name_recall ~ treat_ad )])
}

# To calculate and return confidence interval for treatment variable
calculate_conf_int <- function(mod){
  return(coefci(mod, vcov. = vcovHC)[2,] * 100)
}

# To calculate and return robust std errors
robust_se <- function(mod){
  # Adjust standard errors
  cov1  <- vcovHC(mod, type = "HC1")
  return(sqrt(diag(cov1)))
}

# To calculate and return clustered std errors
clustered_se <- function(mod, study=NULL){
  mod$vcovCL_ <- NULL
```

```
  if (is.null(study)){
    mod$vcovCL_ <- vcovCL(mod, cluster = d[, cluster])
  }else{
    mod$vcovCL_ <- vcovCL(mod, cluster = d[studyno == study, cluster])
  }
  return(sqrt(diag(mod$vcovCL_)))
}

# Function to convert coeftest results object to data frame
# using approach adopted from https://stackoverflow.com/questions/35341821/extracting
#                 -significance-score-from-aer-coeftest-results-in-r
coeftest_to_df <- function(x){
  rt=list()                              # generate empty results list
  for(c in 1:dim(x)[2]) rt[[c]]=x[,c]    # writes column values of x to list
  rt=as.data.frame(rt)                   # converts list to data frame object
  names(rt)=names(x[1,])                 # assign correct column names
  names(rt)[4]<-paste("p_value")
  rt[,"sig"]=symnum(rt$p_value, corr = FALSE,
                    na = FALSE,cutpoints =
                      c(0, 0.001, 0.01, 0.05, 0.1, 1),
                      symbols = c("***", "**", "*", ".", " "))
  return(rt)
}
```

1. 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 (the dependent variable is `name_recall`). After you estimate your model, write a narrative description about what you've learned.

- **Note**: Ignore the blocking the article mentions throughout this problem.
- **Note**: You will estimate something different than is reported in the study.

```
mod_study1 <- calculate_lm(d, study = 1)
mod_study1$vcovHC_ <- vcovHC(mod_study1)

calculate_conf_int(mod_study1)
```

```
##  2.5 % 97.5 %
##  -5.12   3.16
```

```
summary(mod_study1) #display t-test and summary
```

```
##
## Call:
## lm(formula = name_recall ~ treat_ad)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.182 -0.182 -0.173 -0.173  0.827
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1825     0.0161   11.30   <2e-16 ***
## treat_ad     -0.0098     0.0210   -0.47     0.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.382 on 1362 degrees of freedom
## Multiple R-squared:  0.00016,    Adjusted R-squared:  -0.000574
## F-statistic: 0.217 on 1 and 1362 DF,  p-value: 0.641
```

Just looking at the treatment variable `treat_ad` for study 1, the `namerecall` dependent variable is not explained by the treatment. It appears that the ads had little or no effect on candidate name recognition. This is evident by the p-value of 0.643 which indicates that the treatment effect is not statistically significant. The standard errors are robust standard errors (and not clustered).

Also note that the adjusted R2 is negative here which is indicating that the chosen model fits worse than a horizontal line. This is indicating that the chosen model does not follow the trend of the data, so it fits worse than a horizontal line.

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

   The study is clustered on demographic groupings of individuals. Taking clustering into account penalizes the model and increases the standard errors, beause uncertainty at the clustered level must be taken into account in the standard errors also. Hence we see an increases in the clustered standard errors.

3. Estimate a regression that estimates the effect of the ad on candidate name recognition in Study 1, but this time take take clustering into account when you compute the standard errors.

- The estimation of the *model* does not change, only the estimation of the standard errors.
- You can estimate these clustered standard errors using `sandwich::vcovCL`, which means: "The `vcovCL` function from the `sandwich` package.
- We talk about this more in code that is availbe in the course repo.

```
mod_study1 <- calculate_lm(d, study = 1)
clustered_se(mod_study1, study=1)
```

```
## (Intercept)    treat_ad
##      0.0185      0.0238
```

4. Change the context: estimate the treatment effect in Study 2, using clustered standard errors. If you've written your code for part 3 carefully, you should be able to simply change the row-scoping that you're calling. If you didn't write it carefully, for legibility for your colleagues, you might consider re-writting your solution to the last question. Descriptively, do the treatment effects look different between the two studies? Are you able to conduct a formal test by comparing these coefficients? Why, or why not?

```
mod_study2 <- calculate_lm(d, study = 2)
clustered_se(study=2, mod_study2)
```

```
## (Intercept)    treat_ad
##      0.0182      0.0355
```

```
stargazer(
  mod_study1,
  mod_study2,
  type = 'text',
  se=list(clustered_se(mod_study1, 1), clustered_se(mod_study2, 2)),
  header=F
  )
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
```

```
##                                    name_recall
##                              (1)                  (2)
## ------------------------------------------------------------
## treat_ad                    -0.010               -0.003
##                            (0.024)              (0.036)
##
## Constant                   0.182***             0.606***
##                            (0.018)              (0.018)
##
## ------------------------------------------------------------
## Observations                1,364                1,337
## R2                          0.0002              0.00001
## Adjusted R2                 -0.001               -0.001
## Residual Std. Error  0.382 (df = 1362)    0.489 (df = 1335)
## F Statistic         0.217 (df = 1; 1362) 0.008 (df = 1; 1335)
## ============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

We present stargazer summary here to show the treatment coefficients for both studies using clustered standard errors. However, this is not a test and only presents results of two completely different studies. In order to test to compare coefficients, even though they are different studies and different contexts – we can run a test by creating a model with interaction terms with respect to the treatment variable and and using the difference in difference effects between study 1 and study 2 to compare these coefficients.

