Problem Set 3

QTM 200: Applied Regression Analysis

Due: February 17, 2020

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the 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 the course GitHub page in `.pdf` form.
- This problem set is due at the beginning of class on Monday, February 17, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

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.

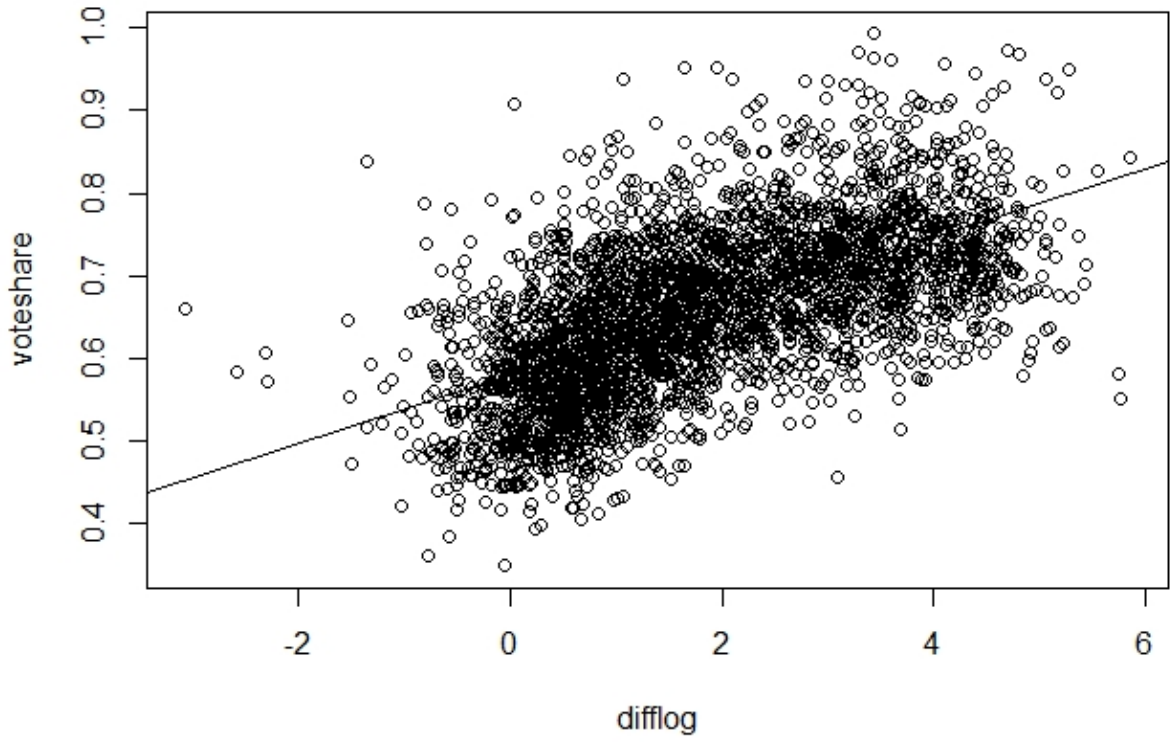
1. Run a regression where the outcome variable is `voteshare` and the explanatory variable is `difflog`.

```
#1a. Run a regression where the outcome variable is voteshare and the
    explanatory variable is difflog.
#calculate sums and means
mean.difflog <- mean(incumbents_subset$difflog)
mean.vote <- mean(incumbents_subset$voteshare)
sum(incumbents_subset$difflog)
```

```
sum(incumbents_subset$voteshare)
numerator <- sum((incumbents_subset$voteshare - mean(incumbents_subset$
    voteshare))*(incumbents_subset$difflog - mean(incumbents_subset$
    difflog)))
denominator <- sum((incumbents_subset$difflog - mean(incumbents_subset$
    difflog))^2)
#calculate regression coefficient
beta.hat <- numerator/denominator
beta.hat
alpha.hat <- mean.vote - (beta.hat*mean.difflog)
alpha.hat
#calculate p-value
sd.vote <- sd(incumbents_subset$voteshare)
se.vote <-  sd.vote/sqrt(sum((incumbents_subset$voteshare - mean.difflog)
    ^2))
TS <- (beta.hat - 0) / se.vote
TS
p <- 2*pt(abs(TS), df=(length(incumbents_subset$voteshare)), lower.tail =
     F)
p
#check model
lm1 <- lm(voteshare~difflog, data=incumbents_subset)
summary(lm1)
```

