Assignment 5 - COMM 290

Lisette del Pino

Goal: Evaluate whether Brett Kavanaugh’s confirmation to the Supreme Court of US, in light of the allegations of sexual assault, eroded the public confidence in the Supreme Court. We are particularly interested in investigating the possible erosion of trust among Democratic voters and women.

The survey recorded people’s trust in the Supreme Court directly before and directly after the confirmation of B. Kavanaugh.

# Question 1

First load in the bk.dta data.

library(tidyverse)

## -- Attaching packages ------------------------------------ tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts --------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(haven) # to read dta  
library(labelled) # to manipulate metadata  
library(psych) # for use of Croenbach's Alpha for inter-item reliability

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

library(stargazer) # for regression tables

## Warning: package 'stargazer' was built under R version 4.0.3

##   
## 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

#bk = Brett K.  
bk<-read\_dta("bk.dta")

Each column tracks a rated response to a question regarding trust in the Supreme Court, or the gender / political learning of the survey responder.

### Step 1: Create the Outcome Variable

Create the outcome variable - “trust in the Supreme Court” - using the follow up survey questions coded in a21,a22,a23.

First I recoded the three post-decision survey questions to fit within 0 and 1, where 0 means that they strongly disagree and 1 means that they strongly agree (so 0 would imply a lower confidence in the Supreme Court while 1 would imply the highest confidence in the Supreme court, post- Brett K.’s appointment)

# need to recode to be between 0 and 1  
bk$post\_trust\_1 <- case\_when(bk$a21 >= 1 & bk$a21 <= 7 ~ (7-bk$a21)/6)  
  
  
bk$post\_trust\_2 <- case\_when(bk$a22 >= 1 & bk$a22 <= 7 ~ (7-bk$a22)/6)  
# from the code book this one might be negatively correlated with trust in   
# the supreme court so I will reverse it.   
bk$post\_trust\_2 <- 1 - bk$post\_trust\_2  
  
  
bk$post\_trust\_3 <- case\_when(bk$a23 >= 1 & bk$a23 <= 7 ~ (7-bk$a23)/6)  
  
# then need to do a Croenbach's Alpha inter-reliability analysis  
alpha(select(bk,post\_trust\_1,post\_trust\_2,post\_trust\_3))

## Number of categories should be increased in order to count frequencies.

##   
## Reliability analysis   
## Call: alpha(x = select(bk, post\_trust\_1, post\_trust\_2, post\_trust\_3))  
##   
## raw\_alpha std.alpha G6(smc) average\_r S/N ase mean sd median\_r  
## 0.67 0.7 0.73 0.44 2.3 0.009 0.6 0.2 0.25  
##   
## lower alpha upper 95% confidence boundaries  
## 0.66 0.67 0.69   
##   
## Reliability if an item is dropped:  
## raw\_alpha std.alpha G6(smc) average\_r S/N alpha se var.r med.r  
## post\_trust\_1 0.39 0.39 0.25 0.25 0.65 0.0174 NA 0.25  
## post\_trust\_2 0.92 0.92 0.85 0.85 11.01 0.0024 NA 0.85  
## post\_trust\_3 0.35 0.35 0.21 0.21 0.55 0.0185 NA 0.21  
##   
## Item statistics   
## n raw.r std.r r.cor r.drop mean sd  
## post\_trust\_1 3882 0.84 0.87 0.88 0.64 0.59 0.24  
## post\_trust\_2 3876 0.66 0.62 0.26 0.24 0.62 0.29  
## post\_trust\_3 3881 0.86 0.88 0.89 0.66 0.59 0.25

Since Cronbach’s raw alpha value is , these three variables are okay indicators of the same thing - public confidence (at least in the sample) of the Supreme Court - but they are not great indicators of confidence.

I suspect that value for these three items is not at least because question 2 () in the survey isn’t as strong an indicator of a low level of confidence in the Supreme Court as it is perhaps a willingness to change the fundamental structure of the system of government in the US.

To make the index, let’s take the row means.

bk$post\_trust\_final <- rowMeans(select(bk,post\_trust\_1,post\_trust\_2,post\_trust\_3))

This column post\_trust\_final indexes the overall positive public opinion in the Supreme Court (for this sample) after the Brett Kavanaugh’s confirmation, rated on a scale of 0 to 1.

