Problem Set 3 Answers

QTM 200: Applied Regression Analysis

Vanessa Wong

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`.

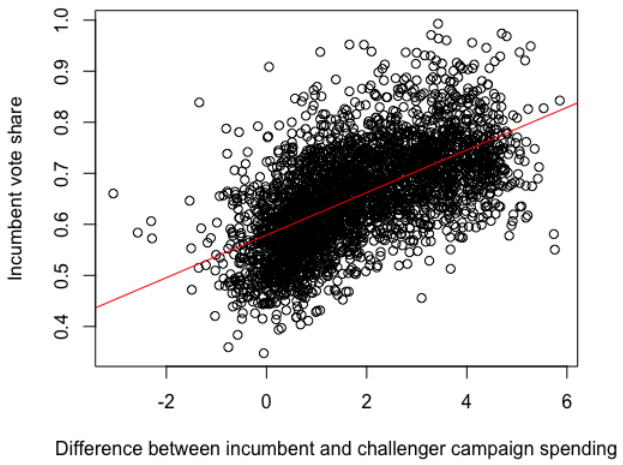
   Linear model: $y = 0.5790 + 0.0417x$

```
# x = difflog (explanatory)
# y = voteshare (outcome)

# run a regression where the outcome variable is voteshare and the
    explanatory variable is difflog.
ymean <- mean(incumbents$voteshare)
```

```r
xmean <- mean(incumbents$difflog)
ysum <- sum(incumbents$voteshare)
xsum <- sum(incumbents$difflog)
yy <- (incumbents$voteshare) - (ymean)
xx <- (incumbents$difflog) - (xmean)
yyxxsum <- sum(yy*xx)
xxsq <- (xx)^2
sumxxsq <- sum(xxsq)
betaincumbs <- yyxxsum/sumxxsq
betaincumbs
# beta = 0.0417
alphaincumbs <- ymean - (betaincumbs*xmean)
alphaincumbs
# alpha = 0.5790
# linear model: y = 0.5790 + 0.0417x
# check work
incumbreg <- lm(incumbents$voteshare ~ incumbents$difflog)
incumbreg
```

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




```r
plot(incumbents$difflog, incumbents$voteshare,
     xlab="Difference between incumbent and challenger campaign spending"
    , ylab="Incumbent vote share")
abline(a=0.5790, b=0.0417, col="red")
```

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

```
residscheck <- residuals(incumbreg)
residscheck
```

4. Write the prediction equation.

$$\hat{y} = 0.5790 + 0.0417x$$

Question 2 (20 points)

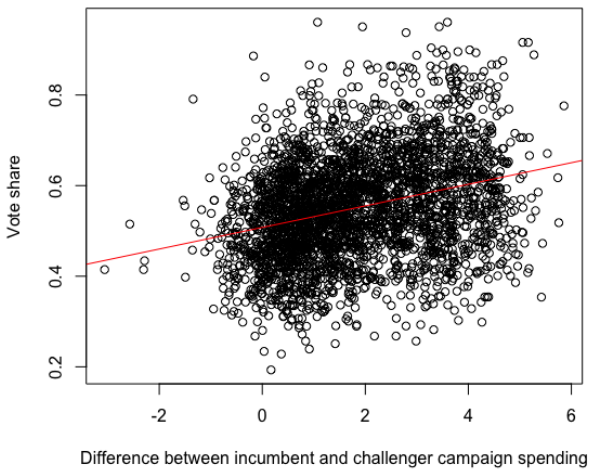
We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

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

   Linear model: y = 0.5076 + 0.0238x

```
y2mean <- mean(incumbents$presvote)
y2sum <- sum(incumbents$presvote)
yy2 <- (incumbents$presvote) - y2mean
yyxxsum2 <- sum(xx*yy2)
betaincumbs2 <- yyxxsum2/sumxxsq
betaincumbs2
# beta = 0.0238
alphaincumbs2 <- y2mean - (betaincumbs2*xmean)
alphaincumbs2
# alpha = 0.5076
# linear model: y = 0.5076 + 0.0238x
# check work
incumbreg2 <- lm(incumbents$presvote ~ incumbents$difflog)
incumbreg2
```

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




```
plot(incumbents$difflog, incumbents$presvote,
```

```
     xlab="Difference between incumbent and challenger campaign spending"
    , ylab="Vote share")
abline(a=0.5076, b=0.0238, col="red")
```

