Data Science HW2 Total: 20 points Due: January 18th, 2019, NOON 1.

1. **In your own words, specify the research question(s) (2 sentences or less) for both studies [4]**

Study 1: Did voter intention change as a result of the Ebola outbreak?

Study 2: Did state specific voter intention change as a result of the Ebola outbreak?

1. **Include and explain IVs, DVs for both studies. [4]**

For study 1, the independent variable was date and the dependent variables were the voter intention index scores. For study 2, the independent was the date (before vs after ebola outbreak) and the dependent variable was the state-specific voter-intention difference score.

1. **In your own words, why did the authors run a regression for this particular analysis? [4]**

The authors ran a regression to determine the importance of each contributing variable (PVI, pre-outbreak polling difference) on the dependent variable (state-specific voter-intention difference score).

#libraries

library(tidyverse)

library(ggplot2)

library(tidyr)

library(dplyr)

library(lubridate)

library(effsize)

library(powerAnalysis)

```

```{r}

study\_02 <- read.csv("Study2.csv")

study\_01 <- read.csv("Study1.csv")

study\_01$date = paste('2015', study\_01$month, study\_01$date)

study\_01$date = as.Date(study\_01$date, format = '%Y %B %d')

outbreakDate = as.Date('2015 September 30', format = '%Y %B %d')

```

```{r}

#change state senate race to string

study\_02$StateSenateRace = as.character(study\_02$StateSenateRace)

#exclude outliers

exclude = c("Hawaii", "Rhode Island")

study\_02 = study\_02[!(study\_02$StateSenateRace %in% exclude),]

```

```{r}

#Reproduce all 4 plots on page 599 using ggplot2

#vector of days

#day <- mday(study\_01$date)

sept <- study\_01 %>%

filter(month == "September") %>%

mutate()

oct <- study\_01 %>%

filter(month == "October")

sep24 = as.Date('2015-09-24', format = '%Y-%m-%d')

sep30 = as.Date('2015-09-30', format = '%Y-%m-%d')

sepweek <- sept %>%

filter(date >= '2015-09-24' & date <= '2015-09-30')

oct1 = as.Date('2015-10-01', format = '%Y-%m-%d')

oct7 = as.Date('2015-10-07', format = '%Y-%m-%d')

octweek <- oct%>%

filter(date >='2015-10-01' & date<='2015-10-07')

```

```{r}

#Reproduce all 4 plots on page 599 using ggplot2 (do not worry about axes units accuracy or title of plot)

ggplot(data = sept, aes(x = date, y = VoterIntentionIndex))+

geom\_point() +

labs(title = "a",

subtitle = "September",

y = "Voter-Intention-Index Score",

x = "Day") +

geom\_smooth(method = lm, aes(color="red"), se=F)+

scale\_y\_continuous(breaks = c(-1,-.5,0,.5,1,1.5,2,2.5,3))+

theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),

panel.background = element\_blank(), axis.line = element\_line(colour = "black"),

legend.position="none"

)

ggplot(data = oct, aes(x = date, y = VoterIntentionIndex))+

geom\_point() +

labs(title = "a",

subtitle = "October",

y = "Voter-Intention-Index Score",

x = "Day") +

geom\_smooth(method = lm, aes(color="red"), se=F)+

scale\_y\_continuous(breaks = c(-1,-.5,0,.5,1,1.5,2,2.5,3))+

theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),

panel.background = element\_blank(), axis.line = element\_line(colour = "black"),

legend.position="none"

)

ggplot(data=sepweek, aes(x=date, y = VoterIntentionIndex))+

geom\_jitter()+

geom\_smooth(method = lm, aes(color = "red"), se=F)+

theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),

panel.background = element\_blank(), axis.line = element\_line(colour = "black"),

legend.position="none"

)+

labs(title = "b", subtitle = "September")

ggplot(data=octweek, aes(x=date, y = VoterIntentionIndex))+

geom\_jitter()+

geom\_smooth(method = lm, aes(color = "red"), se=F)+

theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),

panel.background = element\_blank(), axis.line = element\_line(colour = "black"),

legend.position="none")+

labs(title = "b", subtitle = "October")

```

```{r}

#Mean voter-intention difference score t test (bonus 1 pt for producing code for cohen d)

t.test(study\_02$StateSpecificVoterIntentionIndexDifferenceScore, mu = 0)