# Question 2:

Test the hypothesis that Kavanaugh’s confirmation undercut trust in the Court, by comparing those responded just before the confirmation (cond=0) and just after it (cond=1) using a regression model. Report the point estimates of the intercept and slope coefficients, their standard errors, t-value, p-value, and 95% confidence intervals. Describe the key findings in 2-3 sentences. (2 points)

The model has binary independent/predictor variable cond, with cond = 0 meaning the person was interviewed before confirmation, and cond = 1 meaning the person was interviewed after the confirmation (before and after treatment)

m1<-lm(post\_trust\_final~factor(cond),bk)  
m1

##   
## Call:  
## lm(formula = post\_trust\_final ~ factor(cond), data = bk)  
##   
## Coefficients:  
## (Intercept) factor(cond)1   
## 0.61520 -0.02505

# think i need to make it a factor variable, not continuous.   
  
# negative correlation between time surveyed and trust  
summary(m1)

##   
## Call:  
## lm(formula = post\_trust\_final ~ factor(cond), data = bk)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.61520 -0.11520 0.02096 0.13207 0.40985   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.615198 0.004611 133.413 < 2e-16 \*\*\*  
## factor(cond)1 -0.025047 0.006464 -3.875 0.000108 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2012 on 3873 degrees of freedom  
## (883 observations deleted due to missingness)  
## Multiple R-squared: 0.003862, Adjusted R-squared: 0.003605   
## F-statistic: 15.01 on 1 and 3873 DF, p-value: 0.0001084

confint(m1)

## 2.5 % 97.5 %  
## (Intercept) 0.60615753 0.62423892  
## factor(cond)1 -0.03772036 -0.01237409

#### Least Squares Estimates for the Fitted Line:

Notice we have a negative slope, corresponding to a decrease in trust for those surveyed after the confirmation of Brett Kavanaugh.

* Intercept:0.6151982
* Slope:-0.0250472

#### Standard Error of the Estimates and t value/p value:

Used to calculate the p-value/ Confidence Interval:

* Standard Error of intercept: 0.004611
* Standard Error of slope: 0.006464
* t-val of intercept: 133.413
* t-val of slope: -3.875 #### p values of intercept and slope for the model:

Less than a statistically significant threshold tells us we can reject .

* p-val of intercept: < 2e-16
* p-val of slope: 0.000108

#### Confidence Interval (95%)

Gives us a range of values that the true parameter will be in 95% of the time (not the case that there is a 95% chance that the true parameter is within this range):

* 95% CI of intercept: (Intercept) 0.60615753 0.62423892
* 95% CI of slope: factor(cond)1 -0.03772036 -0.01237409

### Results for Question 2:

From these results, particularly the p-value of the slope (), we get that we must reject the null hypothesis that the confirmation of Brett Kavanaugh did not erode the public confidence in the Supreme Court. Since the p value is less than , we can be confident in rejecting the null hypothesis .

We also note that the true parameter for the intercept and slope will be in the range of values listed for the 95% confidence interval exactly 95% of the time, so the slope will be in between -0.03772036 and -0.01237409 95% of the time. I am not claiming that there is 95% chance that the trust slope is in between these two numbers.

# Question 3

Create another variable measuring people’s baseline trust in the Supreme Court measured at the baseline (sc1, sc2, sc3). Then control for this variable (i.e., lagged DV) in each of the models. Briefly comment on how the point estimates and standard errors change as a result. (2 points)

First I want to recode the variable to be between 0 and 1, where 0 represents the least trust in the supreme court (corresponds to 7, the “strongly agree” response to the questions measuring distrust in the Supreme Court. 1 represents the other end of the scale, namely the “strongly disagree” response to the questions measuring distrust in the Supreme Court.

For the final index, the scores closer to 0 represent a lower confidence in the Supreme Court while the scores closer to 1 represent a higher confidence in the Supreme Court.