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

```
resids2 <- residuals(incumbreg2)
resids2
```

4. Write the prediction equation.

$$\hat{y} = 0.5076 + 0.0238x$$

Question 3 (20 points)

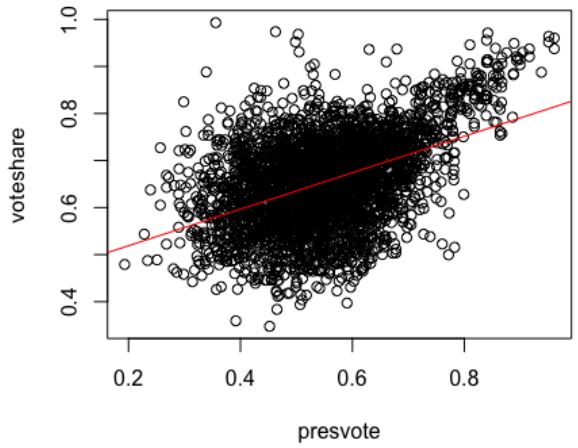
We are interested in knowing how the vote share of the presidential candidate of the incumbent's party is associated with the incumbent's electoral success.

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

   Linear model: y= 0.4413 + 0.3880x

```
xmean3 <- mean(incumbents$presvote)
ymean3 <- mean(incumbents$voteshare)
xsum3 <- sum(incumbents$presvote)
ysum3 <- sum(incumbents$voteshare)
yy3 <- (incumbents$voteshare) - (ymean3)
xx3 <- (incumbents$presvote) - (xmean3)
yyxxsum3 <- sum(yy3*xx3)
xxsq3 <- (xx3)^2
sumxxsq3 <- sum(xxsq3)
beta3 <- yyxxsum3/sumxxsq3
beta3
# beta = 0.3880
alpha3 <- ymean3 - (beta3*xmean3)
alpha3
# alpha = 0.4413
# linear model: y= 0.4413 + 0.3880x
# check work
reg3 <- lm(incumbents$voteshare~incumbents$presvote)
reg3
```

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



```r
plot(incumbents$presvote, incumbents$voteshare,
     xlab="presvote", ylab="voteshare")
abline(a=0.4413, b=0.3880, col="red")
```

3. Write the prediction equation.

$$\hat{y} = 0.4413 + 0.3880x$$

Question 4 (20 points)

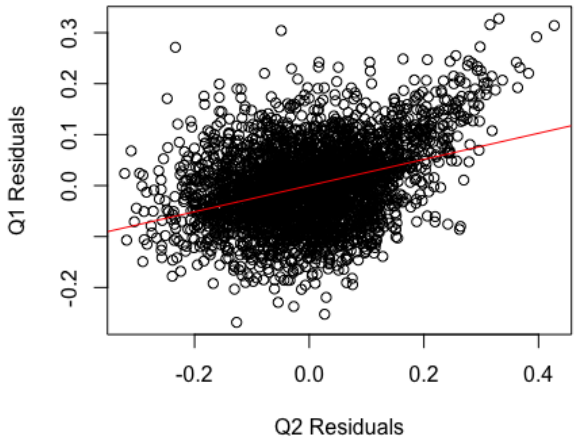
The residuals from part (a) tell us how much of the variation in `voteshare` is *not* explained by the difference in spending between incumbent and challenger. The residuals in part (b) tell us how much of the variation in `presvote` is *not* explained by the difference in spending between incumbent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

   Linear model: y = -4.860 x $10^{-18} + 0.2569x$

```
xmean4 <- mean(resids2)
ymean4 <- mean(residscheck)
xsum4 <- sum(resids2)
ysum4 <- sum(residscheck)
yy4 <- residscheck - (ymean4)
xx4 <- resids2 - (xmean4)
yyxxsum4 <- sum(yy4*xx4)
xxsq4 <- (xx4)^2
sumxxsq4 <- sum(xxsq4)
beta4 <- yyxxsum4/sumxxsq4
beta4
# beta = 0.2569
alpha4 <- ymean4 - (beta4*xmean4)
alpha4
# alpha = -4.860 x 10^(-18)
# linear model: y = -4.860 x 10^(-18) + 0.2569x
# check work
reg4 <- lm(residscheck ~ resids2)
reg4
summary(reg4)
```

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




```
plot(resids2, residscheck,
     xlab="Q2 Residuals", ylab="Q1 Residuals")
abline(reg4, col="red")
```

