Problem Set 3

Applied Stats/Quant Methods 1

Due: November 20, 2021

Data

```
dat = read_csv("../../../datasets/incumbents_subset.csv")
```

Question 1

We are interested in knowing how the difference in campaign spending between incumbent and challenger affects the incumbent's vote share.

1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `difflog`. The function call to generate the model is:

   ```
   mod_vote_spend <- lm(voteshare ~ difflog, data = dat)
   ```

   The results are:

Table 1: Vote share as a function of Differental Spending

	Dependent variable:
	voteshare
difflog	0.041666***
	(0.000968)
Constant	0.579031***
	(0.002251)
Observations	3,193
R^2	0.367341
Adjusted R^2	0.367143
Residual Std. Error	0.078673 (df = 3191)
F Statistic	1,852.791000*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

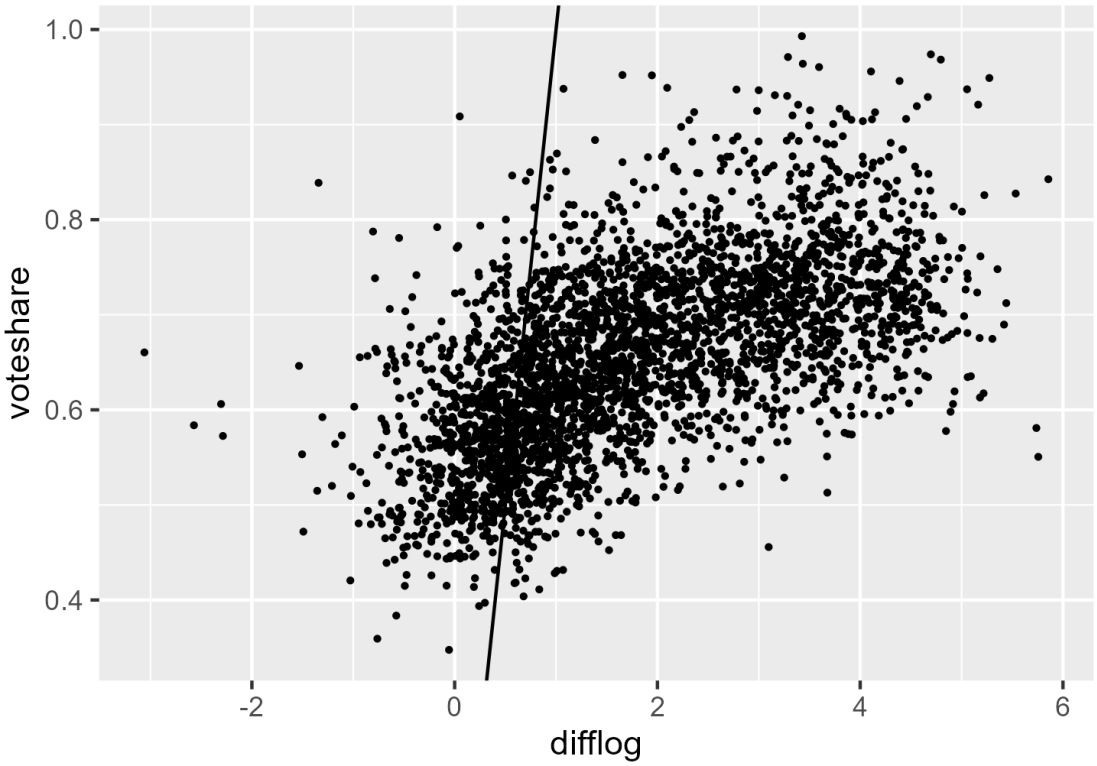

Figure 1: Vote share as a function of Differental Spending

2. Make a scatterplot of the two variables and add the regression line.

3. Save the residuals of the model in a separate object.

```
resid_vote_spend <- mod_vote_spend$residuals
```

4. Write the prediction equation.

   Prediction Equation $voteshare = 0.579031 + (0.041666) * difflog$ ie the voteshare is 0.579031 when difflog is 0 it increases by 0.041666 for each unit increase in difflog

Question 2

We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

1. Run a regression where the outcome variable is `presvote` and the explanatory variable is `difflog`.

   The function call to generate the model is:

```
mod_pres_spend <- lm(presvote ~ difflog, data = dat)
```