# need to recode to be between 0 and 1  
bk$base\_trust\_1 <- case\_when(bk$sc1 >= 1 & bk$sc1 <= 7 ~ (7-bk$sc1)/6)  
  
  
bk$base\_trust\_2 <- case\_when(bk$sc2 >= 1 & bk$sc2 <= 7 ~ (7-bk$sc2)/6)  
# from the code book this one might be negatively correlated with trust in   
# the supreme court so I will reverse it.   
bk$base\_trust\_2 <- 1 - bk$base\_trust\_2  
  
  
bk$base\_trust\_3 <- case\_when(bk$sc3 >= 1 & bk$sc3 <= 7 ~ (7-bk$sc3)/6)  
  
  
alpha(select(bk,base\_trust\_1,base\_trust\_2,base\_trust\_3))

## Number of categories should be increased in order to count frequencies.

##   
## Reliability analysis   
## Call: alpha(x = select(bk, base\_trust\_1, base\_trust\_2, base\_trust\_3))  
##   
## raw\_alpha std.alpha G6(smc) average\_r S/N ase mean sd median\_r  
## 0.63 0.66 0.7 0.39 1.9 0.01 0.6 0.19 0.21  
##   
## lower alpha upper 95% confidence boundaries  
## 0.61 0.63 0.65   
##   
## Reliability if an item is dropped:  
## raw\_alpha std.alpha G6(smc) average\_r S/N alpha se var.r med.r  
## base\_trust\_1 0.34 0.34 0.21 0.21 0.52 0.019 NA 0.21  
## base\_trust\_2 0.90 0.90 0.81 0.81 8.77 0.003 NA 0.81  
## base\_trust\_3 0.27 0.27 0.16 0.16 0.37 0.021 NA 0.16  
##   
## Item statistics   
## n raw.r std.r r.cor r.drop mean sd  
## base\_trust\_1 4758 0.82 0.85 0.85 0.59 0.59 0.23  
## base\_trust\_2 4753 0.64 0.59 0.21 0.19 0.61 0.28  
## base\_trust\_3 4756 0.85 0.87 0.87 0.63 0.60 0.24

# the alphas might be so low because question 2 might not be a great indicator  
# of distrust in the SC  
bk$base\_trust\_final <- rowMeans(select(bk,base\_trust\_1,base\_trust\_2,base\_trust\_3))

Now we use lagged dependent variable to control for the baseline trust variable.

# recall model 1, a simple linear regression  
m1

##   
## Call:  
## lm(formula = post\_trust\_final ~ factor(cond), data = bk)  
##   
## Coefficients:  
## (Intercept) factor(cond)1   
## 0.61520 -0.02505

# now [lagged DV] method gives us:  
m2<-lm(post\_trust\_final~factor(cond)+base\_trust\_final,bk)  
m2

##   
## Call:  
## lm(formula = post\_trust\_final ~ factor(cond) + base\_trust\_final,   
## data = bk)  
##   
## Coefficients:  
## (Intercept) factor(cond)1 base\_trust\_final   
## 0.09894 -0.01717 0.84630

summary(m2)

##   
## Call:  
## lm(formula = post\_trust\_final ~ factor(cond) + base\_trust\_final,   
## data = bk)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.69299 -0.06048 0.00785 0.06762 0.67884   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.098939 0.006594 15.005 < 2e-16 \*\*\*  
## factor(cond)1 -0.017170 0.003793 -4.527 6.16e-06 \*\*\*  
## base\_trust\_final 0.846304 0.009857 85.855 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1179 on 3865 degrees of freedom  
## (890 observations deleted due to missingness)  
## Multiple R-squared: 0.6574, Adjusted R-squared: 0.6572   
## F-statistic: 3708 on 2 and 3865 DF, p-value: < 2.2e-16

stargazer(m1,m2, type = 'text', style = 'ajps')

##   
## -----------------------------------------------------------------------  
## post\_trust\_final   
## Model 1 Model 2   
## -----------------------------------------------------------------------  
## factor(cond)1 -0.025\*\*\* -0.017\*\*\*   
## (0.006) (0.004)   
## base\_trust\_final 0.846\*\*\*   
## (0.010)   
## Constant 0.615\*\*\* 0.099\*\*\*   
## (0.005) (0.007)   
## N 3875 3868   
## R-squared 0.004 0.657   
## Adj. R-squared 0.004 0.657   
## Residual Std. Error 0.201 (df = 3873) 0.118 (df = 3865)   
## F Statistic 15.015\*\*\* (df = 1; 3873) 3707.716\*\*\* (df = 2; 3865)  
## -----------------------------------------------------------------------  
## \*\*\*p < .01; \*\*p < .05; \*p < .1