5. Run a regression to test for the effect of the ad on candidate name recognition, but this time use 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.

- Does this estimate tell you anything useful?
- Why or why not?
- Can you say that the treatment assignment procedure used is fully random when you estimate this model? Or is there some endogeneous process that could be confounding your estimate?

```
mod_pooled <- d[, lm(name_recall ~ treat_ad)]
mod_pooled$vcovCL_ <- vcovCL(mod_pooled, cluster = d[, cluster])
coefci(mod_pooled, vcov = mod_pooled$vcovCL_)
```

```
##              2.5 % 97.5 %
## (Intercept)  0.418  0.491
## treat_ad    -0.207 -0.103
```

```
clustered_se(mod_pooled)
```

```
## (Intercept)    treat_ad
##      0.0186      0.0267
```

```
model_metrics_df <- coeftest_to_df(coeftest(mod_pooled, mod_pooled$vcovCL1_))
model_metrics_df
```

```
##             Estimate Std. Error t value   p_value sig
## (Intercept)    0.454     0.0122   37.26 1.48e-245 ***
## treat_ad      -0.155     0.0188   -8.27  2.16e-16 ***
```

The intercept and treatment effect are statistically significant, but we cannot rely on these results. Hence these are not useful for interpretation.

The reason why the estimates of coefficients are not useful here is because the way this linear regression test has been written is incorrect and not considering various differences in both studies.

There is a difference in context to how data was collected for studies 1 and 2. For example, study 1 was conducted for name recognition of a relatively unknown candidate, while study 2 was for a well-known candidate. Study 1 was conducted at the town level, however, study 2 was conducted at the county level (which can have very wide differences from economic and political standpoints).

Furthermore, there is a major timing difference between these studies (since study 1 took place a month before the election versus study 2 was undertaken a week before the election). As such this could violate the inference and excludability requirements of running linear models.

From a non-interference perpective, combining data of two relatively dissimilar but different studies, in a single test violates randomization and non-interference requirement. We can imagine a group of 24-year olds hanging out in local bars and talking about local politics, thus influencing each other's responses in the studies. The timing component mentioned above (with study 1 being conducted a month before election and study 2 being conducted a week before election), also poses spillover and inference problems,

6. Estimate a model that uses all the data, but this time include a variable that identifies whether an observation was generated during Study 1 or Study 2.

- What is estimated in the "Study 2 Fixed Effect"?
- What is the treatment effect estimate and associated p-value?
- Think a little bit more about the treatment effect that you've estimated: Can this treatment effect, as you've entered it in the model be *different* between Study 1 and Study 2?
- Why or why not?

```
mod_fe <- d[, lm(name_recall ~ treat_ad + as.factor(studyno))]
model_metrics_df <- coeftest_to_df(coeftest(mod_fe, mod_fe$vcovCL1_))

model_metrics_df
```

```
##                     Estimate Std. Error t value   p_value sig
## (Intercept)          0.18068     0.0160  11.297  5.99e-29 ***
## treat_ad            -0.00678     0.0182  -0.373  7.09e-01
## as.factor(studyno)2  0.42610     0.0180  23.731 3.20e-113 ***
```

```
stargazer(
  mod_fe,
  type = 'text',
  se=list(clustered_se(mod_fe)),
  header=F
)
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                              name_recall
## ------------------------------------------------
## treat_ad                      -0.007
##                               (0.020)
##
## as.factor(studyno)2          0.426***
##                               (0.021)
##
## Constant                     0.181***
##                               (0.017)
##
```

```
## ------------------------------------------------
## Observations                    2,701
## R2                              0.193
## Adjusted R2                     0.193
## Residual Std. Error       0.438 (df = 2698)
## F Statistic           323.000*** (df = 2; 2698)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The study 2 fixed effects are estimating the time-invariant un-observable factors related to study 2. Including the study fixed effects is helping us remove omitted variable bias and account for within-group variation over time. Across-group variation is not used to estimate the regression coefficients, because this variation might reflect omitted variable bias.

```
  * study 2 fixed effect estimate and p-value: 0.426 and \ensuremath{3.196\times 10^{-113}}
  * treatment effect estimate and p-value: -0.007 and 0.709
```

The treatment effect can be different between study 1 and 2 due to hetergeneous treatment effects and different inteactions of the treatment variable with both studies. Note that we DO NOT have an interaction term in our model here; and considering the fixed effects solely without accounting for their heterogeneous interaction with the treatment variable. The proper and correct way to use the fixed effects in this case would be to measure the difference in difference estimate between study 2 vs study 1 (study 1 being baseline) using an interaction term with the treatment effect.

7. Estimate a model that lets the treatment effects be different between Study 1 and Study 2. With this model, conduct a formal test – it must have a p-value associated with the test – for whether the treatment effects are different in Study 1 than Study 2.

