

# ImpactDataViz 241 Final Project

```
# load packages
library(data.table)
library(foreign)
library(sandwich)
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)

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.6.2

##
## Attaching package: 'zoo'

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

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.6.2

library(knitr)
library('ggplot2')
library(fastDummies)
```

## Common functions

```

# function to return confidence intervals with robust se
get_confint_robust = function(model, vcovCL) {
  t<-qt(.975, model$df.residual)
  ct<-coeftest(model, vcovCL)
  est<-cbind(ct[,1], ct[,1]-t*ct[,2], ct[,1]+t*ct[,2], ct[,4])
  colnames(est)<-c("Estimate", "LowerCI", "UpperCI", "pValue")
  return(est)
}

# parse out the regression results using robust standard errors
get_regression_results_robust_se = function(model, df, variable_names, showAsTibble) {
  model$vcovHC = vcovHC(model, type="HC1")

  robust_se_all <- sqrt(diag(model$vcovHC))

  est = get_confint_robust(model, model$vcovHC)

  robust_se = c(rep(0, length(variable_names)))
  i = 1
  for (variable_name in variable_names) {
    robust_se_single <- sqrt(diag(model$vcovHC))[variable_name]
    robust_se[i] = robust_se_single
    i = i + 1
  }

  coef = est[variable_names, 'Estimate']
  ci_lower_robust = est[variable_names, 'LowerCI']
  ci_upper_robust = est[variable_names, 'UpperCI']
  p_value = est[variable_names, 'pValue']
  results = data.table(id = variable_names)
  results[, coef := round(coef, 4)]
  results[, ci_lower := round(ci_lower_robust, 4)]
  results[, ci_upper := round(ci_upper_robust, 4)]
  results[, p_value := signif(p_value, 5)]
  results[, robust_se := round(robust_se, 4)]

  if (showAsTibble) {
    print(as_tibble(results))
  }
  return( list('estimates'=results, 'robust_se_all'=robust_se_all))
}

# Multiple plot function
#
# ggplot objects can be passed in ..., or to plotlist (as a list of ggplot objects)
# - cols: Number of columns in layout
# - layout: A matrix specifying the layout. If present, 'cols' is ignored.
#
# If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
# then plot 1 will go in the upper left, 2 will go in the upper right, and
# 3 will go all the way across the bottom.
#
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {

```

```

library(grid)

# Make a list from the ... arguments and plotlist
plots <- c(list(...), plotlist)

numPlots = length(plots)

# If layout is NULL, then use 'cols' to determine layout
if (is.null(layout)) {
  # Make the panel
  # ncol: Number of columns of plots
  # nrow: Number of rows needed, calculated from # of cols
  layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                    ncol = cols, nrow = ceiling(numPlots/cols))
}

if (numPlots==1) {
  print(plots[[1]])
} else {
  # Set up the page
  grid.newpage()
  pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

  # Make each plot, in the correct location
  for (i in 1:numPlots) {
    # Get the i,j matrix positions of the regions that contain this subplot
    matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

    print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                     layout.pos.col = matchidx$col))
  }
}

```

## 1. Parse the survey data into a data.table

```

parse_survey_data = function(filename, treatment_only=FALSE) {
  cat(filename, "\n")
  raw <- fread(filename)

  # covariates
  setnames(raw, 'Q3', 'state')
  setnames(raw, 'Q4', 'gender')
  setnames(raw, 'Q5', 'age')
  setnames(raw, 'Q6', 'ethnicity_multi')
  setnames(raw, 'Q7', 'political_party')
  setnames(raw, 'Q8', 'education')
  setnames(raw, 'Q9', 'covid_sick')
  setnames(raw, 'Q10', 'covid_hospitalized')
  setnames(raw, 'Q11', 'covid_died')

```

```

# duration of survey time
setnames(raw, 'Duration (in seconds)', 'duration_of_survey')

# which block was active? (did the user see the treatment or
# control data viz)
setnames(raw, 'Q15', 'treatment_viz_is_accurate')
if (!treatment_only) {
  setnames(raw, 'Q17', 'control_viz_is_accurate')
}

# outcome questions about COVID attitudes
setnames(raw, 'Q18', 'outcome_spread')
setnames(raw, 'Q19', 'outcome_death')

# which block was active determines if
# subject received treatment data viz or control
# data viz
cat(" number of responses", nrow(raw), '\n')
raw[, treatment := ifelse(is.na(treatment_viz_is_accurate), 0, 1)]
cleaned = raw[!is.na(outcome_spread) & !is.na(outcome_death),]
cat(" number of responses after dropping na", nrow(cleaned), '\n')

# ethnicity allows for multiple choice
# for covariates, just grab the first one
ethnicity_single = rep(0, nrow(cleaned))
i = 1
for (eth_entry in cleaned[, ethnicity_multi]) {
  eth_tokens = unlist(strsplit(eth_entry, ","))
  ethnicity_single[i] = as.numeric(eth_tokens[1])
  i = i + 1
}
cleaned[, ethnicity := ethnicity_single ]

# counts in control vs treatment
n_control = nrow(cleaned[treatment == 0, ])
n_treatment = nrow(cleaned[treatment == 1, ])

cat(" number in treatment", n_treatment, "\n")
cat(" number in control", n_control, "\n\n")

return(cleaned)
}