We can interpret these coefficients in this way. represents the predicted mean when all predicting variables are equal to zero (so this is the predicted mean for those people that were interviewed before the confirmation and also for the people who scored a 0, meaning low opinion, in the baseline opinion measure). The base\_trust\_ final coefficient, , shows that those who had a higher opinion of the Supreme Court in the baseline retained that higher opinion in the follow-up.

The interesting coefficient change here happens with cond, representing the mean difference in trust between those in the control group and treatment group. The standard errors for cond in the new model, 0.003793 for the slope, is less than the standard error for the slope for cond in model 1 (0.006464), it’s almost halved. This means that the coefficients in the second model for cond will be closer to the population’s true mean. Indeed, we see the new mean difference between the control and treatment group, -0.017170, is still negative but shallower once we control for the baseline trust index. So controlling for baseline trust yields a shallower slope on the line of best fit on the model, meaning the effect of the treatment is not as strong. (-0.025047 for model 1 slope vs -0.017170 for model 2).

# Question 4

Investigate whether the effect was stronger among Democrats. Also investigate whether the effect was stronger among women. Report the point estimates of the intercept and slope coefficients, their standard errors, t-value, p-value, and 95% confidence intervals. Describe the key findings in 3-4 sentences. (2 points)

Part 1: To see if the effect was stronger among Democrats, need to observe the interactions in the model. First I made political orientation a binary condition (1 if the respondent identified as a Democrat, 0 if they identified as not a Democrat). The female binary variable was already coded for us.

# is dem is 0 if not democrat, 1 if democrat.   
bk$is\_dem<-case\_when(((bk$pid >=4) & (bk$pid <=6)) ~ 1,  
 ((bk$pid < 4) | (bk$pid > 6)) ~ 0)  
m\_dem\_test<-lm(post\_trust\_final~factor(cond),subset(bk,is\_dem==1))  
coef(m\_dem\_test)

## (Intercept) factor(cond)1   
## 0.58795021 -0.04236872

m\_not\_dem\_test<-lm(post\_trust\_final~factor(cond),subset(bk,is\_dem==0))  
coef(m\_not\_dem\_test)

## (Intercept) factor(cond)1   
## 0.651956760 -0.003635743

m\_dem<-lm(post\_trust\_final~factor(cond)\*is\_dem,bk)  
summary(m\_dem)

##   
## Call:  
## lm(formula = post\_trust\_final ~ factor(cond) \* is\_dem, data = bk)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.65196 -0.10114 0.01471 0.13427 0.45442   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.651957 0.006901 94.475 < 2e-16 \*\*\*  
## factor(cond)1 -0.003636 0.009630 -0.378 0.70580   
## is\_dem -0.064007 0.009118 -7.020 2.61e-12 \*\*\*  
## factor(cond)1:is\_dem -0.038733 0.012766 -3.034 0.00243 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1966 on 3868 degrees of freedom  
## (886 observations deleted due to missingness)  
## Multiple R-squared: 0.04851, Adjusted R-squared: 0.04777   
## F-statistic: 65.73 on 3 and 3868 DF, p-value: < 2.2e-16

confint(m\_dem)

## 2.5 % 97.5 %  
## (Intercept) 0.63842719 0.66548633  
## factor(cond)1 -0.02251660 0.01524511  
## is\_dem -0.08188217 -0.04613093  
## factor(cond)1:is\_dem -0.06376098 -0.01370498

## Treatment effects (Tracking If Democrat):

#### Coefficients:

* Intercept of Model: $ 0.651957$
* Slope of Model:

Notice that this is statistically significant, so must reject null hypothesis that being a Democrat is not a moderating effect on the effect that Kavanaugh’s confirmation had on the public opinion of the Supreme Court.