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

```
#1b. Make a scatterplot of the two variables and add the regression line.
library(ggplot2)
plot(voteshare~difflog, data=incumbents_subset)
abline(lm(voteshare ~ difflog, data=incumbents_subset))
```



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

```
residual.lm1 <- incumbents_subset$voteshare - 0.579031 -  (0.041666*
    incumbents_subset$difflog)
```

4. Write the prediction equation. Yi = 0.579031 + 0.041666Xi + errori

Question 2 (20 points)

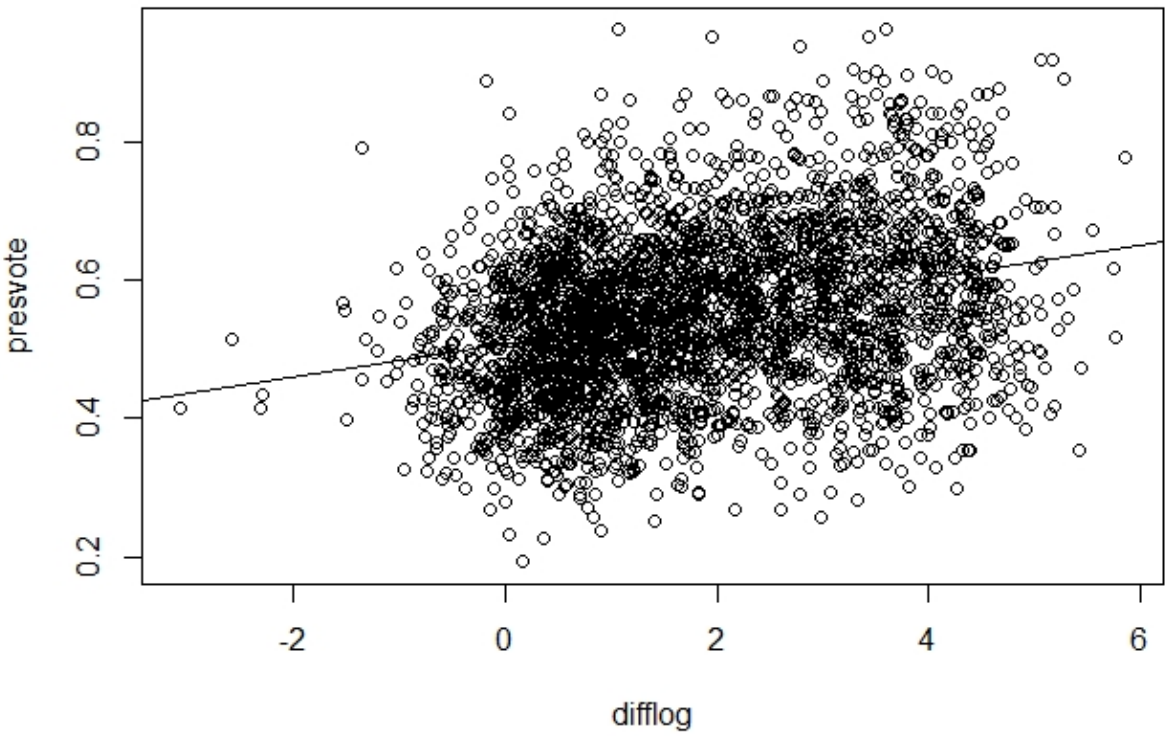
We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