```
mod_interaction <- d[, lm(name_recall ~ treat_ad + as.factor(studyno) +
                            (as.factor(studyno) * treat_ad))]

# calculating model coeftest metrics to be displayed as a dataframe
model_metrics_df <- coeftest_to_df(coeftest(mod_interaction,
                                            mod_interaction$vcovCL1_))
model_metrics_df
```

```
##                               Estimate Std. Error t value  p_value sig
## (Intercept)                    0.18247     0.0185   9.845 1.71e-22 ***
## treat_ad                      -0.00980     0.0241  -0.406 6.85e-01
## as.factor(studyno)2            0.42332     0.0231  18.299 1.32e-70 ***
## treat_ad:as.factor(studyno)2   0.00699     0.0367   0.191 8.49e-01
```

```
# showing final results for interaction model
stargazer(
  mod_interaction,
  type = 'text',
  se=list(clustered_se(mod_interaction)),
  header=F
)
```

```
##
## ============================================================
##                                  Dependent variable:
##                              ---------------------------
##                                       name_recall
## ------------------------------------------------------------
## treat_ad                                -0.010
##                                         (0.024)
```

```
##
## as.factor(studyno)2                    0.423***
##                                         (0.026)
##
## treat_ad:as.factor(studyno)2            0.007
##                                         (0.043)
##
## Constant                               0.182***
##                                         (0.018)
##
## ------------------------------------------------------------
## Observations                              2,701
## R2                                        0.193
## Adjusted R2                               0.192
## Residual Std. Error                 0.438 (df = 2697)
## F Statistic                    215.000*** (df = 3; 2697)
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

We create a linear model with interaction term between study 2 fixed effects with the treatment variable. This indicates the difference of how much the treatment varies for study 2 over study 1.

However, we see the p-value of this HTE (using clustered standard errors for the model) as 0.849 which doesn't indicate a statistically significant heterogeneous treatment effect.

2. Peruvian Recycling

Look at this article about encouraging recycling in Peru. The paper contains two experiments, a "participation study" and a "participation intensity study." In this problem, we will focus on the latter study, whose results are contained in Table 4 in this problem. You will need to read the relevant section of the paper (starting on page 20 of the manuscript) in order to understand the experimental design and variables. (*Note that "indicator variable" is a synonym for "dummy variable," in case you haven't seen this language before.*)

1. 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.

   ATE is the same as what is mentioned as the coefficient estimate by the model at 0.187 Confidence interval 0.251 and 0.123

2. 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.

   ATE is the same as what is mentioned as the coefficient estimate by the model at -0.024 Confidence interval 0.054 and 0.054

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

   - Percentage of visits turned in a bag
   - Avg. no. of bins turned in per week
   - Avg. weight (in kg) of recyclables turned in per week

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

   None

5. 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 six weeks of treatment? 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.

   $Y_{AB} = [\beta_{A1}B + \beta_{A2}S + \lambda Ybl_A + P_A + \alpha_j + \epsilon_i] - [\beta_{B1}B + \beta_{B2}S + \lambda Ybl_j + P_B + \alpha_j + \epsilon_i]$ represents the difference in model output for households A and B

   α_j cancels out because of fixed street effects that are same for both households A and B

   ϵ_i cancels out because experiments are randomized and observations are identical and independently distributed (i.i.d), hence we can assume error terms are homoskedastic (constant variance of error terms).

   B, S, P_i are indicator variables, hence cancel out between A and B households (as they are identical being in the same treatment groups)

   Therefore, we only have difference in model prediction due to λYbl_{AB} only.

   Which can be calculated as 0.281 X 2 that comes to 0.562

6. 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 our reasoning.

   This will be an example of omitted variable bias. If this were an observational study, and we omitted baseline variable, its effects would influence Y only through the other independent variables present in the equation. Assuming that $\alpha.Ybl$ was positively correlated to the treatment variable (on providing a recycle bin). Hence, in this case, the effect of removing the baseline variable on the resuls of providing a recycle bin would have be an overstatement of effects on the latter.

   However, this is an experimental study with randomized treatment assignment: With randomized treatment assignment, we know that treatment is uncorrelated with everything else: both observable covariates and unobservables we can't measure. So in an experiment, we don't have to worry about omitted variable bias, because we should get approximately the same answer no matter how many covariates we include. However, with additional explained covariates the standard error of the treatment effect variable shrinks, leading to higher certainty.

   Hence to summarize, missing the omitted baseline variable would have no impact on the coefficient of the treatment variable (providing a recycle bin), but would expand its standard error, leading to a reduction in the variables' significance.

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

   Including `hascellphone` is not a bad control since it doesn't directly impact the dependent variable. This does seem to be a good control since it allows people to get the flyer treatment and help narrow the treatment variable slightly.

   However this probably represents a design flaw because `havecell` may indicate household affluence, which could directly or indirectly represent the wealth of a household, and therefore, their ability of buying recyclable good more than those households with no cell phone.

   Therefore as a final design thought, the designers of this experiment should have probably blocked the treatment of cell phone ownership.

8. 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.

   On a theoritical standpoint, coefficients of a treatment effect (in this case `sms`) should not change with the adddition or removal or pre-treatment covariates (their standard errors can get tightened). Our recyclcing experiment is randomized, and hence there should be no relation between the coefficient of `sms` and `nocell`.

   However, is it possible that we may see increase in coefficient of `sms` treatment effect variable upon the removal of `nocell` covariate. We know that `nocell` covariate is statistically significant at the 5% confidence level. Hence removing this covariate will shift it's impact to some of the covariates. Additionally, `nocell` indicates a causal relationship between affluence and recycling (to a certain extent), therefore removing that covariate could shift some of it's causal interpretation to `sms` which is closely interlinked to someone owing a cell phone.

3. Multifactor Experiments

Staying with the same experiment, now think about multifactor experiments.

1. 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 3x3 dimension experiment with the following axes: 3 bin attributes [bin with or without sticker or no bin at all] and 3 attributes for SMS [personalized sms, generic sms or no sms at all].

   The `havecell` is not included in the design dimensions because it is a pre-treatment covariate which is not included in the administration of treatment to subjects; a.k.a. not an intervention (i.e. designers cannot control who has a cell or not – similar to how they are giving out bins and text messages).