   The results are in table 2

Table 2: Presidential vote share as a function of Differental Spending

	Dependent variable:
	presvote
difflog	0.023837***
	(0.001359)
Constant	0.507583***
	(0.003161)
Observations	3,193
R^2	0.087951
Adjusted R^2	0.087665
Residual Std. Error	0.110442 (df = 3191)
F Statistic	307.715400*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

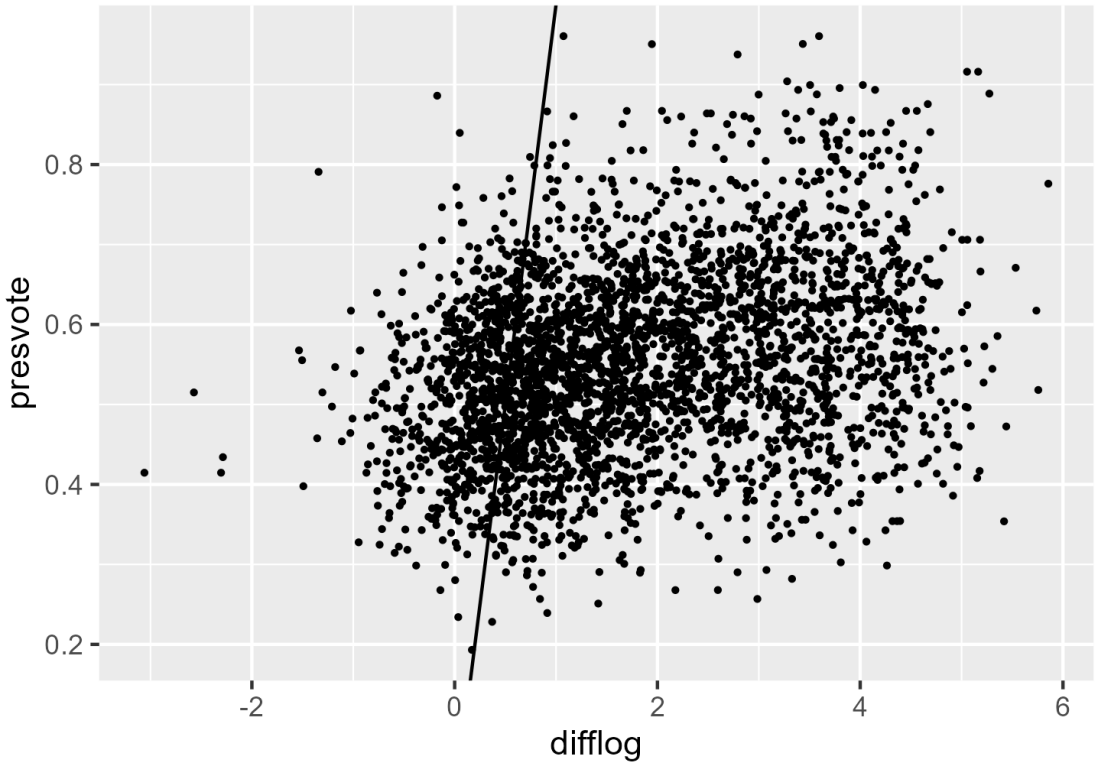

Figure 2: Presidential vote share as a function of Differental Spending

2. Make a scatterplot of the two variables and add the regression line.

3. Save the residuals of the model in a separate object.

```
resid_pres_spend <- mod_pres_spend$residuals
```

4. Write the prediction equation. **Prediction Equation** $presvote = 0.507583 + (0.023837) * diff log$ ie the presvote is 0.507583 when difflog is 0 it increases by 0.023837 for each unit increase in difflog

Question 3

We are interested in knowing how the vote share of the presidential candidate of the incumbent's party is associated with the incumbent's electoral success.