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

```
#question 2a. Run a regression where the outcome variable is presvote and
    the explanatory variable is difflog.
#calculate sums and means
mean.difflog <- mean(incumbents_subset$difflog)
mean.presvote <- mean(incumbents_subset$presvote)
sum(incumbents_subset$difflog)
sum(incumbents_subset$presvote)
numerator2 <- sum((incumbents_subset$presvote - mean(incumbents_subset$
    presvote))*(incumbents_subset$difflog - mean(incumbents_subset$difflog
    )))
denominator2 <- sum((incumbents_subset$difflog - mean(incumbents_subset$
    difflog))^2)
#calculate regression coefficient
beta.hat2 <- numerator2/denominator2
beta.hat2
alpha.hat2 <- mean.presvote - (beta.hat2*mean.difflog)
alpha.hat2
#calculate p-value
sd.presvote <- sd(incumbents_subset$presvote)
se.presvote <-  sd.presvote/sqrt(sum((incumbents_subset$presvote - mean.
    difflog)^2))
TS2 <- (beta.hat2 - 0) / se.presvote
TS2
p2 <- 2*pt(abs(TS2), df=(length(incumbents_subset$presvote)), lower.tail
    = F)
p2
#check model
lm2 <- lm(presvote~difflog, data=incumbents_subset)
summary(lm2)
```

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

```
plot(presvote~difflog, data=incumbents_subset)
abline(lm(presvote ~ difflog, data=incumbents_subset))
```



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

```
residual.lm2 <- incumbents_subset$presvote - 0.507583 -  (0.023837*
    incumbents_subset$difflog)
```

4. Write the prediction equation. Yi = 0.507583 + 0.023837Xi + errori

Question 3 (20 points)

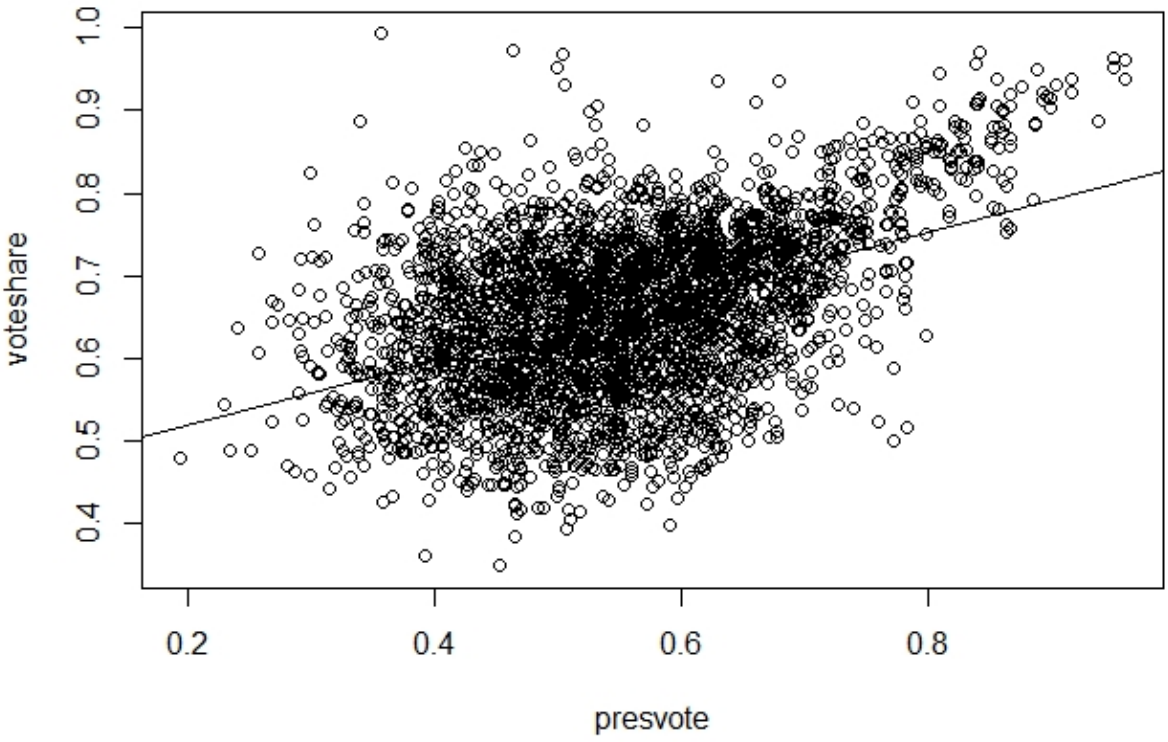
We are interested in knowing how the vote share of the presidential candidate of the incumbent's party is associated with the incumbent's electoral success.

