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.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(data.table)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(lmtest)
library(sandwich)
library(data.table)
library(stargazer)
library(ggplot2)
library(lme4)
```

```
## Loading required package: Matrix
```

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

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

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

```
robust_se <- function(mod){
  # Adjust standard errors
  cov1  <- vcovHC(mod, type = "HC1")
  return(sqrt(diag(cov1)))
}
```

```
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
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),syn
  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 %
## -5.121351  3.161774
```

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

```
##
## Call:
## lm(formula = name_recall ~ treat_ad)
##
## 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
```

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 0.6426573 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
##  0.01849151  0.02375363
```

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
##  0.01818893  0.03550334
```

```
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)  0.4177709  0.4906211
## treat_ad    -0.2074875 -0.1026589
```

```
clustered_se(mod_pooled)
```

```
## (Intercept)    treat_ad
##  0.01857624  0.02673048
```

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

```
##              Estimate Std. Error    t value       p_value sig
## (Intercept)  0.4541960 0.01218920 37.262160 1.482473e-245 ***
## treat_ad    -0.1550732 0.01876226 -8.265167  2.160639e-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. Furthermore, the treatment assignment is not fully random (you shouldn't take out the study that someone else did), because of the timing issue with 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.

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)         0.180684806 0.01599385 11.2971428   5.992829e-29 ***
## treat_ad           -0.006775249 0.01817672 -0.3727432   7.093688e-01
## as.factor(studyno)2 0.426098820 0.01795546 23.7308813 3.195530e-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         322.848*** (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: 0.4260988 and $3.1955299 \times 10^{-113}$
- treatment effect estimate and p-value: -0.0067752 and 0.7093688

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)
model_metrics_df <- coeftest_to_df(coeftest(mod_interaction, mod_interaction$vcovCL1_))

model_metrics_df
```

```
##                                 Estimate Std. Error    t value      p_value
## (Intercept)                   0.182468694 0.01853379  9.8451922 1.713013e-22
## treat_ad                     -0.009797887 0.02412534 -0.4061243 6.846835e-01
## as.factor(studyno)2           0.423319729 0.02313299 18.2993960 1.318995e-70
## treat_ad:as.factor(studyno)2  0.006994538 0.03669948  0.1905896 8.488615e-01
##                                 sig
## (Intercept)                     ***
## treat_ad
## as.factor(studyno)2             ***
## treat_ad:as.factor(studyno)2
```

```
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             215.167*** (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 0.8488615 which doesn't indicate a statistically significant heterogeneous treatment effect.