1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `presvote`. The function call to generate the model is:

```
mod_pres_spend <- lm(presvote ~ difflog, data = dat)
```

   and the results are in 3

Table 3: Vote share as a function of Presidential vote share

	Dependent variable:
	voteshare
presvote	0.388018***
	(0.013493)
Constant	0.441330***
	(0.007599)
Observations	3,193
R^2	0.205814
Adjusted R^2	0.205565
Residual Std. Error	0.088146 (df = 3191)
F Statistic	826.950200*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

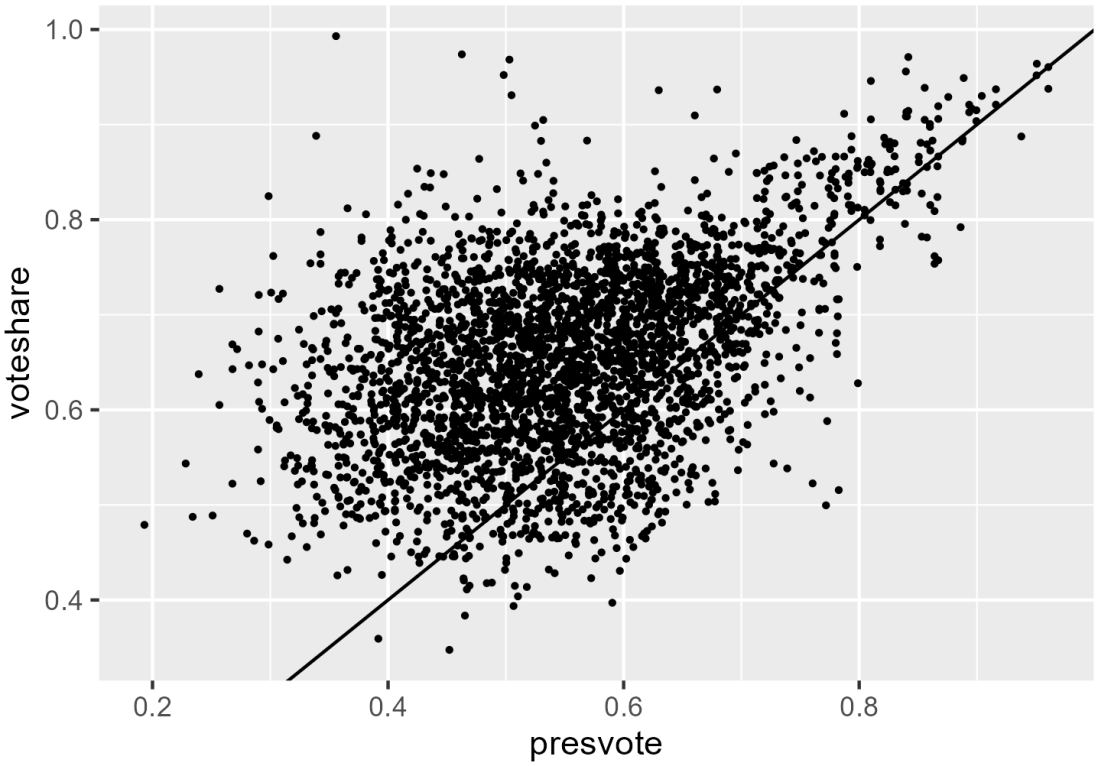

Figure 3: Vote share as a function of Presidential vote share

2. Make a scatterplot of the two variables and add the regression line.

3. Write the prediction equation.

   Prediction Equation

   $voteshare = 0.441330 + (0.388018) * presvote$ ie the voteshare is 0.441330 when presvote is 0 it increases by 0.388018 for each unit increase in presvote

Question 4

The residuals from part (a) tell us how much of the variation in `voteshare` is *not* explained by the difference in spending between incumbent and challenger. The residuals in part (b) tell us how much of the variation in `presvote` is *not* explained by the difference in spending between incumbent and challenger in the district.

1. Run a regression where the outcome variable is the residuals from Question 1 and the explanatory variable is the residuals from Question 2.

   The function call to generate the model is:

   ```
   1 mod_vote_pres <- lm(voteshare ~ presvote, data = dat)
   ```

   The results of the linear model are:

Table 4: Vote share residuals as a function of Presidential vote share residuals

	Dependent variable:
	resid_vote_spend
resid_pres_spend	0.256877***
	(0.011762)
Constant	−0.000000
	(0.001299)
Observations	3,193
R^2	0.130038
Adjusted R^2	0.129765
Residual Std. Error	0.073380 (df = 3191)
F Statistic	476.974700*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