3. Write the prediction equation.

   $\hat{y}$ = -4.860 x 10^{-18} + 0.2569x

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`.

```
# [,2] = presvote
Y3 <- matrix(incumbents$voteshare, nrow=3193)
Y3
XtX <- t(X3) %*% X3
Xty <- t(X3) %*% Y3
Xty
XtX.inv <- solve(XtX)
XtX
b <- XtX.inv %*% Xty
b
dim(incumbents)[1]

# check work
lm(incumbents$voteshare ~ incumbents$difflog + incumbents$presvote)$
    coefficients
```

```
multireg2 <- lm(incumbents$voteshare ~ incumbents$difflog + incumbents$
    presvote)
summary(multireg2)
# linear model: y = 0.4486 + 0.0355xdifflog + 0.2569xpresvote
```

2. Write the prediction equation.

$$\hat{y} = 0.4486 + 0.0355x_{difflog} + 0.2569x_{presvote}$$

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

   Aswered in tex file

NOTE: The final R file with answers is called PS3_answers4. I had to resave the file multiple times because I was making changes to the R file as I was compiling.

```
######################
# load libraries
# set wd
# clear global .envir
######################

# 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)
}

# here is where you load any necessary packages
# ex: stringr
# lapply(c("stringr"), pkgTest)
```

```
29
30 lapply(c("stringr"),  pkgTest)
31
32 # set working directory, import datasets
33 setwd("~/GitHub/QTM200Spring2020/problem_sets/PS3")
34 incumbents <- read.csv("incumbents_subset.csv")
35
36 ########################
37 # Problem 1:
38 ########################
39 # x = difflog (explanatory)
40 # y = voteshare (outcome)
41
42 # run a regression where the outcome variable is voteshare and the explanatory
       variable is difflog.
43 ymean <- mean(incumbents$voteshare)
44 xmean <- mean(incumbents$difflog)
45 ysum <- sum(incumbents$voteshare)
46 xsum <- sum(incumbents$difflog)
47 yy <- (incumbents$voteshare) - (ymean)
48 xx <- (incumbents$difflog) - (xmean)
49 yyxxsum <- sum(yy*xx)
50 xxsq <- (xx)^2
51 sumxxsq <- sum(xxsq)
52 betaincumbs <- yyxxsum/sumxxsq
53 betaincumbs
54 # beta = 0.0417
55 alphaincumbs <- ymean - (betaincumbs*xmean)
56 alphaincumbs
57 # alpha = 0.5790
58 # linear model: y = 0.5790 + 0.0417x
59 # check work
60 incumbreg <- lm(incumbents$voteshare ~ incumbents$difflog)
61 incumbreg
62
63 # make a scatterplot of the two variables and add the regression line
64 plot(incumbents$difflog, incumbents$voteshare,
65      xlab="Difference between incumbent and challenger campaign spending",
       ylab="Incumbent vote share")
66 abline(a=0.5790, b=0.0417, col="red")
67
68 # save the residuals of the model in a separate object
69 residscheck <- residuals(incumbreg)
70 residscheck
71
72 # write the prediction equation
73 #yhat = 0.5790 + 0.0417x
74
75 ########################
76 # Problem 2:
77 ########################
```

```
# x = difflog (explanatory)
# y = presvote (outcome)

# run a regression where the outcome variable is presvote and the explanatory
    variable is difflog
y2mean <- mean(incumbents$presvote)
y2sum <- sum(incumbents$presvote)
yy2 <- (incumbents$presvote) - y2mean
yyxxsum2 <- sum(xx*yy2)
betaincumbs2 <- yyxxsum2/sumxxsq
betaincumbs2
# beta = 0.0238
alphaincumbs2 <- y2mean - (betaincumbs2*xmean)
alphaincumbs2
# alpha = 0.5076
# linear model: y = 0.5076 + 0.0238x
# check work
incumbreg2 <- lm(incumbents$presvote ~ incumbents$difflog)
incumbreg2

# make a scatterplot of the two variables and add the regression line
plot(incumbents$difflog, incumbents$presvote,
     xlab="Difference between incumbent and challenger campaign spending",
    ylab="Vote share")
abline(a=0.5076, b=0.0238, col="red")

#save the residuals of the model in a separate object