```

```

# A large run of 260 subjects run on 7/21
run1 <- parse_survey_data("data/run1.2020.07.21.csv")

```

```

## data/run1.2020.07.21.csv
## number of responses 265
## number of responses after dropping na 259
## number in treatment 131
## number in control 128

```

```
run1 = run1[, run := 0]
run1_control = run1[treatment == 0, ]
run1_treatment = run1[treatment == 1, ]
run1[, condition := treatment]

# One small run was done in evening 7/24 treatment only
run2_small <- parse_survey_data("data/run2.small.2020.07.24.csv", TRUE)
```

```
## data/run2.small.2020.07.24.csv
## number of responses 33
## number of responses after dropping na 30
## number in treatment 30
## number in control 0
```

```
run2_small[, run := 2]
run2_small[, control_viz_is_accurate := ""]

# A large run of 270 run on 7/25 treatment and control
run2_large <- parse_survey_data("data/run2.2020.07.25.csv")
```

```
## data/run2.2020.07.25.csv
## number of responses 270
## number of responses after dropping na 262
## number in treatment 130
## number in control 132
```

```
run2_large = run2_large[, run := 2]
run2 = rbind(run2_small, run2_large)
run2[, condition := treatment]
```

```
# Combine the runs on 7/21 and 7/25
combined = rbind(run1, run2)
```

## 2. EDA

### 2.1 Check Duration of Survey

```
show_duration = function(d) {
  cat('duration for control ', mean(d[condition == 0, duration_of_survey]), '\n')
  cat('duration for treatment', mean(d[condition == 1, duration_of_survey]), '\n\n')
}
show_duration(run1)
```

```
## duration for control 153.1875
## duration for treatment 172.2061
```

```
show_duration(run2)
```

```
## duration for control    136.7955  
## duration for treatment 205.5813
```

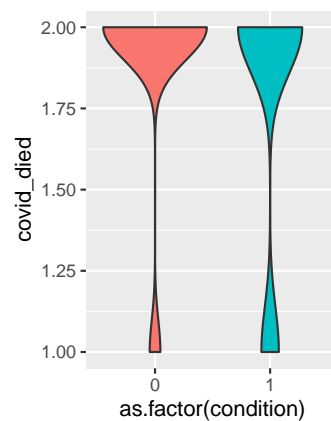
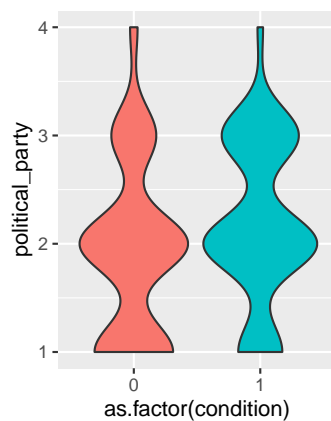
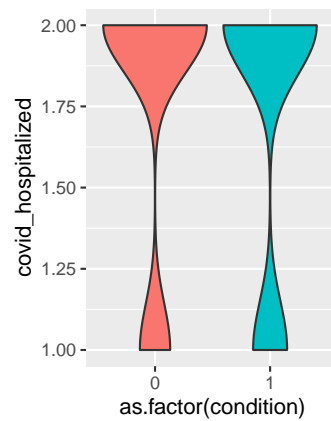
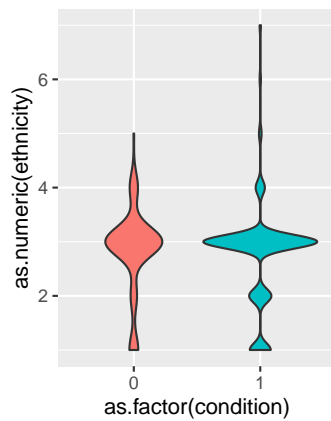
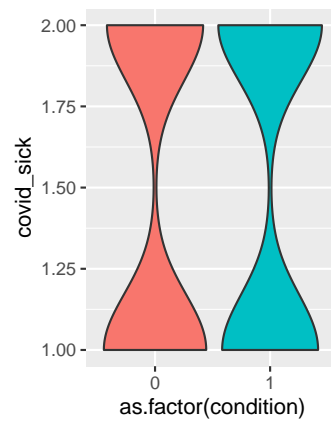
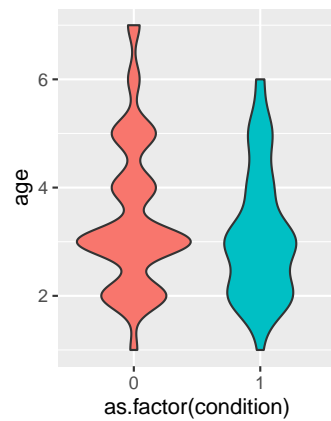
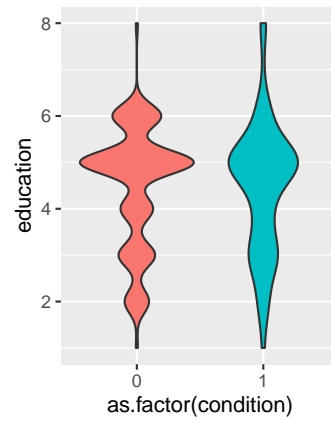
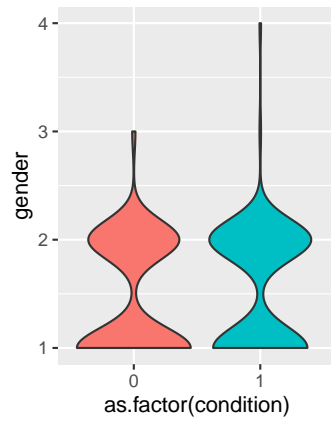
```
show_duration(combined)
```

```
## duration for control    144.8654  
## duration for treatment 190.5567
```