#### Standard error of Estimates and t values:

* standard error of intercept:
* standard error of slope:
* t value of intercept:
* t value of slope: -3.034

#### p values:

-p value of intercept: < 2e-16

-p value of slope: 0.00243

#### 95% CI:

* CI of intercept: 0.63842719 to 0.66548633
* CI of slope: -0.06376098 to -0.01370498

#### Results for Democrat Interaction model:

We can see that identifying as a Democrat is a moderating factor in the the effect Kavanaugh’s confirmation had on trust in the Supreme Court. We know this because the p-value for the slope tracking the interactions is statistically significant(less than actually). From the confidence interval, we can see that the true population parameter is within -0.06376098 and -0.01370498 95% of the time.

# is women model  
m\_women<-lm(post\_trust\_final~factor(cond)\*female,bk)  
summary(m\_women)

##   
## Call:  
## lm(formula = post\_trust\_final ~ factor(cond) \* female, data = bk)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.62748 -0.12748 0.00973 0.14167 0.41945   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.627483 0.006673 94.027 < 2e-16 \*\*\*  
## factor(cond)1 -0.026099 0.009374 -2.784 0.00539 \*\*   
## female -0.023253 0.009247 -2.515 0.01195 \*   
## factor(cond)1:female 0.002416 0.012966 0.186 0.85218   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2009 on 3845 degrees of freedom  
## (909 observations deleted due to missingness)  
## Multiple R-squared: 0.006814, Adjusted R-squared: 0.006039   
## F-statistic: 8.793 on 3 and 3845 DF, p-value: 8.281e-06

confint(m\_women)

## 2.5 % 97.5 %  
## (Intercept) 0.61439957 0.640567322  
## factor(cond)1 -0.04447779 -0.007720269  
## female -0.04138192 -0.005124731  
## factor(cond)1:female -0.02300521 0.027837912

## Treatment effects (Tracking If Woman):

#### Coefficients:

* Intercept of Model:
* Slope of Model:

Notice that this is not statistically significant, so we can not reject null hypothesis.

#### Standard error of Estimates and t values:

* standard error of intercept:
* standard error of slope:

#### p values:

* p values of intercept:
* p values of slope:

#### 95% CI:

* CI of intercept:
* CI of slope:

#### Results (Heterogenous Effects - Women):

From this, we cannot reject the null hypothesis that being a woman did not have an effect on the effect of Kavanaugh’s confirmation in public trust of the Supreme Court.

## Overall Results:

Overall, being a Democrat led the treatment to have a much stronger effect on the treatment eroding the respondent confidence in the Supreme Court.

# Question 5:

Question 5: Use the Penn student survey data to answer your group’s research question. The data file penn.dta will be uploaded on 11/21. First, propose two specific, testable hypotheses bearing on your general research question. Each hypothesis should state your prediction about the relationship between two variables. For example, if your research question was “which factors are related to Y”, you should name two specific factors (e.g., gender, ethnicity, etc.). Run regression models to test your hypotheses. Report the point estimates of the intercept and slope coefficients, their standard errors, t-value, p-value, and 95% confidence intervals. Describe what these numbers mean in a few sentences. Finally, offer a succinct answer to your research question (2 points)

* note: I ended up using the penn.sav file uploaded recently, not the penn.dta file uploaded in April 2020.

Our team wanted to investigate whether social media engagement rose after the start of the pandemic in the US (March 2020)

Hypothesis 1: Those who went on more social media sites pre-pandemic were likely to go on more social media sites after the start of the pandemic. Hypothesis 2: If the respondent selected more social media sites as sites they usually visit, they believed they spent too much time on social media.

penn<-read\_sav("penn.sav")  
setwd("/Users/Lizard Empress/Documents/Code/r-statistical-inference")  
  
head(penn)