#(d = cohen.d(), pooled = TRUE, na.rm = TRUE))

```

```{r}

#State-specific voter-intention difference score between Republican vs. Democrat candidate (bonus 1 pt for producing code for cohen d)

study\_02$rep <- study\_02$PollingDifferenceInFinalPollInSeptember > 0

study\_02$dem <- study\_02$PollingDifferenceInFinalPollInSeptember < 0

t.test(study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$rep],

study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$dem], var.equal = TRUE)

cohen.d(study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$rep],

study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$dem], pooled = TRUE, na.rm = T)

#filtering by PVI (last paragraph)

republican <- study\_02 %>%

filter(PVI >= 0, PVI != 0)

democrat <- study\_02 %>%

filter(PVI <= 0, PVI != 0)

t.test(republican$StateSpecificVoterIntentionIndexDifferenceScore,

democrat$StateSpecificVoterIntentionIndexDifferenceScore,

var.equal = TRUE)

```

```{r}

#Regression: difference score, PVI score, polling difference (do not worry about beta values, but your R2, p values should still match the article results).

#y = StateSpecificVoterIntentionIndexDifferenceScore

#x = PVI + PollingDifferenceInFinalPollInSeptember

fitpoll = lm(StateSpecificVoterIntentionIndexDifferenceScore ~ PVI + PollingDifferenceInFinalPollInSeptember , data = study\_02)

fitpoll

summary(fitpoll)

confint(fitpoll)

hw\_02

Abby Bergman

1/15/2019

#libraries  
library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.1.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.7  
## ✔ tidyr 0.8.2 ✔ stringr 1.3.1  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ──────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
library(tidyr)  
library(dplyr)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(effsize)  
library(powerAnalysis)

study\_02 <- read.csv("Study2.csv")  
study\_01 <- read.csv("Study1.csv")  
  
study\_01$date = paste('2015', study\_01$month, study\_01$date)  
  
study\_01$date = as.Date(study\_01$date, format = '%Y %B %d')  
  
outbreakDate = as.Date('2015 September 30', format = '%Y %B %d')

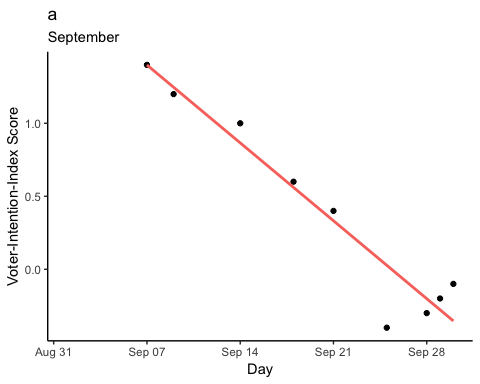
#change state senate race to string  
study\_02$StateSenateRace = as.character(study\_02$StateSenateRace)  
  
#exclude outliers  
exclude = c("Hawaii", "Rhode Island")  
study\_02 = study\_02[!(study\_02$StateSenateRace %in% exclude),]

#Reproduce all 4 plots on page 599 using ggplot2  
#vector of days  
#day <- mday(study\_01$date)  
  
sept <- study\_01 %>%  
 filter(month == "September") %>%  
 mutate()  
oct <- study\_01 %>%  
 filter(month == "October")  
  
sep24 = as.Date('2015-09-24', format = '%Y-%m-%d')  
sep30 = as.Date('2015-09-30', format = '%Y-%m-%d')  
  
sepweek <- sept %>%  
 filter(date >= '2015-09-24' & date <= '2015-09-30')  
  
oct1 = as.Date('2015-10-01', format = '%Y-%m-%d')  
oct7 = as.Date('2015-10-07', format = '%Y-%m-%d')  
  
octweek <- oct%>%  
 filter(date >='2015-10-01' & date<='2015-10-07')

