Problem Set 4

Quantitative Political Methodology (U25 363)

Due: April 24, 2018

# Instructions

* Put your name at the top of your written document. Complete and show your work in a written Word or .pdf document that you will upload to GitHub. Please show your work if possible. 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 have plots, attach them as well within your written document. Make sure you label clearly which question the codes correspond to. If you are not sure if work needs to be shown for a particular problem, please ask me.
* Your homework should be submitted electronically on the course GitHub page. • This problem set is due before the beginning of class on Wednesday April 24, 2019.

No late assignments will be accepted.

* Total available points for this homework is 100.

# Question 1 (10 points)

The following table was created using the data from a study run in a major Latin American city.[[1]](#footnote-1) As part of the study, confederate made illegal left turns across traffic to draw the attention of the police officers. Two of the confederates were upper class drivers and two were lower class drivers. The researchers were interested in whether officers were more or less likely to solicit a bribe from drivers depending on their class (officers use phrases like, “We can solve this the easy way” to draw a bribe). The table below shows the resulting data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Not Stopped | Bribe requested | Stopped/given warning |
| Upper class | 14 | 6 | 7 |
| Lower class | 7 | 7 | 1 |

1. Calculate the *χ*[[2]](#footnote-2) test statistic by hand (even better if you can do ”by hand” in R).

Fe= (row total/grand total) x column total x2 = ∑(fo-fe)^2/fe

X2 =[(14-13.5)^2/13.5 + (6-8.4)^2/8.4 + (7-5.1)^2/5.1 + (7-7.5)^2/7.5 + (7-4.6)^2/4.6 + (1-2.9)^2/2.9]

X2= 3.94(by hand) or 3.79(in R)

x <- matrix(c(14,6,7,7,7,1),ncol = 3, nrow = 2, byrow = TRUE)

rownames(x) <- c("upper", "lower")

colnames(x) <- c("NS", "B", "W")

x <- Table1

chisq.test(Table1, correct = FALSE)

1. Now calculate the p-value (in R).2 What do you conclude if *α* = *.*1?
   1. P-value=.1502, fail to reject the null hypothesis
2. chisq.test(Table1, correct = FALSE)
3. Calculate the standardized residuals for each cell and put them in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Not Stopped | Bribe requested | Stopped/given warning |
| Upper class  Lower class | 1.55  -3.23 | -1.67  1.68 | 1.57  -1.55 |

1. How might the standardized residuals help you interpret the results?

Helps us see how if our expected values are held relatively constant from our observed values. Letting us know if any outliers are present or not.

# Question 2 (15 points)

Researchers are interested in learning the effect of all of those yard signs on voting preferences.[[3]](#footnote-3) Working with a campaign in Fairfax County, Virginia, 131 precincts were randomly divided into a treatment and control group. In 30 precincts, signs were posted around the precinct that read, “For Sale: Terry McAuliffe. Don’t Sellout Virgina on November 5.” Below is the result of a regression with two variables and a constant. The dependent variable is the proportion of the vote that went to McAuliff’s opponent Ken Cuccinelli. The first variable indicates whether a precinct was randomly assigned to have the sign against McAuliffe posted. The second variable indicates a precinct that was adjacent to a precinct in the treatment group (since people in those precincts might be exposed to the signs).

**Impact of lawn signs on vote share**

Precinct assigned lawn signs (n=30) 0.042

(0.016) Precinct adjacent to lawn signs (n=76) 0.042

(0.013)

Constant 0.302

(0.011)

*Notes: R*2=0.094, N=131

1. Use the results to determine whether having these yard signs in a precinct affects vote share (e.g., conduct a hypothesis test with *α* = *.*05).

Ho: μo= μo

Ha: μo ≠ μo

Z=(.0135 - .016)/(.0233)= -0.107

P=2\*pnorm(.107, lower.tail=F)=.915

.915>.05

We cannot reject the null hypothesis

Does not affect vote share

1. Use the results to determine whether being next to precincts with these yard signs affects vote share (e.g., conduct a hypothesis test with *α* = *.*05).