## # A tibble: 6 x 68  
## q2 q3 q4 q5 q6 q7 q8 q9 q10\_1  
## <dbl+l> <dbl+lb> <dbl> <dbl+lb> <dbl> <dbl+l> <dbl+l> <dbl+l> <dbl+lb>  
## 1 3 [Eng~ NA NA 12 [Sys~ NA 3 [Jun~ 1 [Mal~ 1 [Whi~ 1 [Fac~  
## 2 1 [Col~ 13 [ Co~ NA NA NA 2 [Sop~ 2 [Fem~ 6 [Oth~ NA   
## 3 1 [Col~ 8 [ Bi~ NA NA NA 3 [Jun~ 2 [Fem~ 6 [Oth~ 1 [Fac~  
## 4 3 [Eng~ NA NA 6 [Com~ NA 3 [Jun~ 1 [Mal~ 1 [Whi~ 1 [Fac~  
## 5 3 [Eng~ NA NA 10 [Mec~ NA 4 [Sen~ 1 [Mal~ 2 [His~ 1 [Fac~  
## 6 1 [Col~ 14 [ Co~ NA NA NA 3 [Jun~ 1 [Mal~ 5 [Asi~ 1 [Fac~  
## # ... with 59 more variables: q10\_2 <dbl+lbl>, q10\_3 <dbl+lbl>,  
## # q10\_4 <dbl+lbl>, q10\_5 <dbl+lbl>, q10\_6 <dbl+lbl>, q10\_7 <dbl+lbl>,  
## # q10\_8 <dbl+lbl>, q11\_1 <dbl>, q12\_1 <dbl+lbl>, q12\_2 <dbl+lbl>,  
## # q12\_3 <dbl+lbl>, q12\_4 <dbl+lbl>, q12\_5 <dbl+lbl>, q12\_6 <dbl+lbl>,  
## # q12\_7 <dbl+lbl>, q13\_1 <dbl>, q14 <dbl+lbl>, q15 <dbl+lbl>,  
## # q16\_1 <dbl+lbl>, q16\_2 <dbl+lbl>, q16\_3 <dbl+lbl>, q16\_4 <dbl+lbl>,  
## # q16\_5 <dbl+lbl>, q16\_6 <dbl+lbl>, q16\_7 <dbl+lbl>, q16\_8 <dbl+lbl>,  
## # q17 <dbl>, q18 <dbl+lbl>, q19 <dbl+lbl>, q20 <dbl+lbl>, q21 <dbl+lbl>,  
## # q22 <dbl+lbl>, q23 <dbl+lbl>, q24 <dbl+lbl>, q25 <dbl+lbl>, q26\_1 <dbl>,  
## # q27\_1 <dbl+lbl>, q27\_2 <dbl+lbl>, q27\_3 <dbl+lbl>, q27\_4 <dbl+lbl>,  
## # q28 <dbl+lbl>, q29 <dbl+lbl>, q30\_1 <dbl+lbl>, q30\_2 <dbl+lbl>,  
## # q30\_3 <dbl+lbl>, q30\_4 <dbl+lbl>, q31 <dbl+lbl>, q32 <dbl+lbl>,  
## # q33 <dbl+lbl>, q34 <dbl+lbl>, q35 <dbl+lbl>, q36 <dbl+lbl>, q37 <dbl+lbl>,  
## # q38 <dbl+lbl>, q39 <dbl+lbl>, q40 <dbl+lbl>, q41 <dbl+lbl>, q42 <dbl+lbl>,  
## # q43 <dbl+lbl>

# overall number of sites visited pre-march 2020 and post-march 2020  
penn$num\_sites\_visited\_pre<-rowSums(select(penn, q10\_1:q10\_8),na.rm=T)  
penn$num\_sites\_visited\_post<-rowSums(select(penn, q12\_1:q12\_7),na.rm=T)  
  
summary(penn$num\_sites\_visited\_pre)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 3.0 4.0 3.9 5.0 6.0

summary(penn$num\_sites\_visited\_post)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 3.00 3.50 3.45 4.00 6.00

# sums only go from 1 to 6 in both variables, so recode:  
  
penn$pre\_predictor <- case\_when(penn$num\_sites\_visited\_pre >= 0 ~ (penn$num\_sites\_visited\_pre/6))  
penn$post\_outcome <- case\_when(penn$num\_sites\_visited\_post >= 0 ~ (penn$num\_sites\_visited\_post/6))  
  
# Let's try predicting social media usage after the start of the pandemic based on social media usage before the pandemic  
m\_social\_media<-lm(post\_outcome~pre\_predictor,penn)  
summary(m\_social\_media)