### 3. Covariate balance

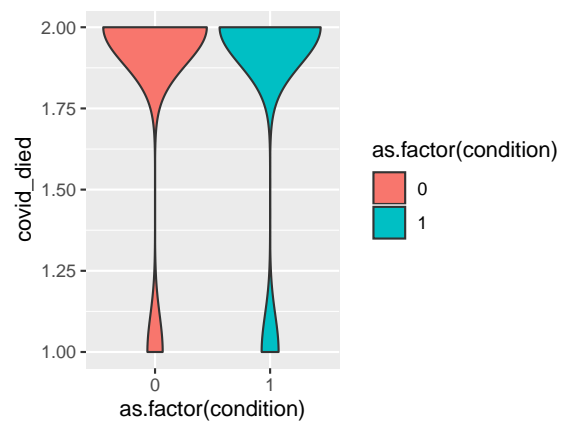
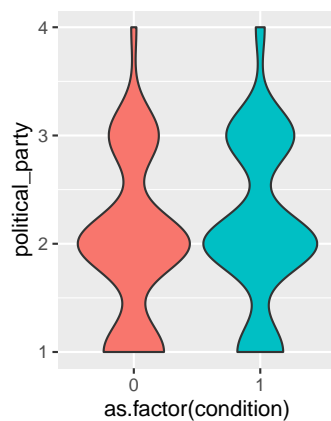
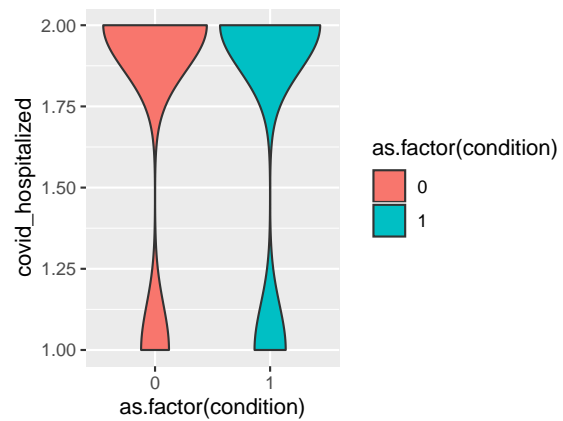
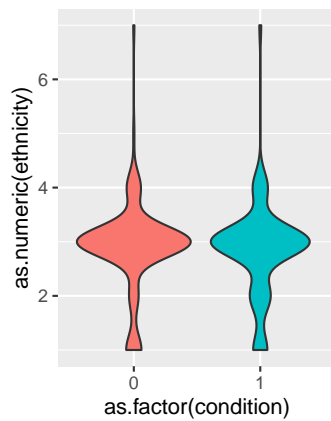
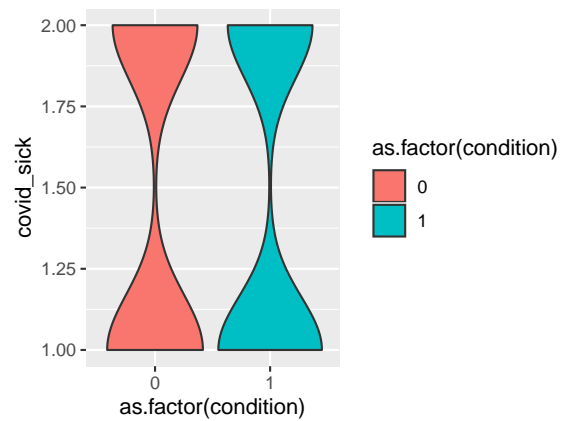
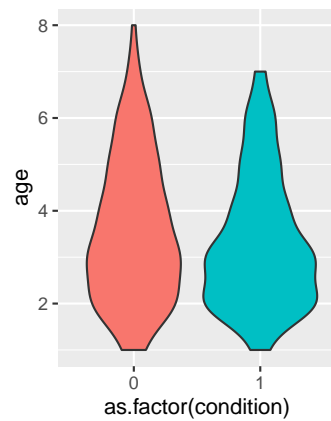
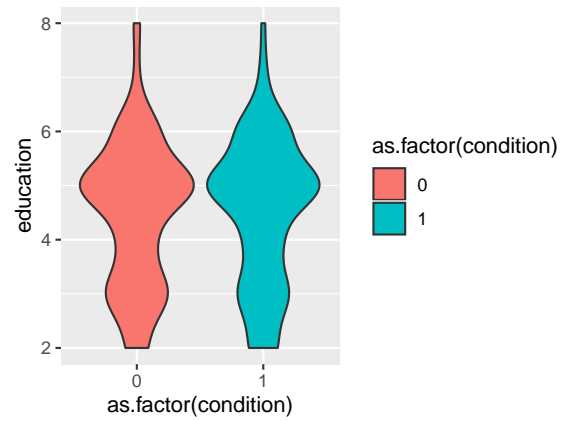
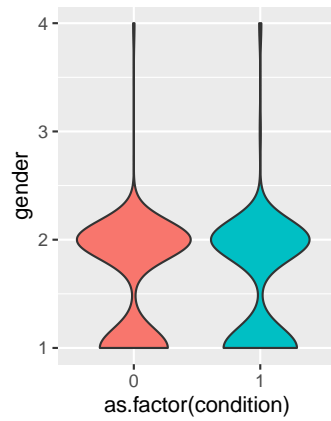
#### 3.1 Compare distributions using violin plots

```
check_covariate_balance = function(d) {  
  options(repr.plot.width = 14, repr.plot.height = 8)  
  
  p1 = ggplot(d, aes(x=as.factor(condition), y=gender, fill=as.factor(condition))) +  
    geom_violin()  
  
  p2 = ggplot(d, aes(x=as.factor(condition), y=age, fill=as.factor(condition))) +  
    geom_violin()  
  
  p3 = ggplot(d, aes(x=as.factor(condition), y=as.numeric(ethnicity), fill=as.factor(condition))) +  
    geom_violin()  
  
  p4 = ggplot(d, aes(x=as.factor(condition), y=political_party, fill=as.factor(condition))) +  
    geom_violin()  
  
  p5 = ggplot(d, aes(x=as.factor(condition), y=education, fill=as.factor(condition))) +  
    geom_violin()  
  
  p6 = ggplot(d, aes(x=as.factor(condition), y=covid_sick, fill=as.factor(condition))) +  
    geom_violin()  
  
  p7 = ggplot(d, aes(x=as.factor(condition), y=covid_hospitalized, fill=as.factor(condition))) +  
    geom_violin()  
  
  p8 = ggplot(d, aes(x=as.factor(condition), y=covid_died, fill=as.factor(condition))) +  
    geom_violin()  
  
  multiplot(p1, p2, p3, p4, p5, p6, p7, p8, cols=2)  
}  
check_covariate_balance(run1)
```