#Reproduce all 4 plots on page 599 using ggplot2 (do not worry about axes units accuracy or title of plot)  
  
ggplot(data = sept, aes(x = date, y = VoterIntentionIndex))+  
 geom\_point() +  
 labs(title = "a",   
 subtitle = "September",  
 y = "Voter-Intention-Index Score",   
 x = "Day") +  
 geom\_smooth(method = lm, aes(color="red"), se=F)+  
 scale\_y\_continuous(breaks = c(-1,-.5,0,.5,1,1.5,2,2.5,3))+  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),  
panel.background = element\_blank(), axis.line = element\_line(colour = "black"),  
legend.position="none"  
 )

## Warning: Removed 21 rows containing non-finite values (stat\_smooth).

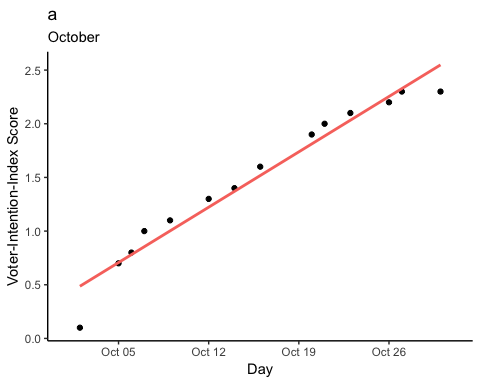
## Warning: Removed 21 rows containing missing values (geom\_point).



ggplot(data = oct, aes(x = date, y = VoterIntentionIndex))+  
 geom\_point() +  
 labs(title = "a",   
 subtitle = "October",  
 y = "Voter-Intention-Index Score",   
 x = "Day") +  
 geom\_smooth(method = lm, aes(color="red"), se=F)+  
 scale\_y\_continuous(breaks = c(-1,-.5,0,.5,1,1.5,2,2.5,3))+  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),  
panel.background = element\_blank(), axis.line = element\_line(colour = "black"),  
legend.position="none"  
 )

## Warning: Removed 17 rows containing non-finite values (stat\_smooth).

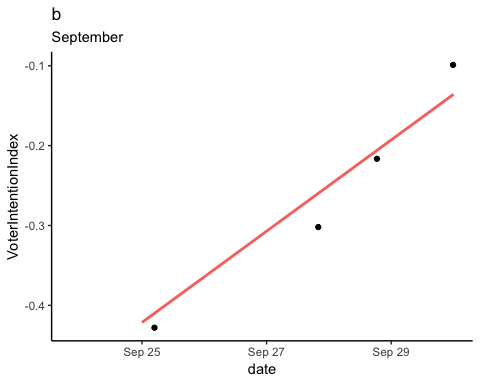
## Warning: Removed 17 rows containing missing values (geom\_point).



ggplot(data=sepweek, aes(x=date, y = VoterIntentionIndex))+  
 geom\_jitter()+  
 geom\_smooth(method = lm, aes(color = "red"), se=F)+  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),  
panel.background = element\_blank(), axis.line = element\_line(colour = "black"),  
legend.position="none"  
 )+  
 labs(title = "b", subtitle = "September")

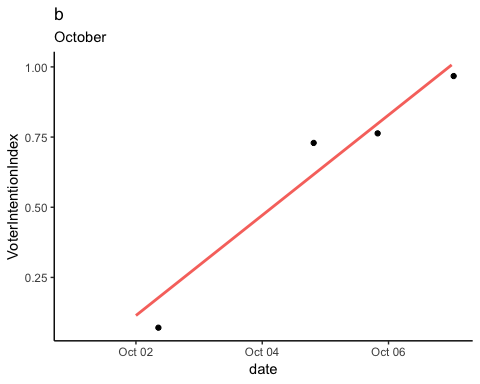
## Warning: Removed 3 rows containing non-finite values (stat\_smooth).

## Warning: Removed 3 rows containing missing values (geom\_point).



ggplot(data=octweek, aes(x=date, y = VoterIntentionIndex))+  
 geom\_jitter()+  
 geom\_smooth(method = lm, aes(color = "red"), se=F)+  
 theme(panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),  
panel.background = element\_blank(), axis.line = element\_line(colour = "black"),   
 legend.position="none")+  
 labs(title = "b", subtitle = "October")

## Warning: Removed 3 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 3 rows containing missing values (geom\_point).