Ho: μo= μo

Ha: μo ≠ μo

Z=(.012 - .013)/(.013)= -0.0769

P=2\*pnorm(.0769, lower.tail=F)=.9387

.9387>.05

We cannot reject the null

Does not affect vote share

1. Interpret the coefficient for the constant term substantively.

Of those untreated, there was a close to zero correlation to having no lawn signs and the share of vote for cuccinelli

1. Evaluate the model fit for this regression. What does this tell us about the importance of yard signs versus other factors?

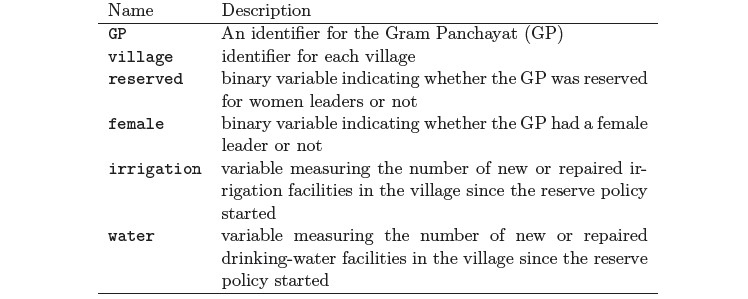
Yard signs play an almost insignificant role in the affect on voter share than other factors

# Question 3 (10 points)

Chattopadhyay and Duflo were interested in studying the causal effect of having female politicians on policy outcomes.[[4]](#footnote-4) Do women promote different policies than men? Answering this question with observational data is pretty difficult due to potential confounding problems (e.g. the districts that choose female politicians are likely to systematically differ in other aspects too). Hence, they exploit a randomized policy experiment in India, where since the mid-1990s,  of village council heads have been **randomly** reserved for women. A subset of the data from West Bengal can be found at the following link: <https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv>

Each observation in the data set represents a village and there are two villages associated with one GP (i.e. a level of government is called ”GP”). Figure 1 below shows the names and descriptions of the variables in the “women.csv” dataset. The authors hypothesize that female politicians are more likely to support policies female voters want. Researchers found that more women complain about the quality of drinking water than men. You will be asked to estimate the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages.

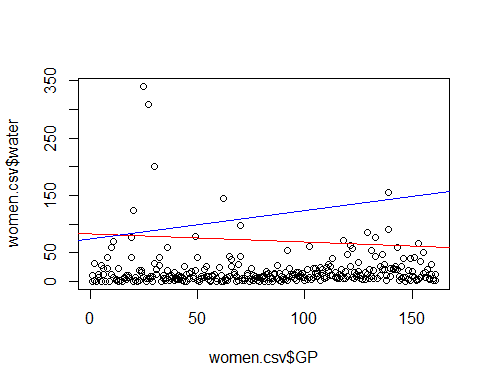
Figure 1: Names and description of variables from Chattopadhyay and Duflo (2004).



1. State a null and alternative (two-tailed) hypothesis.

Ho: μ=0

Ha: μ≠0

1. Run a bivariate regression to test this hypothesis in R (include your code!).

reserved <- subset(women.csv, women.csv$reserved=="1")

plot(reserved$GP, reserved$water)

corline <- lm(reserved$GP ~ reserved$water)

abline(corline, col = "red")

unreserved <- subset(women.csv, women.csv$reserved=="0")

unrescorline <- lm(unreserved$GP ~ unreserved$water)

plot(unreserved$GP, unreserved$water)

abline(unrescorline, col = "blue")

plot(women.csv$GP, women.csv$water)

abline(corline, col = "red")

abline(unrescorline, col = "blue")

unrescorline

corline

1. Interpret the coefficient estimate for reservation policy.

Contrary to hypothesized outcomes it was seen that villages that reserved positions for women were less likely to build more water facilities and villages that did not have a reservation policy were more likely to build water facilities.