2. 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?

   All dummy variables equal to zero in the case of results depicted in Table 4B, represent households in the control group (i.e. households with no treatment variables for receipt of recycle bins and no SMS messages; along with households that have no cells).

   Therefore this baseline category of households in control group represents the intercept.

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

   In column (1) of Table 4B, we find that households that received a bin but without any sticker were 3.5 percentage points more likely to turn in recyclables over the control group.

4. 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?

   Recycling bin with sticker. Because of the 2% excess coefficient for recycling bin with sticker – and same level of statistical significance in both cases (as seen in the standard errors of 0.015 in both cases).

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

   Difference between both treatment effects is 2%. We know that the standard error for the difference between both treatment effects is the sum of the standard errors of individual treatment effects. Therefore, the standard error of the difference between both treatment effects is 0.06. We

see that this is not less than 2 standard deviations from the difference in the coefficients, and hence not significant.

Another approach of checking the statistical significance of the difference betwen both treatment effects (with and without sticker) would be to conduct an T-test (or two-sides T-test) with a sharp null hypothesis that both treatments are the same.

Table 4b already provides us with a p-value for such an F-test 1 = 2 (at the bottom). The p-value for F-test 1=2 is 0.31 which indicates the differences in treatment effects are not significant at the 5% significance level.

6. 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."

   Table 4C represents every distinct combination of bin and SMS message treatment; the omitted category being no phone and receiving no bin.

4. Now! Do it with data

Download the data set for the recycling study in the previous problem, obtained from the authors. We'll be focusing on the outcome variable Y="number of bins turned in per week" (avg_bins_treat).

```
d <- foreign::read.dta("../data/karlan_data_subset_for_class.dta")
d <- data.table(d)
head(d)
```

```
##    street havecell avg_bins_treat base_avg_bins_treat bin sms bin_s bin_g sms_p
## 1:      7        1          1.042               0.750   1   1     1     0     0
## 2:      7        1          0.000               0.000   0   1     0     0     1
## 3:      7        1          0.750               0.500   0   0     0     0     0
## 4:      7        1          0.542               0.500   0   0     0     0     0
## 5:      6        1          0.958               0.375   1   0     0     1     0
## 6:      8        0          0.208               0.000   1   0     0     1     0
##    sms_g
## 1:     1
## 2:     0
## 3:     0
## 4:     0
## 5:     0
## 6:     0
```

```
## Do some quick exploratory data analysis with this data.
## There are some values in this data that seem a bit strange.

## Determine what these are.
## Don't make an assessment about keeping, changing, or
## dropping just yet, but at any point that your analysis touches
## these variables, you'll have to determine what is appropriate
## given the analysis you are conducting.
```

```
calculate_conf_int <- function(mod){
  return(coefci(mod, vcov. = vcovHC)[2,] * 100)
}

robust_se <- function(mod){
  cov1  <- vcovHC(mod, type = "HC1")
```

```
  return(sqrt(diag(cov1)))
}
```

1. 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, using **of course** robust standard errors (use these throughout).

```
# just to be consistent with the paper
mod_1 <- d[, lm(avg_bins_treat ~ bin)]
mod_1$vcovHC_ <- vcovHC(mod_1)

#showing robust standard errors
robust_se(mod_1)
```

```
## (Intercept)         bin
##      0.0115      0.0208
```

```
#showing confidence interval using robust SE
calculate_conf_int(mod_1)
```

```
##  2.5 % 97.5 %
##   9.45  17.62
```

95% confidence interval for the treatment effect is 9.454, 17.622

2. 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.

```
mod_2 <- d[!is.na(street),lm(avg_bins_treat ~ bin + base_avg_bins_treat)]

#showing confidence interval using robust SE
calculate_conf_int(mod_2)
```

```
##  2.5 % 97.5 %
##   9.06  15.79
```

95% confidence interval for the treatment effect for this updated model is is 9.056, 15.792

This model should also estimate similar treatment effect (as the prior model). However, by adding new covariate, the confidence interval of the treatment effect should get narrow (as it happens in this case).

NOTE: we are removing NA's from model 2, because further down the code, we intend on matching model 4 (with street effects) and model 2 using ANOVA (and would like to have same data rows in both models)

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

```
mod_3 <- d[!is.na(street), lm(avg_bins_treat ~ bin +
                                base_avg_bins_treat + as.factor(street))]

#showing confidence interval using robust SE for model 3 treatment effect
calculate_conf_int(mod_3)
```

```
##  2.5 % 97.5 %
##   7.68  15.09
```

95% confidence interval for the treatment effect for our 3rd model (with fixed street effects) is 7.684, 15.093