#Mean voter-intention difference score t test (bonus 1 pt for producing code for cohen d)  
  
t.test(study\_02$StateSpecificVoterIntentionIndexDifferenceScore, mu = 0)

##   
## One Sample t-test  
##   
## data: study\_02$StateSpecificVoterIntentionIndexDifferenceScore  
## t = 2.3365, df = 31, p-value = 0.0261  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 0.1294591 1.9074159  
## sample estimates:  
## mean of x   
## 1.018438

#(d = cohen.d(), pooled = TRUE, na.rm = TRUE))

#State-specific voter-intention difference score between Republican vs. Democrat candidate (bonus 1 pt for producing code for cohen d)   
study\_02$rep <- study\_02$PollingDifferenceInFinalPollInSeptember > 0  
  
study\_02$dem <- study\_02$PollingDifferenceInFinalPollInSeptember < 0  
  
t.test(study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$rep],  
 study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$dem], var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$rep] and study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$dem]  
## t = 2.427, df = 30, p-value = 0.02144  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.3281352 3.8118648  
## sample estimates:  
## mean of x mean of y   
## 1.73 -0.34

cohen.d(study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$rep],  
 study\_02$StateSpecificVoterIntentionIndexDifferenceScore[study\_02$dem], pooled = TRUE, na.rm = T)

##   
## Cohen's d  
##   
## d estimate: 0.9033137 (large)  
## 95 percent confidence interval:  
## lower upper   
## 0.1089832 1.6976441

#filtering by PVI (last paragraph)  
republican <- study\_02 %>%  
 filter(PVI >= 0, PVI != 0)  
  
democrat <- study\_02 %>%  
 filter(PVI <= 0, PVI != 0)  
  
t.test(republican$StateSpecificVoterIntentionIndexDifferenceScore,   
 democrat$StateSpecificVoterIntentionIndexDifferenceScore,   
 var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: republican$StateSpecificVoterIntentionIndexDifferenceScore and democrat$StateSpecificVoterIntentionIndexDifferenceScore  
## t = 3.0873, df = 29, p-value = 0.004418  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.8195817 4.0367341  
## sample estimates:  
## mean of x mean of y   
## 1.843158 -0.585000

#Regression: difference score, PVI score, polling difference (do not worry about beta values, but your R2, p values should still match the article results).  
  
#y = StateSpecificVoterIntentionIndexDifferenceScore  
#x = PVI + PollingDifferenceInFinalPollInSeptember  
  
fitpoll = lm(StateSpecificVoterIntentionIndexDifferenceScore ~ PVI + PollingDifferenceInFinalPollInSeptember , data = study\_02)  
fitpoll

##   
## Call:  
## lm(formula = StateSpecificVoterIntentionIndexDifferenceScore ~   
## PVI + PollingDifferenceInFinalPollInSeptember, data = study\_02)  
##   
## Coefficients:  
## (Intercept)   
## 0.338251   
## PVI   
## 0.138099   
## PollingDifferenceInFinalPollInSeptember   
## -0.008727

summary(fitpoll)

##   
## Call:  
## lm(formula = StateSpecificVoterIntentionIndexDifferenceScore ~   
## PVI + PollingDifferenceInFinalPollInSeptember, data = study\_02)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.6622 -1.5346 0.2057 1.5409 4.1313   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 0.338251 0.469067 0.721  
## PVI 0.138099 0.075527 1.828  
## PollingDifferenceInFinalPollInSeptember -0.008727 0.038861 -0.225  
## Pr(>|t|)   
## (Intercept) 0.4766   
## PVI 0.0778 .  
## PollingDifferenceInFinalPollInSeptember 0.8239   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.262 on 29 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.2126, Adjusted R-squared: 0.1583   
## F-statistic: 3.914 on 2 and 29 DF, p-value: 0.03127

confint(fitpoll)

## 2.5 % 97.5 %  
## (Intercept) -0.62109977 1.29760114  
## PVI -0.01637161 0.29256975  
## PollingDifferenceInFinalPollInSeptember -0.08820671 0.07075338

```