Reserved policy on water facilities

r = -0.1461

Unreserved positions on number of water facilities

r = 0.4896

# Question 4 (25 points)

In this question, use the prestige dataset in the car library. First, run the following commands:

install.packages(car) library(car) data(Prestige) help(Prestige)

We would like to study whether individuals with higher levels of income have more prestigious jobs. Moreover, we would like to study whether professionals have more prestigious jobs than blue and white collar workers.

1. Create a new variable professional by recoding the variable type so that professionals are coded as 1, and blue and white collar workers are coded as 0 (Hint: ifelse.)

Prestige$binary = ifelse(Prestige$type=="prof",1,0)

1. Run a linear model with prestige as an outcome and income, professional, and the interaction of the two as predictors (Note: this is a continuous × dummy interaction.)

lm(Prestige$prestige ~ Prestige$income + Prestige$binary)

1. Write the prediction equation based on the result. Y=30.62+.001x1+.22.8x2
2. Interpret the coefficient for income.

For every dollar more you make in income, your professions prestige rises .001

1. Interpret the coefficient for professional.

For every professional job, your prestige is on avg 22.8points more prestigious than a wc or bc job

1. What is the effect of a $1,000 increase in income on prestige score for professional occupations? In other words, we are interested in the marginal effect of income when the variable professional takes the value of 1. Calculate the change in ˆ*y* associated with a $1,000 increase in income based on your answer for (c).

Y=30.62+.001x1+22.8x2

Y= 30.62 +.001(1000)+22.8(1)

Y=54.2

Y= 30.62 +.001(0)+22.8(1)

Y= 53.42

For professional occupations, a 1000$ increase in income means a .78 increase in prestige

1. What is the effect of changing one’s occupations from non-professional to professional when her income is $6,000? We are interested in the marginal effect of professional jobs when the variable income takes the value of 6*,*000. Calculate the change in ˆ*y* based on your answer for (c).

Y=30.62+.001x1+22.8x2

Y= 30.62 +.001(6000)+22.8(0)

Y=36.62

Y= 30.62 +.001(6000)+22.8(1)

Y= 59.4

When one who’s salary is 6000$ changes occupation from non professional to professional there is a 22.8 increase in prestige

# Question 5 (15 points)

In this question, you will run several regressions and interpret the model fit (*R*2) using the newhamp dataset in the faraway package.

library("faraway") data("newhamp") colnames(newhamp)

1. Run six regressions using pObama as a dependent variable and the following independent variable(s):
   1. votesys only

votepOb <- lm(newhamp$pObama ~ newhamp$votesys)

summary(votepOb)

* 1. votesys and povrate

sysrate <- lm(newhamp$pObama ~ newhamp$votesys + newhamp$povrate)

summary(sysrate)

* 1. votesys, povrate, and pci

sysratepci <- lm(newhamp$pObama ~ newhamp$votesys + newhamp$povrate + newhamp$pci)

summary(sysratepci)

* 1. votesys, povrate, pci, and Dean

sysratepcidean <- lm(newhamp$pObama ~ newhamp$votesys + newhamp$povrate + newhamp$pci + newhamp$Dean)

summary(sysratepcidean)

* 1. votesys, povrate, pci, Dean, and white

sysratepcidwhite <- lm(newhamp$pObama ~ newhamp$votesys + newhamp$povrate + newhamp$pci + newhamp$Dean + newhamp$white)

summary(sysratepcidwhite)

* 1. Dean only

deanPObama <- lm(newhamp$pObama ~ newhamp$Dean)

summary(deanPObama)

1. Get *R*2 and adjusted *R*2 from each of the six models above and fill in the table below. The first column is for *R*2s, and the second column is for adjusted *R*2s.

• Optional: If you summarize changes in *R*2 and adjusted *R*2 graphically using plot, you will receive 1 extra point, conditional on proper title, labels, and legend. There is no “correct” style!

