Chloe Sutter / Problem Set 3

QTM 200: Applied Regression Analysis

Due: February 17, 2020

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.

Question 1 (20 points)

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

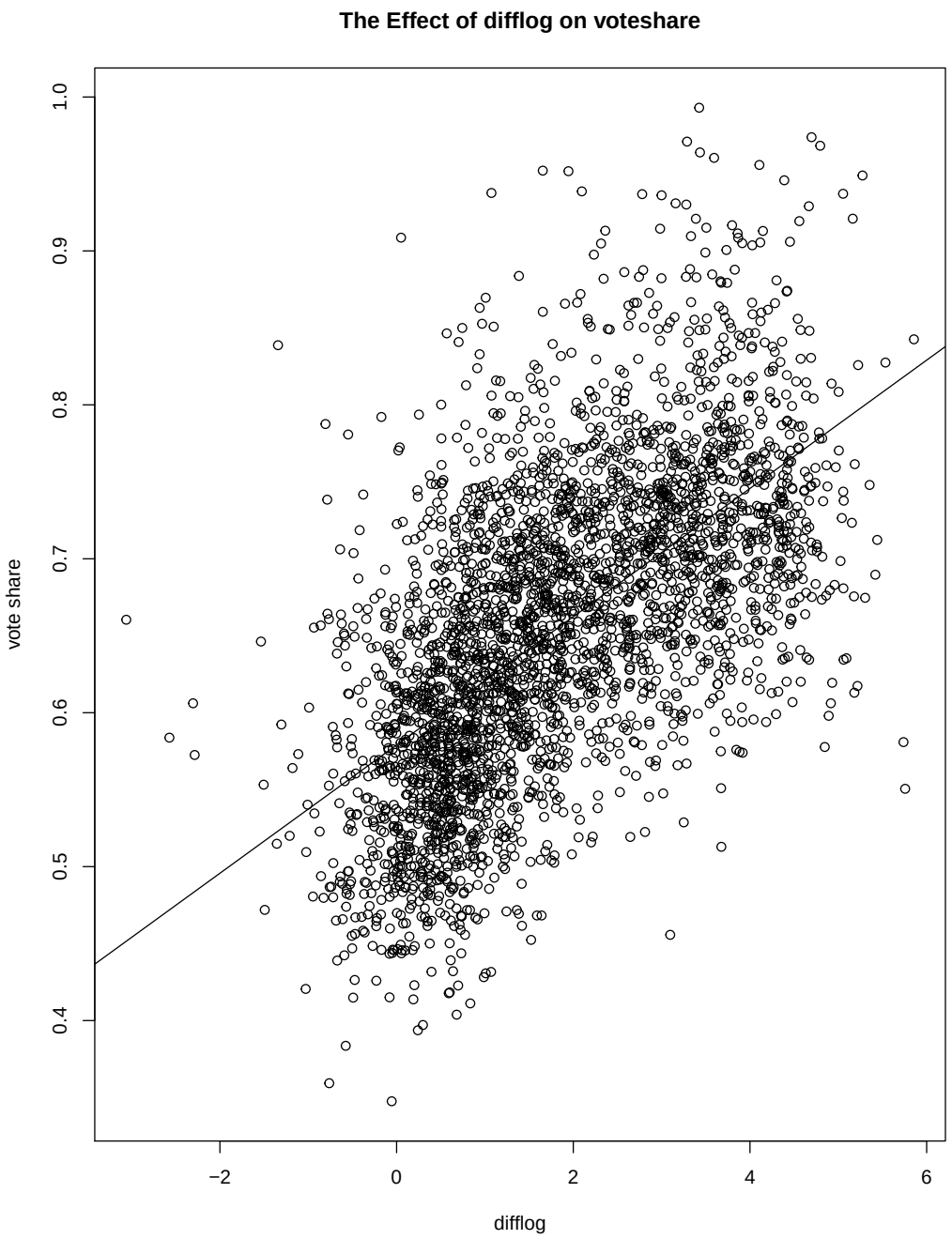
1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `difflog`.

```
#look at data
summary(incumbents_subset)
table1 <- data.frame(incumbents_subset$voteshare, incumbents_subset$
    difflog)
table1
# estimate regression
regress1 <- lm(incumbents_subset$voteshare~ incumbents_subset$difflog)
regress1

# Coefficients:
  # (Intercept)       difflog
  # 0.57903       0.04167
```