4. Recall that the authors described their experiment as "stratified at the street level," which is a synonym for blocking by street. Does including these block fixed effects change the standard errors of the estimates *very much*? Conduct the appropriate test for the inclusion of these block fixed effects, and interpret them in the context of the other variables in the regression.

```
mod_4 <- d[!is.na(street),lm(avg_bins_treat ~ bin +
                               base_avg_bins_treat + as.factor(street))]
```

```
stargazer(mod_1, mod_2, mod_4,
          type = 'text',
          se = list(
                    robust_se(mod_1),
                    robust_se(mod_2),
                    robust_se(mod_4)),
          add.lines = list(
            c('SE Flavor','Robust', 'Robust', 'Robust'),
            c('Street fixed effects', 'No', 'No', 'Yes')
            ),
          omit = 'street',
          model.numbers=FALSE,
          column.labels = c("Model 1", "Model 2", "Model 4"),
          header=F
          )
```

```
##
## ================================================================================================
##                                                Dependent variable:
##                        ---------------------------------------------------------------------
##                                                   avg_bins_treat
##                                Model 1                  Model 2                  Model 4
## ------------------------------------------------------------------------------------------------
## bin                            0.135***                 0.124***                 0.114***
##                                (0.021)                  (0.017)                  (0.018)
##
## base_avg_bins_treat                                     0.390***                 0.374***
##                                                         (0.030)                  (0.027)
##
## Constant                       0.635***                 0.352***                 0.368***
##                                (0.011)                  (0.021)                  (0.035)
##
## ------------------------------------------------------------------------------------------------
## SE Flavor                       Robust                   Robust                   Robust
## Street fixed effects              No                       No                       Yes
## Observations                    1,785                    1,782                    1,782
## R2                              0.024                    0.338                    0.436
## Adjusted R2                     0.024                    0.337                    0.372
## Residual Std. Error      0.405 (df = 1783)        0.333 (df = 1779)        0.324 (df = 1600)
## F Statistic          44.500*** (df = 1; 1783) 454.000*** (df = 2; 1779) 6.840*** (df = 181; 1600)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

```
# Running F-test on models 2 vs. 4 to check for statistically significant differences in parameters
test_fixed_effects <- anova(mod_2, mod_4, test='F')
test_fixed_effects
```

```
## Analysis of Variance Table
```

```
##
## Model 1: avg_bins_treat ~ bin + base_avg_bins_treat
## Model 2: avg_bins_treat ~ bin + base_avg_bins_treat + as.factor(street)
##   Res.Df RSS  Df Sum of Sq    F  Pr(>F)
## 1   1779 197
## 2   1600 168 179      29.3 1.56 9.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As seen in the stargazer output, we are comparing models 1, 2 and 4 — and can observe that the standard errors of the estimates only slightly change (after adding the block fixed effects) – even after we use robust standard errors in our stargazer package.

Comparing models 2 and 4 using ANOVA and running an F-test, we get very low p-value of 9.514×10^{-6} which means that there is significant difference between the residuals of both models at the 5% and 1% significance levels. Since the only difference between both models is the addition of street fixed effects in model 4 (which are statistically significant), we find that the fixed effects and other group of variables in model 4 are jointly significant.

Therefore, we can infer that adding the addition set of covariates (i.e. fixed street effects in model 4) improves our results further. The coefficient of treatment effect is getting more precise, and hence model explanability increases.

In general block randomization at street level should reduce variance of treatment effects because it reduces the chance of lot of similar households at different parts of town getting clubbed together in treatment or control. This is what the street fixed effect is also doing here, because here we are taking the grouped-mean effect of households at same-street as an average coefficient; hence ultimately reducing the variance of treatment effect a.k.a. reducing "dirty variation" in the treatment variable (similar to block randomization).

5. 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.

```
d[, nocell:= 1 - havecell]
```

6. 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.

```
mod_5 <- d[ , lm(avg_bins_treat ~ bin + base_avg_bins_treat +
                   nocell + as.factor(street))]

# printing confidence interval as asked
calculate_conf_int(mod_5)
```

```
##  2.5 % 97.5 %
##    7.8   15.2
```

```
#printing detailed model-by-model comparison for convenience
stargazer(mod_2, mod_4, mod_5,
        type = 'text',
        se = list(
                  robust_se(mod_2),
                  robust_se(mod_4),
                  robust_se(mod_5)),
        add.lines = list(
          c('SE Flavor','Robust', 'Robust', 'Robust'),
```

```
          c('Street fixed effects', 'No', 'Yes', 'Yes')
          ),
        omit = 'street',
        model.numbers=FALSE,
        column.labels = c("Model 2", "Model 4", "Model 5"),
        header=F
        )
```

```
##
## ================================================================================================
##                                              Dependent variable:
##                      ------------------------------------------------------------------------
##                                                 avg_bins_treat
##                               Model 2                 Model 4                 Model 5
## ------------------------------------------------------------------------------------------------
## bin                          0.124***                0.114***                0.115***
##                               (0.017)                 (0.018)                 (0.018)
##
## base_avg_bins_treat          0.390***                0.374***                0.373***
##                               (0.030)                 (0.027)                 (0.027)
##
## nocell                                                                       -0.050***
##                                                                               (0.017)
##
## Constant                     0.352***                0.368***                0.387***
##                               (0.021)                 (0.035)                 (0.035)
##
## ------------------------------------------------------------------------------------------------
## SE Flavor                     Robust                  Robust                  Robust
## Street fixed effects            No                      Yes                     Yes
## Observations                  1,782                   1,782                   1,781
## R2                            0.338                   0.436                   0.439
## Adjusted R2                   0.337                   0.372                   0.375
## Residual Std. Error      0.333 (df = 1779)       0.324 (df = 1600)       0.323 (df = 1598)
## F Statistic          454.000*** (df = 2; 1779) 6.840*** (df = 181; 1600) 6.880*** (df = 182; 1598)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

95% confidence interval for the treatment effect for model 5 (after adding no-cell covariate) is 7.801, 15.219. Model 4 gave us a confidence interval of 7.684, 15.093.

We don't see a significant reduction in confidence intervals even after adding the additional "no cell phone" covariate because the confidence interval of the treatment effect is already quite significant at the 1% significance level.

Addition of new covariate is not adding further statistical power to the test. The `base_avg_bins_treat` variable already bakes in historical recycling measures related to a lot of confounding factors. Hence adding a new covariate (which is perhaps correlated to `base_avg_bins_treat`) doesnt add much statistical precision or power to our model.

7. 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.