##   
## Call:  
## lm(formula = post\_outcome ~ pre\_predictor, data = penn)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36325 -0.08605 -0.02992 0.08061 0.35781   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.14392 0.07858 1.831 0.0722 .   
## pre\_predictor 0.66320 0.11666 5.685 4.48e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1596 on 58 degrees of freedom  
## Multiple R-squared: 0.3578, Adjusted R-squared: 0.3467   
## F-statistic: 32.32 on 1 and 58 DF, p-value: 4.482e-07

confint(m\_social\_media)

## 2.5 % 97.5 %  
## (Intercept) -0.0133800 0.3012138  
## pre\_predictor 0.4296794 0.8967301

From these statistics:

#### Intercept and Slope

Intercept estimate: 0.14392 Slope estimate: 0.66320

From this, we see that those who visited no social media sites before the start of pandemic are expected to increase their social media usage by 0.14392 after the start of the pandemic. Since there is a positive slope, and a relatively steep one at that, we can say that social media usage certainly increased the more someone used social media before the start of the pandemic. Let’s check the veracity of this claim with our p-value calculation:

#### Standard Error, t values, p values

* standard error of intercept/slope: 0.07858 / 0.11666
* t value for intercept/slope: 1.831 / 5.685
* p value for intercept/slope: 0.0722 / 4.48e-07

#### 95 % CI

The true population parameter will be in this range of values, found from the sample, 95% of the time (ok this is a pretty big confidence interval…)

* CI for intercept: -0.0133800 to 0.3012138
* CI for slope: 0.4296794 to 0.8967301

Since the slope estimate has a statistically significant p value, we can reject the null hypothesis that the more social media sites a respondent used before the start of the pandemic, the more they were likely to increase their usage after the start of the pandemic.

Now to test the second hypothesis: that those who increased their social media site usage over the start of the pandemic were more likely to think they used social media excessively:

penn$site\_diff<-NA # tracking the difference in sites visited  
  
penn$site\_diff<-((penn$num\_sites\_visited\_post) - (penn$num\_sites\_visited\_pre))  
  
# now only want those people who went on more sites after the start of the pandemic, let's track this in a binary variable:  
  
penn$pos\_site\_diff<-case\_when(penn$site\_diff > 0 ~ 1,   
 penn$site\_diff <= 0 ~ 0)  
  
# now let's get the binary outcome variable, did they think they spend too much time on social media (valid responses are are answers 1 and 2 to q14)  
  
penn$too\_much\_time<-case\_when(penn$q14 < 3 ~ 1,  
 penn$q14 >= 3 ~ 0)  
  
m\_time<-lm(too\_much\_time~pos\_site\_diff, penn)  
summary(m\_time)

##   
## Call:  
## lm(formula = too\_much\_time ~ pos\_site\_diff, data = penn)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.8 0.0 0.2 0.2 0.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.80000 0.05252 15.232 <2e-16 \*\*\*  
## pos\_site\_diff 0.20000 0.12865 1.555 0.125   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3714 on 58 degrees of freedom  
## Multiple R-squared: 0.04, Adjusted R-squared: 0.02345   
## F-statistic: 2.417 on 1 and 58 DF, p-value: 0.1255

#stargazer(m\_time, type = "text", style = "ajps")

We get some strange results with this linear model, and I suspect it’s because the sameple size ends up being 60 (after discounting everyone whose site variation did not go up after the start of the pandemic)

#### Estimates for intercept and slope:

* intercept: 0.80000
* slope: 0.20000

#### Standard error, t value, p values, CI:

* standard error for intercept/slope: 0.05252 / 0.12865
* t values for intercept and slope: 15.232 / 1.555
* p values for intercept and slope: <2e-16 / 0.125

The true population parameter will be inside the CI 95% of the time: 0.6948646, -0.057528, 0.9051354, 0.457528 CI for intercept and slope here.

Since the slope did not have a statistically significant p value, we cannot reject the null hypothesis, which states that those who increased the variety of sites visited since the start of the pandemic did not find that they spend too much time on social media.

Thus, my team’s research question “Did social media usage increase after the start of the pandemic” remains unanswered. On the one hand, it is clear that those who visited a wider variety of sites before the start of the pandemic were likely to increase the variety of sites visited after March 2020, but it is not clear whether overall usage increased. More analysis needs to be done. I would like to measure average increase in number of hours in the future.