resids2 <- residuals(incumbreg2)
resids2

# write the prediction equation
# yhat = 0.5076 + 0.0238x

######################
# Problem 3:
######################
# x = presvote (explanatory)
# y = voteshare (outcome)

# run a regression where the outcome variable is voteshare and the explanatory
    variable is presvote
xmean3 <- mean(incumbents$presvote)
ymean3 <- mean(incumbents$voteshare)
xsum3 <- sum(incumbents$presvote)
ysum3 <- sum(incumbents$voteshare)
yy3 <- (incumbents$voteshare) - (ymean3)
xx3 <- (incumbents$presvote) - (xmean3)
yyxxsum3 <- sum(yy3*xx3)
xxsq3 <- (xx3)^2
sumxxsq3 <- sum(xxsq3)
beta3 <- yyxxsum3/sumxxsq3
```

```
beta3
# beta = 0.3880
alpha3 <- ymean3 - (beta3*xmean3)
alpha3
# alpha = 0.4413
# linear model: y= 0.4413 + 0.3880x
# check work
reg3 <- lm(incumbents$voteshare~incumbents$presvote)
reg3

# make a scatterplot of the two variables and add the regression line
plot(incumbents$presvote, incumbents$voteshare,
     xlab="presvote", ylab="voteshare")
abline(a=0.4413, b=0.3880, col="red")

# write the prediction equation
# yhat = 0.4413 + 0.3880x

#######################
# Problem 4:
#######################
# x = resids2 (explanatory)
# y = residscheck (outcome)

# run a regression where the outcome variable is the residuals from Q1 and the
    explantory variable is the residuals from Q2
xmean4 <- mean(resids2)
ymean4 <- mean(residscheck)
xsum4 <- sum(resids2)
ysum4 <- sum(residscheck)
yy4 <- residscheck - (ymean4)
xx4 <- resids2 - (xmean4)
yyxxsum4 <- sum(yy4*xx4)
xxsq4 <- (xx4)^2
sumxxsq4 <- sum(xxsq4)
beta4 <- yyxxsum4/sumxxsq4
beta4
# beta = 0.2569
alpha4 <- ymean4 - (beta4*xmean4)
alpha4
# alpha = -4.860 x 10^(-18)
# linear model: y = -4.860 x 10^(-18) + 0.2569x
# check work
reg4 <- lm(residscheck ~ resids2)
reg4
summary(reg4)

# make a scatterplot of the two reisduals and add the regression line
plot(resids2, residscheck,
     xlab="Q2 Residuals", ylab="Q1 Residuals")
abline(reg4, col="red")
```

```
176
177 # write the prediction equation
178 # yhat = -4.860 x 10^(-18) + 0.2569x
179
180 ######################
181 # Problem 5:
182 ######################
183 # x = difflog, presvote (explanatory)
184 # y = voteshare (outcome)
185
186 # run a regression where the outcome variable is the incumbent's voteshare and
      the explanatory variables are difflog and presvote
187 X3 <- cbind(incumbents$difflog, incumbents$presvote)
188 X3
189 # [,1] = difflog
190 # [,2] = presvote
191 dim(incumbents)[1]
192 Y3 <- matrix(incumbents$voteshare, nrow=3193)
193 Y3
194 XtX <- t(X3) %*% X3
195 Xty <- t(X3) %*% Y3
196 Xty
197 XtX.inv <- solve(XtX)
198 XtX
199 b <- XtX.inv %*% Xty
200 b
201 # check work
202 lm(incumbents$voteshare ~ incumbents$difflog + incumbents$presvote)
    $coefficients
203 multireg2 <- lm(incumbents$voteshare ~ incumbents$difflog +
    incumbents$presvote)
204 summary(multireg2)
205 # linear model: y = 0.4486 + 0.0355xdifflog + 0.2569xpresvote
206
207 # write the prediction equation
208 # yhat = 0.4486 + 0.0355xdifflog + 0.2569xpresvote
209
210
211 # what is it in this output that is identical to the output in Q4? why do you
    think this is the case?
212 # the slope of the regression of voteshare on presvote is the same as the
    slope of of the regression of Q1 residuals on Q2 residuals.
213 # i think this is the case because the regression of Q1 residuals on Q2
    residuals tells us how much variation in presvote
214 # *and* voteshare is not explained by difflog. in both cases, the slopes
    describe the relationship between voteshare and presvote, while
215 # holding difflog constant. As such, they have the same slope.
```