```
mod_6 <- d[ , lm(avg_bins_treat ~ bin + sms + nocell +
                   base_avg_bins_treat + as.factor(street))]
```

```
stargazer(mod_6,
        type = 'text',
        se = list(
                  robust_se(mod_6)),
        add.lines = list(
          c('SE Flavor','Robust'),
          c('Street fixed effects', 'Yes')
          ),
        omit = 'street',
        model.numbers=FALSE,
        column.labels = c("Model 6"),
        header=F
        )
```

```
##
## ================================================
##                           Dependent variable:
##                       ---------------------------
##                             avg_bins_treat
##                                 Model 6
## ------------------------------------------------
## bin                            0.115***
##                                 (0.018)
##
## sms                             0.005
##                                 (0.022)
##
## nocell                         -0.047**
##                                 (0.021)
##
## base_avg_bins_treat            0.373***
##                                 (0.027)
##
## Constant                       0.385***
##                                 (0.038)
##
## ------------------------------------------------
## SE Flavor                       Robust
## Street fixed effects              Yes
## Observations                     1,781
## R2                               0.439
## Adjusted R2                      0.375
## Residual Std. Error       0.323 (df = 1597)
## F Statistic           6.830*** (df = 183; 1597)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
calculate_conf_int(mod_6)
```

```
##  2.5 % 97.5 %
##   7.79  15.22
```

We see that adding `sms` doesn't reduce confidence intervals for the treatment effect. This seems to be a design flaw in the experiment. We know that in linear models past performance metrics are always best predictors of future results. Hence adding the `base_avg_bins_treat` has had the most

impact on the regression. Furthermore, it is possible that the sms covariate is correlated to the `base_avg_bins_treat` variable. `base_avg_bins_treat` is thus encompassing various endogenous relationships with other covariates including household affluence, ownership of gadgets such as cell phones etc; which ultimately provide a rough measure of how much a given household recycles on average.

This can also be illustrated by the model results which indicate a p-value of `0.022` for the `sms` variable (which is not statistical significant at the 5% significance level).

8. 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.

```
mod_7 <- d[ , lm(avg_bins_treat ~ bin_g + bin_s +
                   sms_p + sms_g + nocell +
                   base_avg_bins_treat + as.factor(street))]

#providing confidence interval for all treatments and covariates
coefci(mod_7, vcov. = vcovHC)[1:8,]
```

```
##                        2.5 %   97.5 %
## (Intercept)          0.3095  0.46035
## bin_g                0.0540  0.15234
## bin_s                0.0802  0.17545
## sms_p               -0.0633  0.04721
## sms_g               -0.0346  0.07405
## nocell              -0.0916 -0.00119
## base_avg_bins_treat  0.3142  0.43349
## as.factor(street)2  -0.2152  0.01381
```

```
#providing a stargazer analysis to match the study (using robust SE)
stargazer(mod_7,
        type = 'text',
        se = list(
                  robust_se(mod_7)),
        add.lines = list(
          c('SE Flavor','Robust'),
          c('Street fixed effects', 'Yes')
          ),
        omit = 'street',
        model.numbers=FALSE,
        column.labels = c("Model 7"),
        header=F
        )
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                             avg_bins_treat
##                                 Model 7
## ------------------------------------------------
## bin_g                           0.103***
##                                  (0.023)
##
## bin_s                           0.128***
```

```
##                                 (0.022)
##
## sms_p                           -0.008
##                                 (0.026)
##
## sms_g                            0.020
##                                 (0.026)
##
## nocell                         -0.046**
##                                 (0.021)
##
## base_avg_bins_treat            0.374***
##                                 (0.027)
##
## Constant                       0.385***
##                                 (0.038)
##
## ------------------------------------------------
## SE Flavor                        Robust
## Street fixed effects               Yes
## Observations                     1,781
## R2                               0.440
## Adjusted R2                      0.375
## Residual Std. Error       0.323 (df = 1595)
## F Statistic           6.770*** (df = 185; 1595)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

We see from the stargazer results that both `bin_s` and `bin_g` treatment effects are statistically significant. However, from the F-test provided in the study, for a null hypothesis `bin_s` and `bin_g` being different – we see that the difference between both treatment effects is not statistically significant.

Hence the final outcome of treatment effect, as a collective measure of providing bin to customers should be the generic `bin` indicator. We cannot say with confidence that either of the two treatments (`bin with sticker` or `bin without sticker`) is more effective than the other.

Therefore, from this analysis we should expect the confidence interval of the generic `bin` indicator (from model 6) should lie in between the confidence intervals of the `bin_s` and `bin_g` treatment indicators. Which we can confirm is the case here.

5. A Final Practice Problem

Now for a fictional scenario. An emergency two-week randomized controlled trial of the experimental drug ZMapp is conducted to treat Ebola. (The control represents the usual standard of care for patients identified with Ebola, while the treatment is the usual standard of care plus the drug.)

Here are the (fake) data.

```
d <- fread("../data/ebola_rct2.csv")
head(d)
```

```
##    temperature_day0 dehydrated_day0 treat_zmapp temperature_day14
## 1:             99.5               1           0              98.6
## 2:             97.4               0           0              98.0
## 3:             97.0               0           1              97.9
```

```
## 4:             99.7              1           0             98.4
## 5:             99.6              1           1             99.3
## 6:             98.3              1           1             99.8
##    dehydrated_day14 male
## 1:                1    0
## 2:                1    0
## 3:                0    1
## 4:                1    0
## 5:                1    0
## 6:                1    1
```

You are asked to analyze it. Patients' temperature and whether they are dehydrated is recorded on day 0 of the experiment, then ZMapp is administered to patients in the treatment group on day 1. Dehydration and temperature is again recorded on day 14.