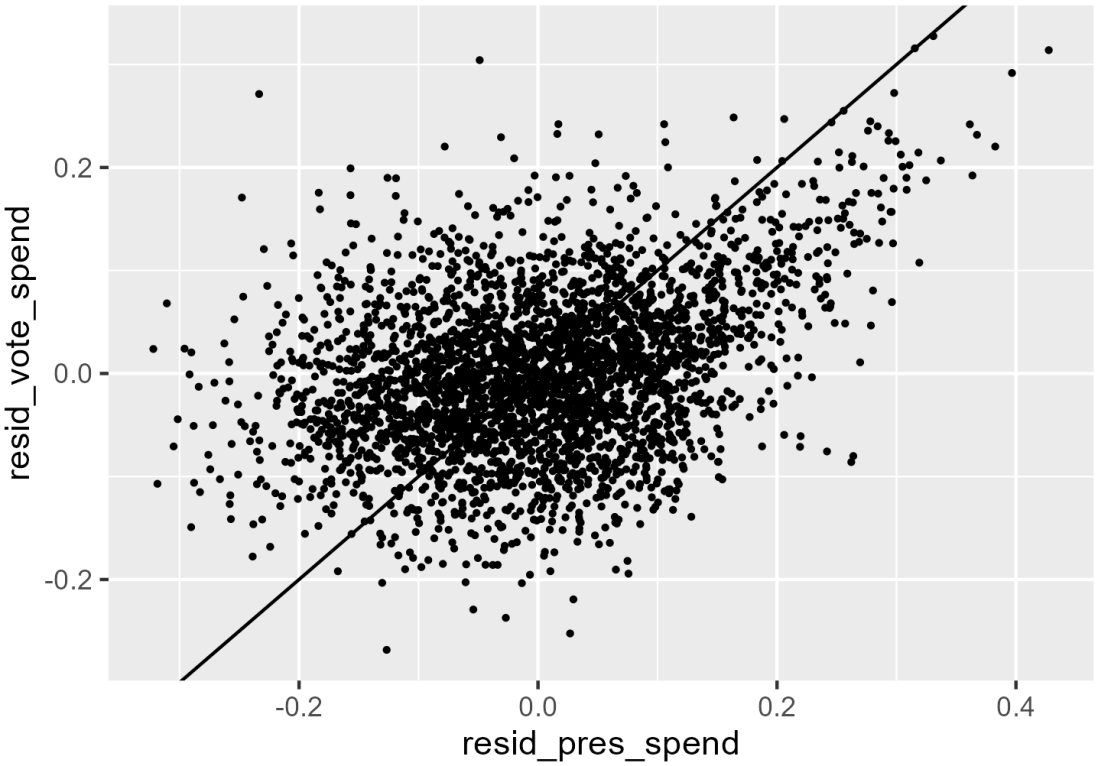

Figure 4: Vote share residuals as a function of Presidential vote share residuals

2. Make a scatterplot of the two residuals and add the regression line.

3. Write the prediction equation.

   Prediction Equation voteshare residuals = -5.207e-18 + (0.2569) * presvote residuals
   ie the voteshare residual value is -5.207e-18 when presvote residual value is 0 it increases by 0.2569 for each unit increase in presvote residuals

   ie the value of the incumbent vote share not accounted for by the difference in incumbent spending increases by 0.2569 for each unit increase in the value of the unidentified factors which cause an increase in presidential vote share.

Question 5

What if the incumbent's vote share is affected by both the president's popularity and the difference in spending between incumbent and challenger?

1. Run a regression where the outcome variable is the incumbent's `voteshare` and the explanatory variables are `difflog` and `presvote`.

   The function call to generate the model is:

```
mod_vote_spend_pres <- lm(voteshare ~ difflog + presvote, data = dat)
```

   The results of the linear model are:

Table 5: Vote share as a function of Presidential vote share and differential spending

	Dependent variable:
	voteshare
difflog	0.035543***
	(0.000946)
presvote	0.256877***
	(0.011764)
Constant	0.448644***
	(0.006330)
Observations	3,193
R^2	0.449610
Adjusted R^2	0.449265
Residual Std. Error	0.073391 (df = 3190)
F Statistic	1,302.947000*** (df = 2; 3190)
Note:	*p<0.1; **p<0.05; ***p<0.01

The additional variables are plotted in 5.