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

```
plot(incumbents_subset$voteshare~incumbents_subset$difflog, xlab="difflog
    ", ylab =
       "vote share", main="The Effect of difflog on voteshare")
# display estimated regression line
abline(regress1)
```

Figure 1:



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

```
residuals(regress1)
regress1resid <- residuals(regress1)
```

4. Write the prediction equation.

   $Y = \alpha + \beta X$

   Y = .57 + .04X where Y is a vector of observed outcomes for voteshare and and X is a vector of difflog.

Question 2 (20 points)

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.

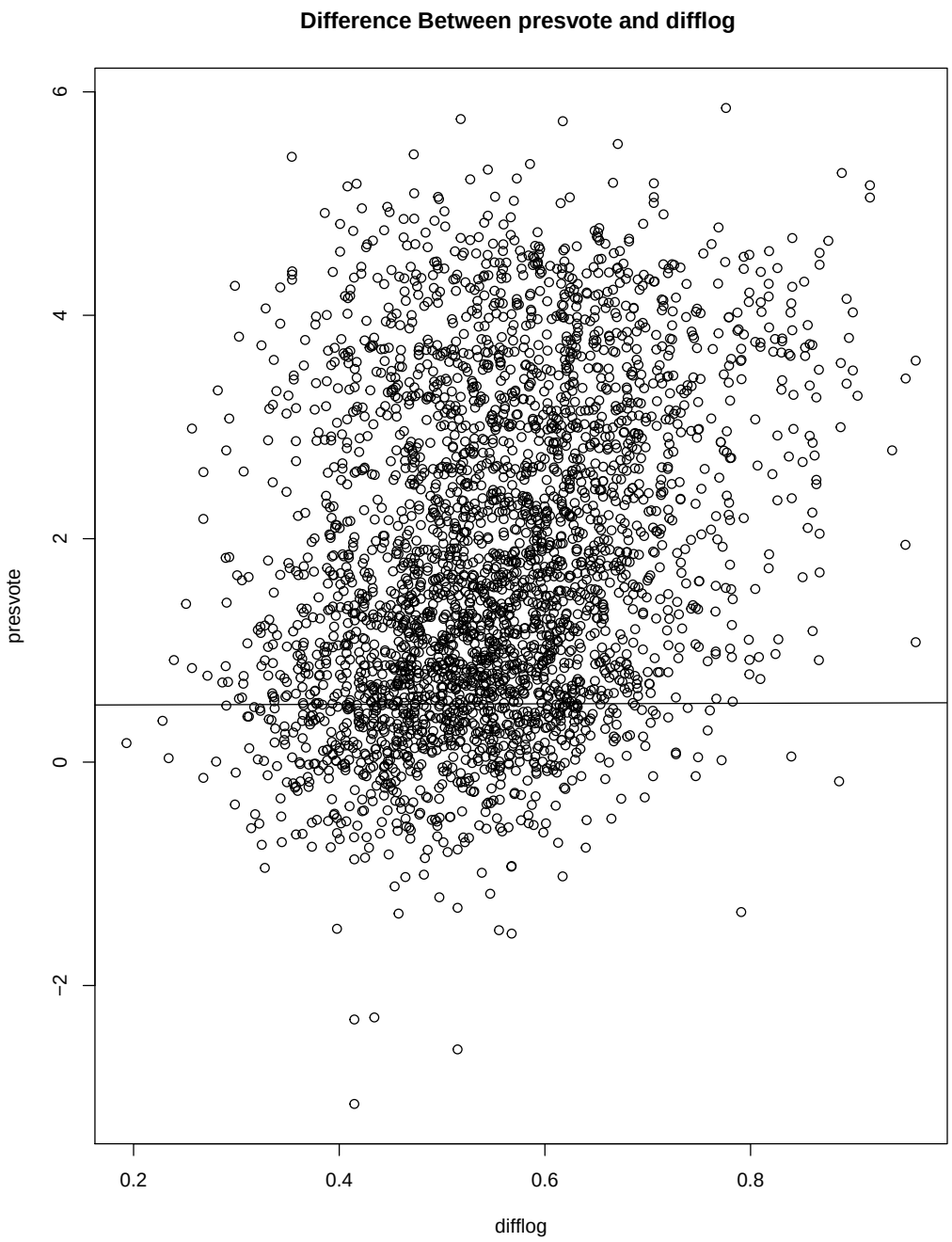
1. Run a regression where the outcome variable is `presvote` and the explanatory variable is `difflog`.

```
# estimate regression
regress2 <- lm(incumbents_subset$presvote~incumbents_subset$difflog)
regress2
#Coefficients:
  #(Intercept)  incumbents_subset$difflog
  #0.50758                     0.02384
```