1. 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 dehydrated on day 14? What is the p-value associated with this estimate?

```
mod_1 <-  d[ , lm(dehydrated_day14 ~ treat_zmapp)]
coeftest(mod_1, vcov. = mod_1$vcovHC_)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8475     0.0548   15.46   <2e-16 ***
## treat_zmapp  -0.2377     0.0856   -2.78   0.0066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

   We use robust standard errors to estimate the effect of ZMapp. The estimated effect of using ZMapp is -0.238, which means there is an overall reduction in 14th day temperature. The standard error estimated by the model is 0.086.

   We see that probability of effect this size if the sharp null were true i.e. the p-value, is 0.66%, this is evidence to say treatment is effective.

2. Add covariates for dehydration 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.

```
mod_2 <-  d[ , lm(dehydrated_day14 ~ treat_zmapp  + dehydrated_day0 + temperature_day0)]
mod_2$vcovHC_ <- vcovHC(mod_2)
coeftest(mod_2, vcov. = mod_2$vcovHC_)
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -19.4697     7.6078   -2.56   0.0121 *
## treat_zmapp       -0.1655     0.0820   -2.02   0.0462 *
## dehydrated_day0    0.0646     0.1780    0.36   0.7177
## temperature_day0   0.2055     0.0781    2.63   0.0099 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

* ATE -0.166
* Standard error 0.082

* p-value 4.62%

3. Do you prefer the estimate of the ATE reported in part (a) or part (b)? Why? Report the results of the F-test that you used to form this opinion.

```
test_object <- anova(mod_1, mod_2, test='F')

test_object
```

```
## Analysis of Variance Table
##
## Model 1: dehydrated_day14 ~ treat_zmapp
## Model 2: dehydrated_day14 ~ treat_zmapp + dehydrated_day0 + temperature_day0
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1     98 17.4
## 2     96 12.9  2      4.47 16.6 6.5e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conduct an F-test on models 1 and 2, and find p-value of the f-test as 6.472×10^{-7}, which indicates that the means between both models are significantly different (and that the group of variables in 2nd model are jointly significant). We prefer model 2 because it is better to add these pre-treatment covariates. This gives model 2 much more explanatory power.

4. The regression from part (2) suggests that temperature is highly predictive of dehydration. Add, temperature on day 14 as a covariate and report the ATE, the standard error, and the p-value.

```
mod_3 <-  d[ , lm(dehydrated_day14 ~ treat_zmapp  +
                    dehydrated_day0 + temperature_day0 + temperature_day14)]
mod_3$vcovHC_ <- vcovHC(mod_3)
coeftest(mod_3, vcov. = mod_3$vcovHC_)[1:5, ]
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -22.5916     7.7460  -2.917  0.00442
## treat_zmapp        -0.1201     0.0858  -1.400  0.16483
## dehydrated_day0     0.0460     0.1732   0.266  0.79093
## temperature_day0    0.1766     0.0770   2.293  0.02403
## temperature_day14   0.0601     0.0258   2.329  0.02200
```

* ATE -0.12
* Standard error 0.086
* p-value 16.48%

5. Do you prefer the estimate of the ATE reported in part (b) or part (d)? What is this preference based on?

We prefer part b (5.2) because adding temperature_day14 is a bad control (this is affected by the treatment itself) and should not be included as explanatory variable in the model.

6. Now let's switch from the outcome of dehydration to the outcome of temperature, and use the same regression covariates as in the chunk titled `add pre-treatment measures`. Test the hypothesis that ZMapp is especially likely to reduce mens' temperatures, as compared to womens', and describe how you did so. What do the results suggest?

```
# mod_4 <-  d[ , lm(temperature_day14 ~ treat_zmapp  +
      #temperature_day0 + dehydrated_day0  +
      #I(male =='0') + (I(male =='0') * treat_zmapp) )]

mod_4 <-  d[ , lm(temperature_day14 ~ treat_zmapp  +
```

```
                    temperature_day0 + dehydrated_day0  +
                    male + (male * treat_zmapp) )]

mod_4$vcovHC_ <- vcovHC(mod_4)

stargazer(mod_4,
          type = 'text',
          se = list(mod_4$robust.se),
          add.lines = list(
            c('SE Flavor','Robust')
            ),
          column.labels = c("Model 4 - HTE"),
          model.numbers=FALSE,
          header=F
          )
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                            temperature_day14
##                              Model 4 - HTE
## ------------------------------------------------
## treat_zmapp                     -0.231*
##                                 (0.119)
##
## temperature_day0               0.505***
##                                 (0.095)
##
## dehydrated_day0                  0.041
##                                 (0.182)
##
## male                           3.080***
##                                 (0.126)
##
## treat_zmapp:male               -2.080***
##                                 (0.192)
##
## Constant                       48.700***
##                                 (9.270)
##
## ------------------------------------------------
## SE Flavor                        Robust
## Observations                      100
## R2                               0.906
## Adjusted R2                      0.901
## Residual Std. Error         0.452 (df = 94)
## F Statistic            181.000*** (df = 5; 94)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

We see that the new model, does show a different heterogeneous treatment effect for males (male * zmapp), leading to a statistically significant reduction in temperature when a treatment is administered for males (male == 1).