2. Write the prediction equation.

   Prediction Equation $voteshare = 0.4486442 + (0.0355431) * difflog + (0.2568770) * presvote$

   ie the voteshare is 0.4486442 when difflog and presvote are 0 it increases by 0.0355431 for each unit increase in difflog (holding presvote constant) it increases by 0.2568770 for each unit increase in presvote (holding difflog constant)

3. What is it in this output that is identical to the output in Question 4? Why do you think this is the case?

   The coefficient for residual president share as a function of incumbent spending (Q4) is the same as the coefficient for presidential vote share.(0.256877)

   In model 5, the coefficient for presvote is a partial predictor, with difflog held constant. In model 2, the residuals represent the variation in the value of the presidential vote, excluding difflog (which was specifically included as a predictor). In both models, we are getting a predictive value for presvote on voteshare, with difflog excluded/controlled.

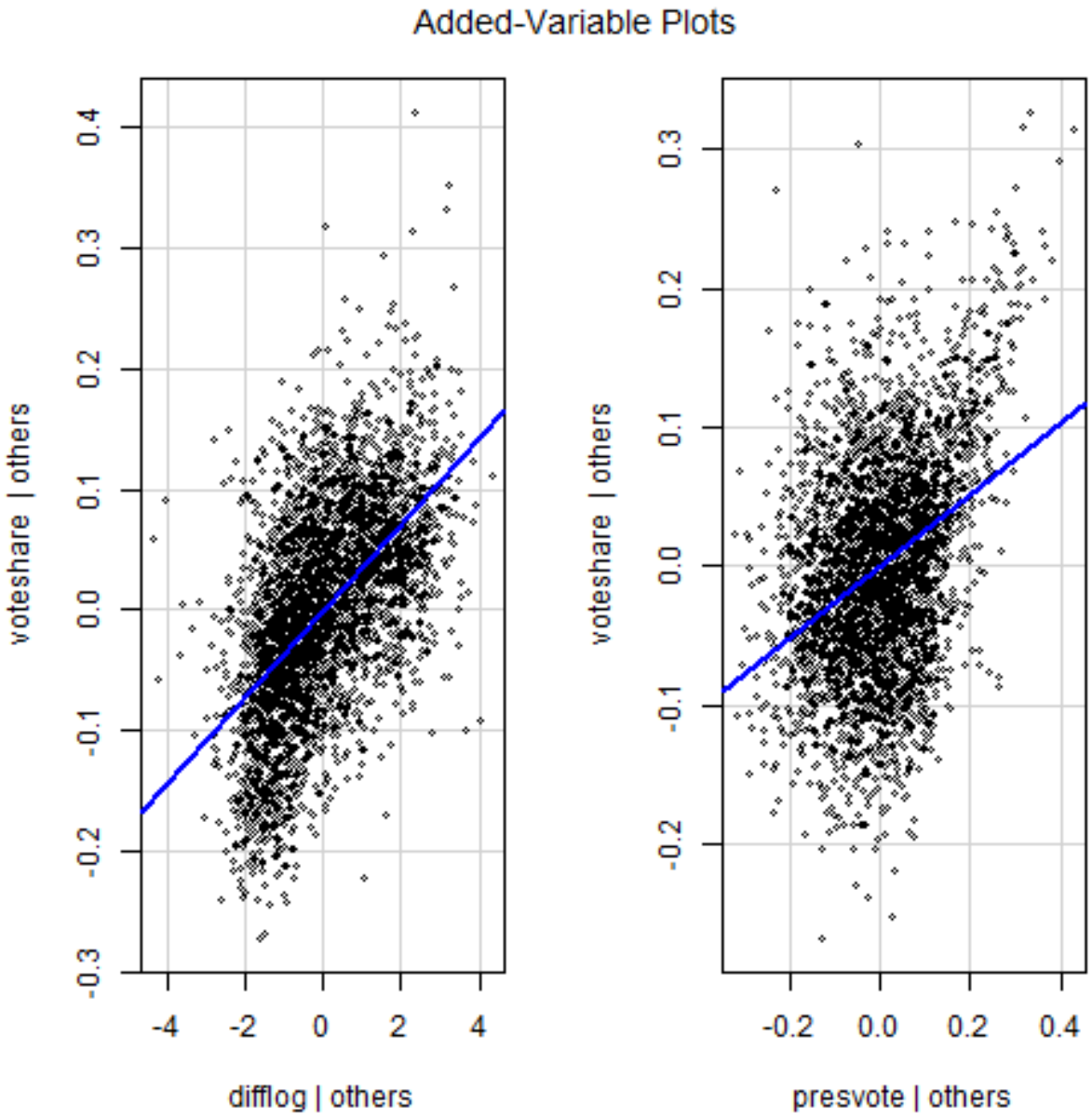

Figure 5: Vote share as a function of differential spending and presidential vote share