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

```
plot(incumbents_subset$presvote,incumbents_subset$difflog, xlab="difflog"
    , ylab="presvote", main="Difference Between presvote and difflog ")
# display estimated regression line
abline(regress2)
```

Figure 2:



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

```
residuals(regress2)
regress2resid <- residuals(regress2)
```

4. Write the prediction equation.

   $Y = \alpha + \beta X$

   $Y = .507 + .023X$ where Y is a vector of the observed outcomes for presvote and X is a vector of difflog.

Question 3 (20 points)

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.

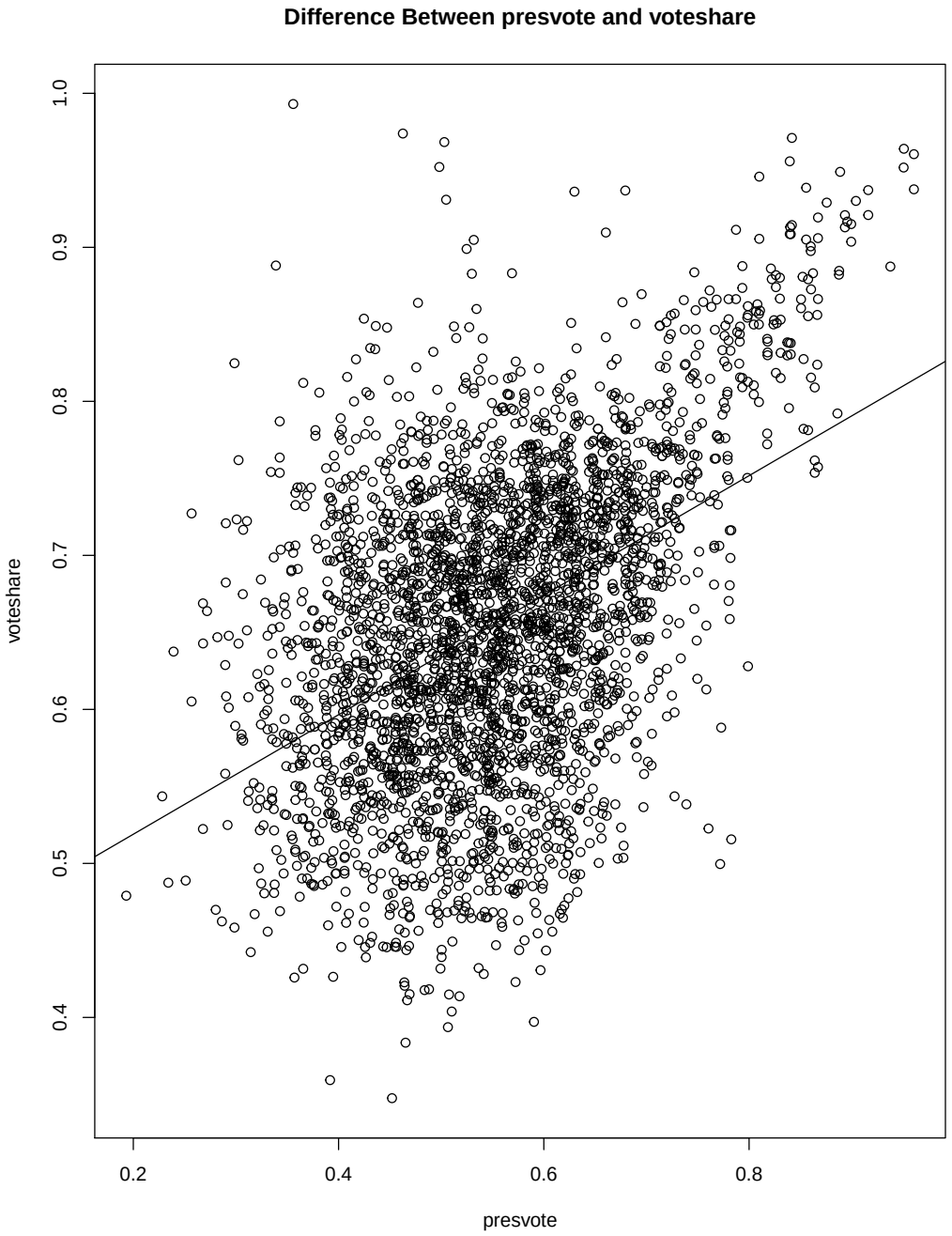
1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `presvote`.

```
regress3 <- lm(incumbents_subset$voteshare~incumbents_subset$presvote)
regress3
# Coefficients:
  # (Intercept)  incumbents_subset$presvote
  # 0.4413                       0.3880
```