The stargazer output shows that the ATE for treatment for females (male == 0) is -0.231. However, for males, the HTE is -2.077 (p-value < 0.01). Which means that for males, the treatment should lead to a -2 degree reduction in day 14 temperature.

Better treatment effects for males, but worse outcomes for men nevertheless (perhaps because the disease hits males harder).

7. Which group – those that are coded as `male == 0` or `male == 1` have better health outcomes in control? What about in treatment? How does this help to contextualize whatever heterogeneous treatment effect you might have estimated?

```
d[, mean(temperature_day14), by=.(male, treat_zmapp)]
```

```
##    male treat_zmapp    V1
## 1:    0           0  98.5
## 2:    1           1  99.1
## 3:    0           1  98.2
## 4:    1           0 101.7
```

The above code block shows that in control (treat_zmapp == 0), males (male == 1) have on an average 2 degrees higher temperature than females. This confirms with our findings with model 4 from 5.6 above where we see that the male (male ==1) indicator has a coefficient of 3.085, indicating a higher baseline average temperature for males.

Therefore, in control, females have a better outcome as opposed to males.

```
# mod_male <- d[ male != 0 , lm(temperature_day14 ~
#       treat_zmapp  + temperature_day0 + dehydrated_day0 )]
# mod_female <- d[ male == 0 , lm(temperature_day14 ~
#       treat_zmapp  + temperature_day0 + dehydrated_day0 )]

# creating dummy prediction data - with 100F baseline temp for all
prediction_data_frame <- data.table(
  id = 1:8,
  male               = c(0, 0, 0, 0, 1, 1, 1, 1),
  dehydrated_day0    = c(1, 1, 0, 0, 1, 1, 0, 0),
  treat_zmapp        = c(1, 0, 1, 0, 1, 0, 1, 0),
  temperature_day0  =  c(100,   100,   100,   100,  100,  100,  100, 100)

)

prediction_data_frame[, pred := predict(
  mod_4, newdata = prediction_data_frame  )]

# displaying predicted outcomes
prediction_data_frame
```

```
##    id male dehydrated_day0 treat_zmapp temperature_day0  pred
## 1:  1    0               1           1              100  99.0
## 2:  2    0               1           0              100  99.2
## 3:  3    0               0           1              100  99.0
## 4:  4    0               0           0              100  99.2
## 5:  5    1               1           1              100 100.0
## 6:  6    1               1           0              100 102.3
## 7:  7    1               0           1              100 100.0
## 8:  8    1               0           0              100 102.3
```

We run predictions to prove the treatment effects on zmap on males and females. We run our

prediction with a universal baseline temperature of 100F for all males and females and present results above.

```
# displaying means of predicted outcomes
prediction_data_frame[, mean(pred), by=.(male, treat_zmapp)]
```

```
##    male treat_zmapp    V1
## 1:    0           1  99.0
## 2:    0           0  99.2
## 3:    1           1 100.0
## 4:    1           0 102.3
```

Finally, as shown above in summary, we see that the treatment effect on males has lead to an overall reduction in temperature by 2F. Therefore, we prove that the treatment has worked in reducing average temperatures in males (thus proving HTE of -2.077 for males)

8. Suppose that you had not run the regression in part (7). Instead, you speak with a colleague to learn about heterogeneous treatment effects. This colleague has access to a non-anonymized version of the same dataset and reports that they looked at heterogeneous effects of the ZMapp treatment by each of 80 different covariates to examine whether each predicted the effectiveness of ZMapp on each of 20 different indicators of health. Across these regressions your colleague ran, the treatment's interaction with gender on the outcome of temperature is the only heterogeneous treatment effect that he found to be statistically significant. They reason that this shows the importance of gender for understanding the effectiveness of the drug, because nothing else seemed to indicate why it worked. Bolstering your colleague's confidence, after looking at the data, they 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 your colleague's 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 conceptual can be answered without performing any computation.)

   This is a multiple-comparisons problem because the colleague tried their fishing expedition on as many variables and specifications; exhausting all possible covariates until they got statistical significant results on the coefficients they liked. Although some amount of searching is inevitable, but a result obtained out of a fishing expedition like this (on 80 covariates, on 20 different indicators) seems more like a coincidence rather than an actual discovery of truth. Therefore, it is likely that the statistical significance obtained on the ZMapp treatment on men is without merit and not true.

9. Now, imagine that your colleague's fishing expedition 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?

   If an independent investigation is performed only on the heterogenous treatment effect of ZMapp and gender - with no influence from the colleague's fishing expedition, then we are inclined to believe the results of our independently performed regression analysis.

10. Now, imagine that your colleague's fishing expedition **did** happen, but that you on your own tested this and only this HTE, discover a positive result and conclude there is an effect. How does your colleague's behavior change the interpretation of your test? Does this seem fair or reasonable?

    In this scenario, we are performed our own analysis first and then fishing expedition comes later. It is plausible that we could get influenced by the fishing expedition and get involved in a multiple-comparisions issue. Hence we would need to do the following to corroborate and validate our experiment:

    Need someone who is not involved with the fishing expedition and validate our test

    State our hypothesis a-priori and then test our model only on that hypothesis, not to pick and choose the hypotheses we like

Implement conservative validation techniques by controlling the family-wise error rate (FWER) using techniques like like Bonferroni correction or Holm's method to avoid overstating statistical significance.

Controlling false-discovery rate – by allowing more false positives and,

Using Bayesian solutions by sufficiently modeling the relationships between corresponding parameters of the model from the beginning