# 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] 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.

---

[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.

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

(a) Calculate the $\chi^2$ test statistic by hand (even better if you can do "by hand" in R).

(b) Now calculate the p-value (in R).[2] What do you conclude if $\alpha = .1$?

---

[2]Remember frequency should be $> 5$ for all cells, but let's calculate the p-value here anyway.

(c) 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 |  |  |  |

(d) How might the standardized residuals help you interpret the results?

# Question 2 (15 points)

Researchers are interested in learning the effect of all of those yard signs on voting preferences.[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

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

---

[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.

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

(c) Interpret the coefficient for the constant term substantively.

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

# Question 3 (10 points)

Chattopadhyay and Duflo were interested in studying the causal effect of having female politicians on policy outcomes.[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, $\frac{1}{3}$ 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).

| Name | Description |
| --- | --- |
| GP | An identifier for the Gram Panchayat (GP) |
| village | identifier for each village |
| reserved | binary variable indicating whether the GP was reserved for women leaders or not |
| female | binary variable indicating whether the GP had a female leader or not |
| irrigation | variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started |
| water | variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started |

[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.

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

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

(c) Interpret the coefficient estimate for reservation policy.

# 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.

(a) 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`.)

(b) 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.)

(c) Write the prediction equation based on the result.

(d) Interpret the coefficient for `income`.

(e) Interpret the coefficient for `professional`.

(f) 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 $\hat{y}$ associated with a \$1,000 increase in income based on your answer for (c).

(g) 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 $\hat{y}$ based on your answer for (c).

# 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)
```

(a) Run six regressions using `pObama` as a dependent variable and the following independent variable(s):

    (a) `votesys` only

    (b) `votesys` and `povrate`

    (c) `votesys`, `povrate`, and `pci`

    (d) `votesys`, `povrate`, `pci`, and `Dean`

    (e) `votesys`, `povrate`, `pci`, `Dean`, and `white`

    (f) `Dean` only

(b) 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 $R^2$s, and the second column is for adjusted $R^2$s.

- 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 |  |  |
| Model 2 |  |  |
| Model 3 |  |  |
| Model 4 |  |  |
| Model 5 |  |  |
| Model 6 |  |  |

(c) Which model has the highest $R^2$? What happens to $R^2$ when you add a lot of variables in your model?

# 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.

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

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

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

    (c) Save the residuals of the model in a separate object.

    (d) Write the prediction equation.

(b) 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.

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

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

    (c) Save the residuals of the model in a separate object.

    (d) Write the prediction equation.

(c) 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.

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

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

  (c) Write the prediction equation.

(d) 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.

  (a) Run a regression where the outcome variable is the residuals from part (a) and the explanatory variable is the residuals from part (b).

  (b) Make a scatterplot of the two residuals and add the regression line.

  (c) Write the prediction equation.

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

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

  (b) Write the prediction equation.

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

  (d) Reflect on your finding. Don't write anything. Just think about it.