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

```
plot(incumbents_subset$voteshare~incumbents_subset$presvote,  xlab="
    presvote", ylab="voteshare", main="Difference Between presvote and
    voteshare")
# display estimated regression line
abline(regress3)
```

Figure 3:



3. Write the prediction equation.

   $Y = \alpha + \beta X$

   Y = .44 + .38X where Y is a vector of observed outcomes for voteshare and X is a

vector of presvote.

Question 4 (20 points)

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.

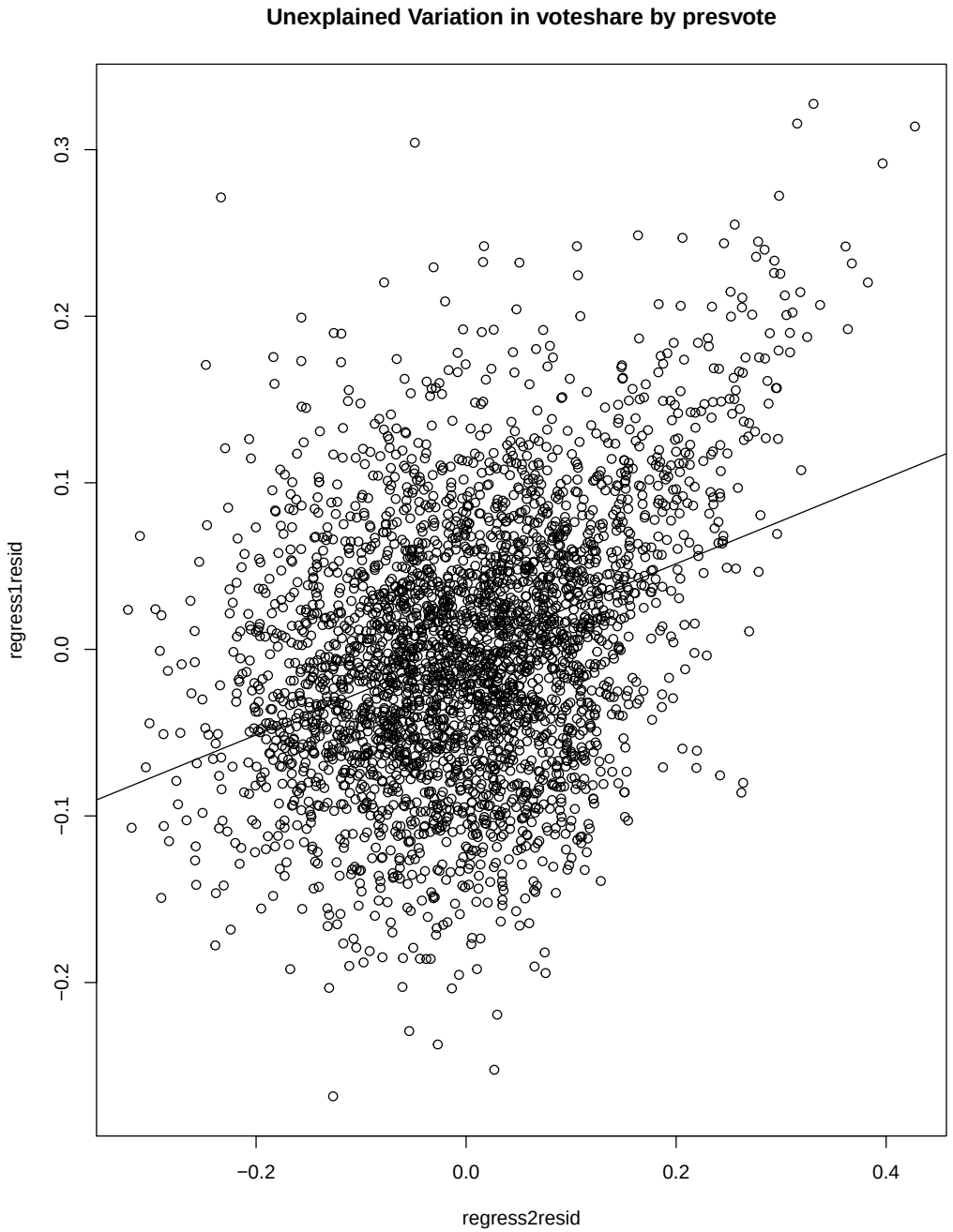
1. Run a regression where the outcome variable is the residuals from Question 1 and the explanatory variable is the residuals from Question 2.

```
# outcome variable is the residuals from Question 1 (Y = regress1resid)
# and the explanatory variable is the residuals from Question 2 (X =
    regress2resid)

lm(regress1resid~regress2resid)
regress4 <- lm(regress1resid~regress2resid)
# find residuals
regress4resid <- residuals(regress4)

# Coefficients:
  # (Intercept)             regress2resid
  # -0.00000000000000000486    0.25687701270009788423
```