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

```
#question 3a.  Run a regression where the outcome variable is voteshare
    and the explanatory variable is presvote.
#calculate sums and means
mean.presvote
mean.vote
sum(incumbents_subset$presvote)
sum(incumbents_subset$voteshare)
numerator3 <- sum((incumbents_subset$voteshare - mean(incumbents_subset$
    voteshare))*(incumbents_subset$presvote - mean(incumbents_subset$
    presvote)))
denominator3 <- sum((incumbents_subset$presvote - mean(incumbents_subset$
    presvote))^2)
#calculate regression coefficient
beta.hat3 <- numerator3/denominator3
beta.hat3
alpha.hat3 <- mean.vote - (beta.hat3*mean.presvote)
alpha.hat3
#calculate p-value
sd.vote
se.vote2 <-  sd.vote/sqrt(sum((incumbents_subset$voteshare - mean.
    presvote)^2))
TS3 <- (beta.hat3 - 0) / se.vote2
TS3
p3 <- 2*pt(abs(TS3), df=(length(incumbents_subset$voteshare)), lower.tail
    = F)
p3
#check model
lm3 <- lm(voteshare~presvote, data=incumbents_subset)
summary(lm3)
```

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

```
plot(voteshare~presvote, data=incumbents_subset)
abline(lm(voteshare ~ presvote, data=incumbents_subset))
```



3. Write the prediction equation. Yi = 0.441330 + 0.388018Xi + errori

Question 4 (20 points)