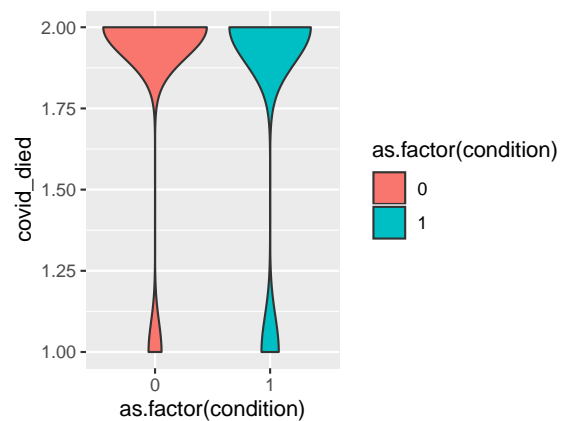
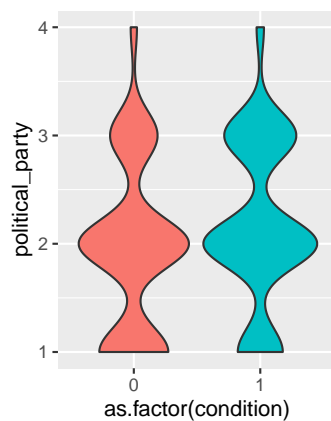
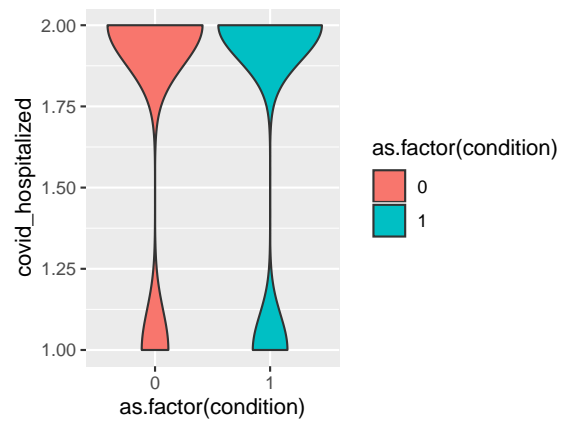
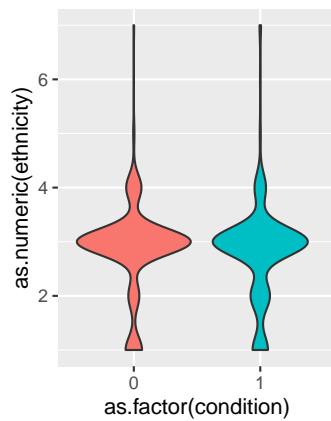
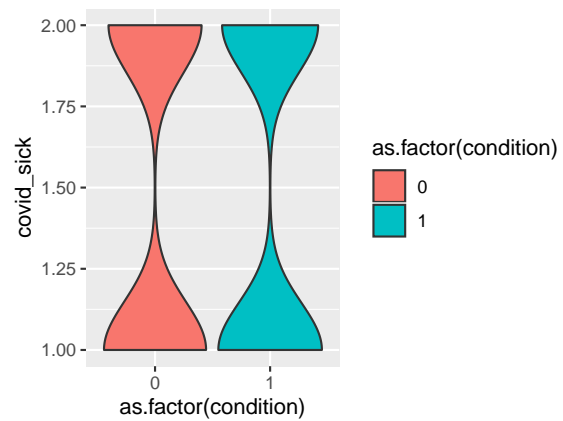
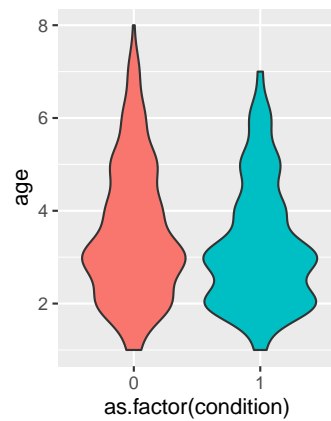
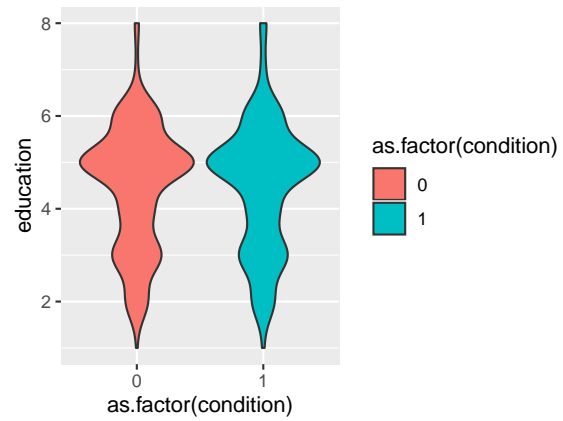
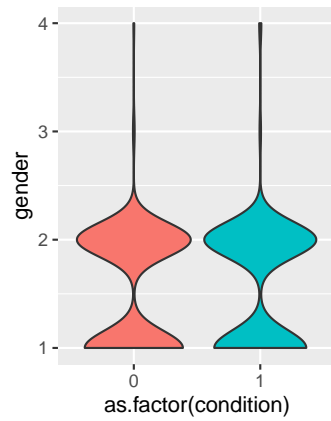


```
check_covariate_balance(run2)
```





```
check_covariate_balance(combined)
```