*R*2 Adjusted *R*2

Model 1 .0834 .0806

Model 2 .08971 .08304

Model 3 .2441 .2358

Model 4 .5091 .5018

Model 5 .5093 .5002

Model 6 .4181 .416

1. Which model has the highest *R*2? What happens to *R*2 when you add a lot of variables in your model?
   1. Model 5 has the highest R2, which is also the model with the most variables. It seems that the more variables added them higher the R2

# Question 6 (25 points)

In this question, you will run several regressions and create an add variable plot (see the lecture slides from week 12) in R using the incumbents subset.csv dataset. Include all of your code.

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.

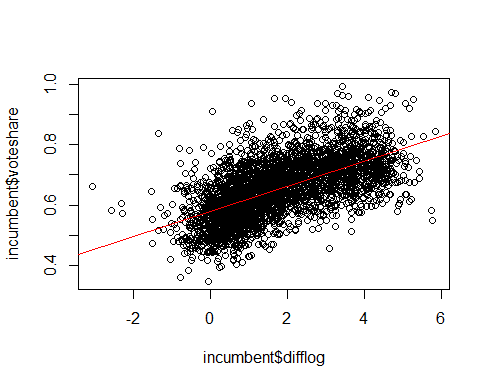
votesharediff <- lm(incumbent$voteshare ~ incumbent$difflog)

summary(votesharediff)

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

plot(incumbent$difflog, incumbent$voteshare)

abline(votesharediff, col = "red")



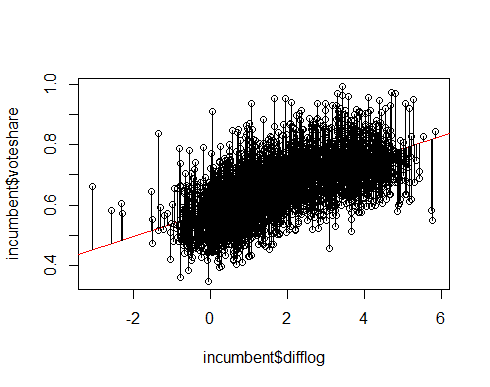
* 1. Save the residuals of the model in a separate object.

plot(incumbent$difflog, incumbent$voteshare)

abline(votesharediff, col = "red")

predshardiff <- predict(votesharediff)

segments(incumbent$difflog, incumbent$voteshare, incumbent$difflog, predshardiff)



* 1. Write the prediction equation

coef(votesharediff)

Y = .579 + .042x

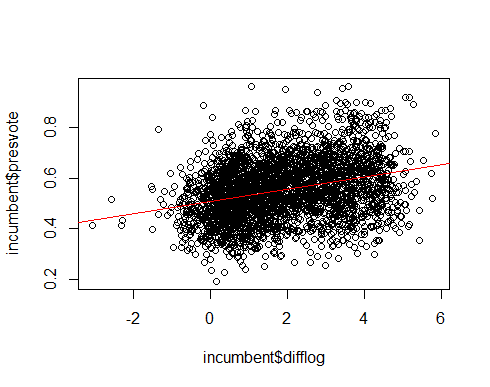
1. We are interested in knowing how the difference between incumbent and challenger’s spending and the vote share of the presidential candidate of the incumbent’s party are related.
   1. Run a regression where the outcome variable is presvote and the explanatory variable is difflog.

Presvotediff <- lm(incumbent$presvote ~ incumbent$difflog)

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

plot(incumbent$difflog, incumbent$presvote)

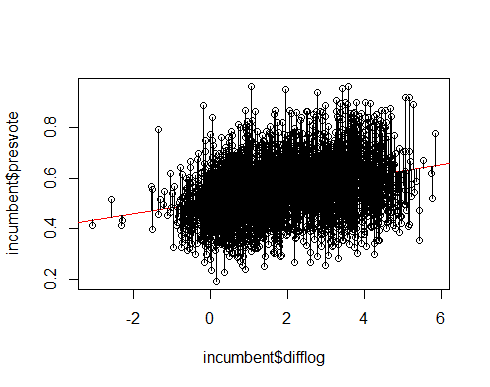
abline(presvotediff, col = “red”)



* 1. Save the residuals of the model in a separate object.