Appendix - Code

```
#####################################################################
# Imelda Finn, 22334657
# POP77003 - Stats I
# clear global .envir, load libraries, set wd
#####################################################################

# remove objects
rm(list=ls())

# detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:grDevices", "package:utils", "package:datasets", "package:methods", "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1, TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0)  for (package in package.list) detach(package, character.only=TRUE)
}
detachAllPackages()

# load libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}

# load necessary packages
lapply(c("ggplot2", "stargazer", "tidyverse", "stringr", "car"), pkgTest)

# function to save output to a file that you can read in later to your docs
output_stargazer <- function(outputFile, appendVal=TRUE, ...) {
  output <- capture.output(stargazer(...))
  cat(paste(output, collapse = "\n"), "\n", file=outputFile, append=appendVal)
}


# set working directory to current parent folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

#####################################################################

#Problem Set 3
#Applied Stats/Quant Methods 1
#Due: November 20, 2021
#Instructions
# Please show your work! You may lose points by simply writing in the answer. If the
```

```r
#problem requires you to execute commands in R, please include the code you used to
#get your answers. Please also include the .R file that contains your code. If you are
#not sure if work needs to be shown for a particular problem, please ask.
# Your homework should be submitted electronically on GitHub.
# This problem set is due before 23:59 on Sunday November 20, 2022. No late assign-
#  ments will be accepted.
# Total available points for this homework is 80.
#In this problem set, you will run several regressions and create an add variable plot (see the
#lecture slides) in R using the incumbents subset.csv dataset. Include all of your code.

######################
# get data
######################
# line in tex = 61
dat = read_csv("../../../datasets/incumbents_subset.csv")
saveRDS(dat, "Data/incumbents_subset.csv")
dat <- readRDS("Data/incumbents_subset.csv")


######################
#Question 1
######################

# difflog is the difference between the incumbents' spending
# and the challengers' spending
dl = sum(dat$difflog - (dat$incspend - dat$chalspend))

png("Graphics/incumbent_subset.png" )
pairs(~voteshare + presvote + difflog, dat)
dev.off() # close output



mean(dat$difflog)
fivenum(dat$difflog)
#-3.0600882  0.6733625  1.6105532  2.9856564  5.8558103'
fivenum(dat$presvote)
#0.1931263 0.4695266 0.5478388 0.6239612 0.9606069
fivenum(dat$voteshare)
#0.3475569 0.5845994 0.6568562 0.7232970 0.9930361


#We are interested in knowing how the difference in campaign spending between incumbent
#and challenger affects the incumbent's vote share.
#1. Run a regression where the outcome variable is voteshare and the
```

```
   explanatory variable
#is difflog.
#1
#2. Make a scatterplot of the two variables and add the regression line.
#3. Save the residuals of the model in a separate object.
#4. Write the prediction equation.
#2

#1.  # line in tex 100
mod_vote_spend <- lm(voteshare ~ difflog, data = dat)

#2.
dat %>% ggplot(aes( x= difflog, y = voteshare)) +
  geom_point(size=0.5) +
  geom_abline()
ggsave("Graphics/vote_spend.png")

#3.  line in tex 109
resid_vote_spend <- mod_vote_spend$residuals

#4.
summary(mod_vote_spend)

output_stargazer("Tables/mod_vote_spend.tex", appendVal = FALSE, mod_vote_
   spend,
                 title="Vote share as a function of Differental Spending",
                 label="tab:vote_spend",  digits = 6 #, summary=FALSE
)

#Prediction Equation