## 4. Estimate ATE

### 4.1 Estimate the ATE for both outcomes

```
estimate_ate = function(d, outcome_field) {
  g <- d[, .(group_mean = mean(get(outcome_field))), keyby = .(condition)]

  ate <- g[, diff(group_mean)]

  res <- NA
  for (i in 1:10000) {
    res[i] <- d[, .(group_mean = mean(get(outcome_field))), keyby = .(sample(condition))][, diff(group_mean)]
  }
  dist_sharp_null <- res
  #hist(dist_sharp_null)
  #abline(v=ate, lwd=3, col='blue')
  #abline(v=abs(ate), lwd=3, col='blue')
  p_value_one_tailed <- mean(dist_sharp_null >= ate)
  p_value_two_tailed <- mean(abs(dist_sharp_null) >= abs(ate))

  cat(outcome_field, '\n')
  cat(' mean control      ', g[condition == '0', group_mean], '\n')
  cat(' mean treatment     ', g[condition == '1', group_mean], '\n')
  cat(' ATE                  ', ate, '\n')
  cat(' p_value 1-tailed    ', p_value_one_tailed, '\n')
  cat(' p_value 2-tailed    ', p_value_two_tailed, '\n\n')
}

cat('run1', '\n')
```

```
## run1
```

```
cat('*****', '\n')
```

```
## *****
```

```
estimate_ate(run1, 'outcome_spread')
```

```
## outcome_spread
## mean control      3.132812
## mean treatment    3.473282
## ATE                0.3404699
## p_value 1-tailed  5e-04
## p_value 2-tailed  0.001
```

```
estimate_ate(run1, 'outcome_death')
```

```
## outcome_death
```

```
## mean control      2.398438
## mean traitement  2.656489
## ATE              0.258051
## p_value 1-tailed 0.0015
## p_value 2-tailed 0.0027
```

```
cat('run2', '\n')
```

```
## run2
```

```
cat('*****', '\n')
```

```
## *****
```

```
estimate_ate(run2, 'outcome_spread')
```

```
## outcome_spread
## mean control      3.151515
## mean traitement  3.30625
## ATE              0.1547348
## p_value 1-tailed 0.0894
## p_value 2-tailed 0.1657
```

```
estimate_ate(run2, 'outcome_death')
```

```
## outcome_death
## mean control      2.5
## mean traitement  2.60625
## ATE              0.10625
## p_value 1-tailed 0.1091
## p_value 2-tailed 0.2021
```

```
cat('combined', '\n')
```

```
## combined
```

```
cat('*****', '\n')
```

```
## *****
```

```
estimate_ate(combined, 'outcome_spread')
```

```
## outcome_spread
## mean control      3.142308
## mean traitement  3.381443
## ATE              0.2391356
## p_value 1-tailed 8e-04
## p_value 2-tailed 0.0018
```

```
estimate_ate(combined, 'outcome_death')
```

```
## outcome_death
## mean control      2.45
## mean treatment    2.628866
## ATE                0.178866
## p_value 1-tailed  0.0018
## p_value 2-tailed  0.0023
```

## 5. Linear Regression

5.1 Perform linear regression the two outcomes (concern about COVID-19 spread, concern about COVID-19 deaths)

```
regression_labels = c('Treatment',
  'Female', 'Non-binary', 'Gender not answered',
  '20-29', '30-39', '40-49', '50-59', '60-69', '70-79',
  'Black/African American', 'Caucasian', 'Hispanic/Latinx',
  'American Indian or Alaskan Native', 'Pacific Islander', 'Ethnicity not answered',
  'Democrat', 'Independent', 'Political-party-other',
  'High school', 'Some college', 'Associates', 'Bachelors', 'Masters', 'Doctoral', 'Professional (J
  'Know someone sick from COVID-19',
  'Know someone hospitalized from COVID-19',
  'Know someone who died COVID-19')

run_regression_outcome1 = function(d) {
  model_spread = lm(outcome_spread ~ condition, d)
  model_spread_adv = lm(outcome_spread ~ condition
    + as.factor(gender)
    + as.factor(age)
    + as.factor(ethnicity)
    + as.factor(political_party)
    + as.factor(education)
    + as.factor(covid_sick)
    + as.factor(covid_hospitalized)
    + as.factor(covid_died), d)

  est_spread      = get_regression_results_robust_se(model_spread, d, c('condition'), FALSE)
  est_spread_adv = get_regression_results_robust_se(model_spread_adv, d, c('condition'), FALSE)
  return ( list('model'= model_spread,
    'model_adv'=model_spread_adv,
    'est'=est_spread,
    'est_adv'=est_spread_adv))
}

mi1 = run_regression_outcome1(run1)
mi2 = run_regression_outcome1(run2)
mi3 = run_regression_outcome1(combined)

stargazer(mi1$model, mi1$model_adv,
  mi2$model, mi2$model_adv,
  mi3$model, mi3$model_adv,
  type="text", report="vcsp*",
```

```

se = list(mi1$est$robust_se_all, mi1$est_adv$robust_se_all,
          mi2$est$robust_se_all, mi2$est_adv$robust_se_all,
          mi3$est$robust_se_all, mi3$est_adv$robust_se_all),
title=paste('Response to COVID-19 Spread'),
dep.var.caption = "Response to COVID-19 Spread",
dep.var.labels.include = FALSE, model.numbers=FALSE,
column.labels = c("July 21, 2020", "July 21, 2020", "July 25, 2020", "July 25, 2020", "Both dates"),
align=TRUE,
covariate.labels = regression_labels)