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

```
plot(regress1resid~regress2resid, main="Unexplained Variation in
    voteshare by presvote")
# display estimated regression line
abline(regress4)
```

Figure 4:



3. Write the prediction equation.

   $Y = \alpha + \beta X$

   Y = -0.0000000000000000048 + 0.25X Where Y = residuals describing how much of

the variation in VOTESHARE is not explained by the difference in spending between incumbent and challenger and where X = residuals describing how much of the variation in PRESVOTE is not explained by the difference in spending between incumbent and challenger in the district.

Question 5 (20 points)

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`. Find regression by-hand.

```
# Y: voteshare
# X1: difflog
# X2: presvote

lm_by_hand <- function(inputDF, covariates, outcome){
#load required packages
require(MASS)
#create matrices
n=nrow(Y)
X <- matrix(c(rep(1,n), incumbents_subset$difflog, incumbents_subset$
  presvote), ncol=3)
Y <- matrix(incumbents_subset$voteshare, ncol=3)
dim(X)
#calculate betas
betas <- solve((t(X)%*%X))%*%(t(X)%*%Y)
rownames(betas)[3] <- "intercept"
k <- ncol(X)
#calculate SEs for betas
#estimate sigma-squared
sigma_squared <- sum((Y-X%*%betas)^2)/(nrow(X)-ncol(X))
#create variance-covariance matrix for betas
var_covar_mat <- sigma_squared*solve(t(X)%*%X)
#standard errors for coefficient estimates
SEs <- sqrt(diag(var_covar_mat))
#get t-stat and p-vals
TS <- (betas-0)/SEs
p_values <- 2*pt(abs(TS), n-k, lower.tail=F)
#regression
reg <- lm_by_hand(incumbents_subset, c("difflog", "presvote"), "
  voteshare")
}
```

Check regression in r

```
regression5 <- lm(voteshare ~ difflog + presvote, data=incumbents_subset)
```

```
regression5
#Coefficients:
  #(Intercept)       difflog      presvote
  #0.44864       0.03554       0.25688
summary(regression5)
regress5resid <- residuals(regression5)
summary(regress5resid)
```

2. Write the prediction equation.

   $\mu y = \beta 0 + \beta 1 X1 + \beta 2 X2$

   $\mu y = 0.449 + 0.035 X1 + 0.256 X2$ where X1 represents difflog and X2 represents presvote.

3. What is it in this output that is identical to the output in Question 4? Why do you think this is the case? Reflect on your finding. Don't write anything. Just think about it.