# voteshare = 0.579031 + (0.041666) * difflog
# ie the voteshare is 0.579031 when difflog is 0
#    it increases by 0.041666 for each unit increase in difflog

#Call:
#  lm(formula = voteshare ~ difflog, data = dat)

#Residuals:
#  Min       1Q   Median       3Q      Max
#-0.26832 -0.05345 -0.00377  0.04780  0.32749

#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept) 0.579031   0.002251  257.19   <2e-16 ***
#  difflog     0.041666   0.000968   43.04   <2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 0.07867 on 3191 degrees of freedom
#Multiple R-squared:  0.3673, Adjusted R-squared:  0.3671
#F-statistic:  1853 on 1 and 3191 DF,  p-value: < 2.2e-16
```

```
141
142
143 #Question 2
144 #We are interested in knowing how the difference between incumbent and
      challenger's
145 # spending and the vote share of the presidential candidate of
146 # the incumbent's party are related.
147 #1. Run a regression where the outcome variable is presvote and the
      explanatory variable
148 #is difflog.
149 #2. Make a scatterplot of the two variables and add the regression line.
150 #3. Save the residuals of the model in a separate object.
151 #4. Write the prediction equation.
152 #3
153
154 #1.   # line in tex 155
155 mod_pres_spend <- lm(presvote ~ difflog, data = dat)
156
157 #2.
158 dat %>% ggplot(aes( x= difflog, y = presvote)) +
159   geom_point(size=0.5) +
160   geom_abline()
161 ggsave("Graphics/pres_spend.png")
162
163 #3.  line in tex 164
164 resid_pres_spend <- mod_pres_spend$residuals
165
166 #4.
167 summary(mod_pres_spend)
168
169 output_stargazer("Tables/mod_pres_spend.tex", appendVal = FALSE,
170                  mod_pres_spend,
171                  title="Presidential vote share as a function of Differental
      Spending",
172                  label="tab:pres_spend",  digits = 6 #, summary=FALSE
173 )
174
175 #Prediction Equation
176 # presvote = 0.507583 + (0.023837) * difflog
177 # ie the presvote is 0.507583 when difflog is 0
178 #    it increases by 0.023837 for each unit increase in difflog
179
180 #Call:
181 #lm(formula = presvote ~ difflog, data = dat)
182
183 #Residuals:
184 #  Min       1Q   Median       3Q      Max
185 #-0.32196 -0.07407 -0.00102  0.07151  0.42743
186
187 #Coefficients:
188 #  Estimate Std. Error t value Pr(>|t|)
```

```r
#(Intercept) 0.507583    0.003161   160.60    <2e-16 ***
#  difflog      0.023837    0.001359    17.54    <2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 0.1104 on 3191 degrees of freedom
#Multiple R-squared:  0.08795,  Adjusted R-squared:  0.08767
#F-statistic: 307.7 on 1 and 3191 DF,  p-value: < 2.2e-16

#Question 3
#We are interested in knowing how the vote share of the presidential
# candidate of the incumbent's party is associated with the
#incumbent's electoral success.
#1. Run a regression where the outcome variable is voteshare and the
    explanatory variable
#is presvote.
#2. Make a scatterplot of the two variables and add the regression line.
#3. Write the prediction equation.
#4

#1)    # line in tex 209
mod_vote_pres <- lm(voteshare ~ presvote, data = dat)

#2)
dat %>% ggplot(aes( x= presvote, y = voteshare)) +
  geom_point(size=0.5) +
  geom_abline()
ggsave("Graphics/vote_pres.png")

#3)
summary(mod_vote_pres)

output_stargazer("Tables/mod_vote_pres.tex", appendVal = FALSE,
                 mod_vote_pres,
                 title="Vote share as a function of Presidential vote share",
                 label="tab:vote_pres",  digits = 6 #, summary=FALSE
)

#Prediction Equation
# voteshare = 0.441330 + (0.388018) * presvote
# ie the voteshare is 0.441330 when presvote is 0
#    it increases by 0.388018 for each unit increase in presvote


#Call:
#  lm(formula = voteshare ~ presvote, data = dat)

#Residuals:
#  Min       1Q   Median       3Q      Max
#-0.27330 -0.05888  0.00394  0.06148  0.41365