```

```

##
## Response to COVID-19 Spread
## =====
##                                     Respon
##                                     -----
##                                     July 21, 2020      July 21, 2020      July 25, 2020
## -----
## Treatment                0.340                0.339                0.155
##                          (0.107)                (0.106)                (0.111)
##                          p = 0.002***            p = 0.002***            p = 0.155
##
## Female                    -0.063
##                          (0.105)
##                          p = 0.550
##
## Non-binary                0.886
##                          (0.420)
##                          p = 0.035**
##
## Gender not answered       0.646
##                          (0.192)
##                          p = 0.001***
##
## 20-29                     0.323
##                          (0.380)
##                          p = 0.395
##
## 30-39                     0.295
##                          (0.381)
##                          p = 0.440
##
## 40-49                     0.181
##                          (0.402)
##                          p = 0.654
##
## 50-59                     0.370
##                          (0.412)
##                          p = 0.370
##
## 60-69                     0.393
##                          (0.441)
##                          p = 0.373
##

```

## 70-79	1.156
##	(0.437)
##	p = 0.009***
##	
## Black/African American	
##	
##	
##	
## Caucasian	0.195
##	(0.201)
##	p = 0.333
##	
## Hispanic/Latinx	-0.087
##	(0.165)
##	p = 0.597
##	
## American Indian or Alaskan Native	0.230
##	(0.205)
##	p = 0.264
##	
## Pacific Islander	-0.471
##	(0.513)
##	p = 0.359
##	
## Ethnicity not answered	-2.519
##	(0.232)
##	p = 0.000***
##	
## Democrat	0.715
##	(0.304)
##	p = 0.019**
##	
## Independent	0.675
##	(0.141)
##	p = 0.00001***
##	
## Political-party-other	0.357
##	(0.171)
##	p = 0.037**
##	
## High school	-0.222
##	(0.370)
##	p = 0.550
##	
## Some college	-0.050
##	(0.800)
##	p = 0.950
##	
## Associates	0.194
##	(0.791)
##	p = 0.806
##	
## Bachelors	0.017
##	(0.795)



```

## p = 0.983
##
## Masters 0.188
## (0.781)
## p = 0.810
##
## Doctoral 0.337
## (0.786)
## p = 0.669
##
## Professional (JD/MD) -0.166
## (0.780)
## p = 0.832
##
## Know someone sick from COVID-19 0.502
## (0.778)
## p = 0.519
##
## Know someone hospitalized from COVID-19 -0.200
## (0.119)
## p = 0.093*
##
## Know someone who died COVID-19 -0.166
## (0.165)
## p = 0.313
##
## as.factor(covid_died)2 0.092
## (0.158)
## p = 0.563
##
## Constant 3.133 2.450 3.152
## (0.085) (0.896) (0.086)
## p = 0.000*** p = 0.007*** p = 0.000***
## -----
## Observations 259 259 292
## R2 0.038 0.314 0.007
## Adjusted R2 0.034 0.227 0.003
## Residual Std. Error 0.858 (df = 257) 0.768 (df = 229) 0.938 (df = 292)
## F Statistic 10.183*** (df = 1; 257) 3.614*** (df = 29; 229) 1.970 (df = 29; 292)
## =====
## Note:

```

```

run_regression_outcome2 = function(d) {
  model_spread = lm(outcome_spread ~ condition, d)
  model_spread_adv = lm(outcome_death ~ condition
    + as.factor(gender)
    + as.factor(age)
    + as.factor(ethnicity)
    + as.factor(political_party)
    + as.factor(education)
    + as.factor(covid_sick)
    + as.factor(covid_hospitalized)
    + as.factor(covid_died), d)
}