Predpresdiff <- predict(presvotediff)

Segments(incumbent$difflog, incumbent$presvote, incumbent$difflog, predpresdiff)



* 1. Write the prediction equation.

Y = .508 + .024x

1. 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.

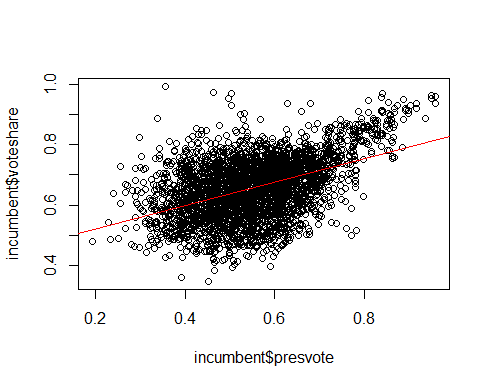
vsharepvote <- lm(incumbent$voteshare ~ incumbent$presvote)

summary(vsharepvote)

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

plot(incumbent$presvote,incumbent$voteshare)

abline(vsharepvote, col = "red")



* 1. Write the prediction equation.

Y= .44 + .39x

1. 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 part (a) and the explanatory variable is the residuals from part (b).

resvsharediff <- resid(votesharediff)

respresvdiff <- resid(presvotediff)

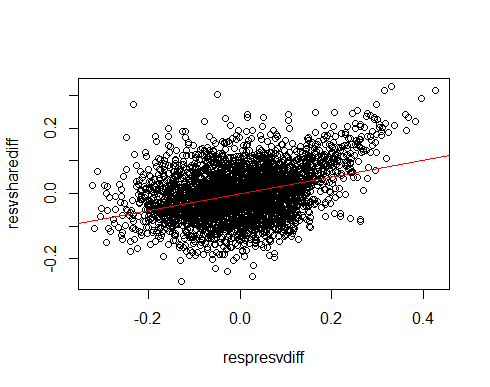
resregab <- lm(resvsharediff ~ respresvdiff)

summary(resregab)

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

plot(respresvdiff, resvsharediff)

abline(resregab, col = "red")



* 1. Write the prediction equation.

Y=.257x, Y intercept basically zero

1. 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.

vsharediffvote <- lm(incumbent$voteshare ~ incumbent$difflog + incumbent$presvote)

summary(vsharediffvote)

* 1. Write the prediction equation.

Y=.449+.036x1+.257x2

* 1. What is it in this output that is identical to the output in part (d)? Answer in one sentence!

The slope of x2 variable is the same as in problem d

* 1. Reflect on your finding. Don’t write anything. Just think about it.

1. Fried, Brian J, Paul Lagunes, and Atheendar Venkataramani. 2010. “Corruption and Inequality at the Crossroad: A Multimethod Study of Bribery and Discrimination in Latin America. *Latin American Research Review*. 45 (1): 76-97. [↑](#footnote-ref-1)
2. Remember frequency should be *>* 5 for all cells, but let’s calculate the p-value here anyway. [↑](#footnote-ref-2)
3. Donald P. Green, Jonathan S. Krasno, Alexander Coppock, Benjamin D. Farrer, Brandon Lenoir, Joshua N. Zingher. 2016. “The effects of lawn signs on vote outcomes: Results from four randomized field experiments.” Electoral Studies 41: 143-150. [↑](#footnote-ref-3)
4. Raghabendra Chattopadhyay and Esther Duflo. (2004). “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. Econometrica, Vol. 72, No. 5, pp. 1409-1443. [↑](#footnote-ref-4)