```

```
#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept) 0.441330   0.007599   58.08   <2e-16 ***
#  presvote    0.388018   0.013493   28.76   <2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 0.08815 on 3191 degrees of freedom
#Multiple R-squared:  0.2058, Adjusted R-squared:  0.2056
#F-statistic:   827 on 1 and 3191 DF,  p-value: < 2.2e-16


#Question 4
#The residuals from part (a) tell us how much of the variation in voteshare is not explained
#by the difference in spending between incumbent and challenger. The residuals in part (b)
#tell us how much of the variation in presvote is not explained by the difference in spending
#between incumbent and challenger in the district.
#1. Run a regression where the outcome variable is the residuals from Question 1 and the
#explanatory variable is the residuals from Question 2.
#2. Make a scatterplot of the two residuals and add the regression line.
#3. Write the prediction equation.

#1.    # line in tex 262
mod_resid_vote_pres <- lm(resid_vote_spend ~ resid_pres_spend)

#2.

ggplot() +
geom_point(aes(resid_pres_spend, resid_vote_spend), size=0.5) +
  geom_abline()
ggsave("Graphics/residuals.png")

#3.
summary(mod_resid_vote_pres)
output_stargazer("Tables/mod_resid_vote_pres.tex", appendVal = FALSE,
                 mod_resid_vote_pres,
                 title="Vote share residuals as a function of Presidential vote share residuals",
                 label="tab:residuals",  digits = 6 #, summary=FALSE
)

# residuals for both are clustered around 0
#Prediction Equation
# voteshare residuals  = -5.207e-18 + (0.2569) * presvote residuals
# ie the voteshare residual value is -5.207e-18 when presvote residual value is 0
#    it increases by 0.2569 for each unit increase in presvote residuals
```

```
284
285 fivenum(resid_pres_spend)
286 fivenum(resid_vote_spend)
287 mean(resid_pres_spend)      # ie 0
288 mean(resid_vote_spend)      # ie 0, as expected
289
290
291 #Call:
292 #  lm(formula = resid_vote_spend ~ resid_pres_spend)
293
294 #Residuals:
295 #  Min        1Q    Median        3Q       Max
296 #-0.25928 -0.04737 -0.00121  0.04618  0.33126
297
298 #Coefficients:
299 #  Estimate Std. Error t value Pr(>|t|)
300 #(Intercept)       -5.207e-18  1.299e-03    0.00        1
301 #resid_pres_spend  2.569e-01  1.176e-02   21.84   <2e-16 ***
302 #  ---
303 #  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
304
305 #Residual standard error: 0.07338 on 3191 degrees of freedom
306 #Multiple R-squared:   0.13,  Adjusted R-squared:  0.1298
307 #F-statistic:   477 on 1 and 3191 DF,  p-value: < 2.2e-16
308
309 #Question 5
310 #What if the incumbent's vote share is affected by both the president's
    popularity and the
311 #difference in spending between incumbent and challenger?
312 #1. Run a regression where the outcome variable is the incumbent's voteshare
    and the
313 #explanatory variables are difflog and presvote.
314 #2. Write the prediction equation.
315 #3. What is it in this output that is identical to the output in Question 4?
    Why do you
316 #think this is the case?
317 #  6
318
319 #1.    # line in tex 320
320 mod_vote_spend_pres <- lm(voteshare ~ difflog + presvote, data = dat)
321
322 #2)
323 #install.packages("car")
324 #library(car)
325 png("Graphics/add_variable.png" )
326 avPlots(mod_vote_spend_pres, col = carPalette()[1], pch = 1, cex=0.5, id=FALSE
    )
327 dev.off() # close output
328
329
330 #dat %>% ggplot(aes( ~voteshare + difflog + presvote)) +
```

```
#  geom_
#  geom_point() +
#  geom_abline()
#ggsave("Graphics/vote_pres.png")

#3)
summary(mod_vote_spend_pres)

output_stargazer("Tables/mod_vote_spend_pres.tex", appendVal = FALSE,
                 mod_vote_spend_pres,
                 title="Vote share as a function of Presidential vote share and differential spending",
                 label="tab:vote_spend_pres",  digits = 6 #, summary=FALSE
)

#Prediction Equation
# voteshare = 0.4486442 + (0.0355431) * difflog + (0.2568770) * presvote
# ie the voteshare is 0.4486442 when difflog and presvote are 0
#    it increases by 0.0355431 for each unit increase in difflog
#          (holding presvote constant)
#    it increases by 0.2568770 for each unit increase in presvote
#          (holding difflog constant)

#Call:
#  lm(formula = voteshare ~ difflog + presvote, data = dat)

#Residuals:
#  Min       1Q   Median       3Q      Max
#-0.25928 -0.04737 -0.00121  0.04618  0.33126

#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept) 0.4486442  0.0063297   70.88   <2e-16 ***
#  difflog     0.0355431  0.0009455   37.59   <2e-16 ***
#  presvote    0.2568770  0.0117637   21.84   <2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 0.07339 on 3190 degrees of freedom
#Multiple R-squared:  0.4496, Adjusted R-squared:  0.4493
#F-statistic:  1303 on 2 and 3190 DF,  p-value: < 2.2e-16
```