```

```

est_spread      = get_regression_results_robust_se(model_spread, d, c('condition'), FALSE)
est_spread_adv = get_regression_results_robust_se(model_spread_adv, d, c('condition'), FALSE)
return ( list('model'= model_spread,
              'model_adv'=model_spread_adv,
              'est'=est_spread,
              'est_adv'=est_spread_adv))
}

mi1d = run_regression_outcome2(run1)
mi2d = run_regression_outcome2(run2)
mi3d = run_regression_outcome2(combined)

stargazer(mi1d$model, mi1d$model_adv,
          mi2d$model, mi2d$model_adv,
          mi3d$model, mi3d$model_adv,
          type="text", report="vcsp*",
          se = list(mi1d$est$robust_se_all, mi1d$est_adv$robust_se_all,
                    mi2d$est$robust_se_all, mi2d$est_adv$robust_se_all,
                    mi3d$est$robust_se_all, mi3d$est_adv$robust_se_all),
          title=paste('Response to COVID-19 Deaths'),
          dep.var.caption = "Response to COVID-19 Deaths",
          dep.var.labels.include = FALSE, model.numbers=FALSE,
          column.labels = c("July 21, 2020", "July 21, 2020", "July 25, 2020", "July 25, 2020", "Both dates"),
          align=TRUE,
          covariate.labels = regression_labels)

```

```

##
## Response to COVID-19 Deaths
## =====
##
##                                     -----
##                                     July 21, 2020      July 21, 2020      July 25, 2020
## -----
## Treatment
##                                     0.340            0.250            0.155
##                                     (0.107)          (0.090)          (0.111)
##                                     p = 0.002***      p = 0.006***      p = 0.100
##
## Female
##                                     0.050
##                                     (0.085)
##                                     p = 0.554
##
## Non-binary
##                                     0.573
##                                     (0.308)
##                                     p = 0.063*
##
## Gender not answered
##                                     0.437
##                                     (0.159)
##                                     p = 0.007***
##
## 20-29
##                                     -0.371
##                                     (0.191)
##                                     p = 0.052*
##

```

## 30-39	-0.323
##	(0.186)
##	p = 0.084*
##	
## 40-49	-0.345
##	(0.209)
##	p = 0.099*
##	
## 50-59	-0.258
##	(0.211)
##	p = 0.221
##	
## 60-69	-0.279
##	(0.260)
##	p = 0.284
##	
## 70-79	0.170
##	(0.254)
##	p = 0.503
##	
## Black/African American	
##	
##	
##	
## Caucasian	0.041
##	(0.168)
##	p = 0.809
##	
## Hispanic/Latinx	-0.103
##	(0.134)
##	p = 0.442
##	
## American Indian or Alaskan Native	-0.019
##	(0.212)
##	p = 0.929
##	
## Pacific Islander	-0.261
##	(0.289)
##	p = 0.367
##	
## Ethnicity not answered	-1.667
##	(0.176)
##	p = 0.000***
##	
## Democrat	0.374
##	(0.239)
##	p = 0.119
##	
## Independent	0.543
##	(0.117)
##	p = 0.00001***
##	
## Political-party-other	0.340
##	(0.134)

##		p = 0.012**	
##			
## High school		-0.107	
##		(0.302)	
##		p = 0.724	
##			
## Some college		0.233	
##		(0.530)	
##		p = 0.660	
##			
## Associates		0.216	
##		(0.516)	
##		p = 0.676	
##			
## Bachelors		0.135	
##		(0.524)	
##		p = 0.797	
##			
## Masters		0.088	
##		(0.512)	
##		p = 0.864	
##			
## Doctoral		0.263	
##		(0.517)	
##		p = 0.612	
##			
## Professional (JD/MD)		-0.475	
##		(0.512)	
##		p = 0.354	
##			
## Know someone sick from COVID-19		0.290	
##		(0.513)	
##		p = 0.572	
##			
## Know someone hospitalized from COVID-19		-0.102	
##		(0.098)	
##		p = 0.298	
##			
## Know someone who died COVID-19		-0.084	
##		(0.125)	
##		p = 0.500	
##			
## as.factor(covid_died)2		0.067	
##		(0.128)	
##		p = 0.603	
##			
## Constant	3.133	2.327	3.152
##	(0.085)	(0.570)	(0.086)
##	p = 0.000***	p = 0.00005***	p = 0.0000
##			
## -----			
## Observations	259	259	292
## R2	0.038	0.252	0.007
## Adjusted R2	0.034	0.157	0.003

```
## Residual Std. Error      0.858 (df = 257)      0.628 (df = 229)      0.938 (df =
## F Statistic      10.183*** (df = 1; 257) 2.657*** (df = 29; 229) 1.970 (df =
## =====
## Note:
```