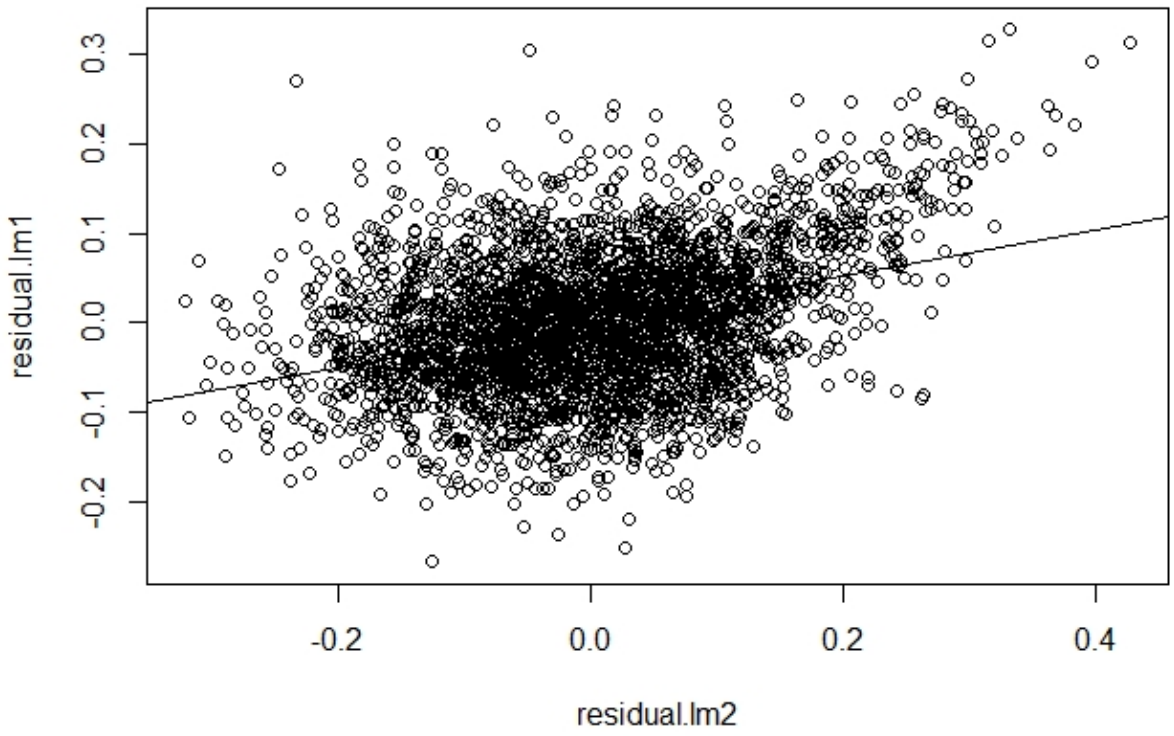
The residuals from part (a) tell us how much of the variation in `voteshare` is *not* explained by the difference in spending between incumbent and challenger. The residuals in part (b) tell us how much of the variation in `presvote` is *not* explained by the difference in spending between incumbent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

```
mean.res1 <- mean(residual.lm1)
mean.res2 <- mean(residual.lm2)
sum(residual.lm1)
sum(residual.lm2)
numerator4 <- sum((residual.lm1 - mean(residual.lm1))*(residual.lm2 -
    mean(residual.lm2)))
denominator4 <- sum((residual.lm2 - mean(residual.lm2))^2)
#calculate regression coefficient
beta.hat4 <- numerator4/denominator4
beta.hat4
alpha.hat4 <- mean.res1 - (beta.hat4*mean.res2)
alpha.hat4
#calculate p-value
sd.r <- sd(residual.lm1)
se4 <-  sd.r/sqrt(sum((residual.lm1 - mean.res2)^2))
TS4 <- (beta.hat4 - 0) / se4
TS4
p4 <- 2*pt(abs(TS4), df=(length(residual.lm1)), lower.tail = F)
p4
#check
lm4 <- lm(residual.lm1~residual.lm2)
summary(lm4)
```

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

```
plot(residual.lm1~residual.lm2)
abline(lm(residual.lm1 ~ residual.lm2))
```



3. Write the prediction equation. Yhat = 1.086e-07 + 0.256877X + error

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`.

```
#question 5a. Run a regression where the outcome variable is the
    incumbent's voteshare and the explanatory variables are difflog and
    presvote.
#subset data
sub.incumb <- incumbents_subset[,c("voteshare", "difflog", "presvote")]
#estimate regression by hand
lm_by_hand <- function(sub.incumb, difflog, presvote, voteshare)
X <- as.matrix(cbind(rep(1, dim(sub.incumb)[2]), sub.incumb["difflog", "
    presvote"]))
Y <- sub.incumb["voteshare"]
#calculate betas
betas <- solve((t(X)%*%X)) %*% (t(X)%*%Y)
#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)
#SEs for coefficient estimates
SEs <- sqrt(diag(var_covar_mat))

#check
lm5 <- lm(voteshare~difflog+presvote, data=incumbents_subset)
summary(lm5)
```

2. Write the prediction equation. Yi = 0.4486442 + 0.0355431X1 + 0.2568770X2 + error

   (X1 is difflog, X2 is 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? The min, 1Q, median, 3Q and max for the residuals are identical.

   I think this is the case because the residual model from question 4 takes into account the same three variables as the multiple regression, so they have the same residual values.