Problem Set 3

Applied Stats/Quant Methods 1

Due: November 12, 2021

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 GitHub in `.pdf` form.
- This problem set is due before class on Friday November 12, 2021. No late assignments will be accepted.
- Total available points for this homework is 80.

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

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`.

```
install.packages("tidyverse") # install tidyverse for readr
library(tidyverse)

incumb <- read_csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2021/main/datasets/incumbents_subset.csv")
# read in data
str(incumb) # explore data
```

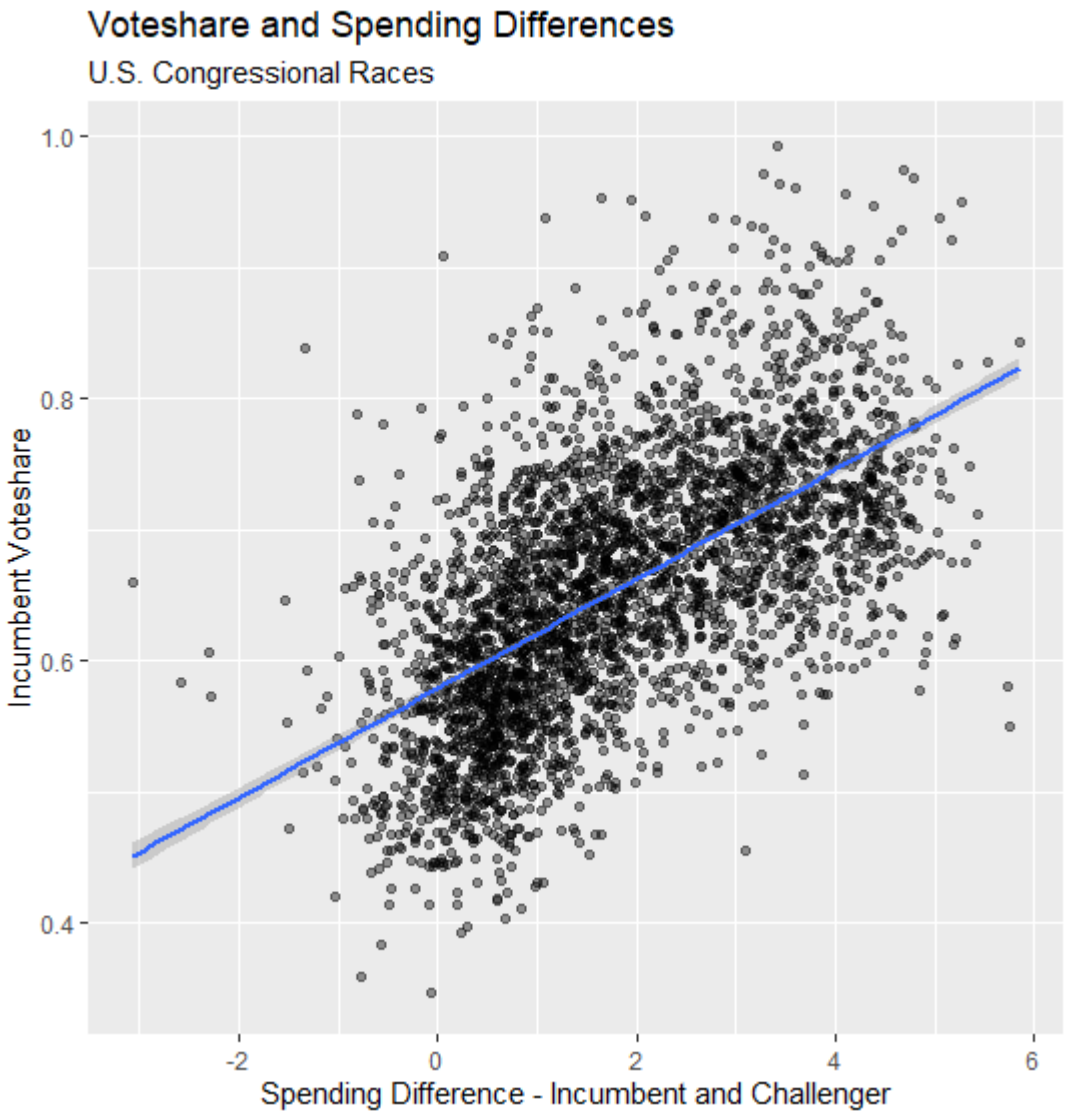
```r
diffvote <- lm(voteshare ~ difflog, data = incumb) # regress voteshare on
    difflog
summary(diffvote) # summarise regression
```

difflog has a coefficient of .042, SE of .00097, t-value of 43.04, and p-value of <0.0000000000000002 (essentially 0). The intercept is .579, with a SE of .00225, t-value of 257.19, and p-value of <0.0000000000000002 (essentially 0).

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

```r
diffvote_plot <- ggplot(aes(difflog, voteshare), data = incumb) +
  geom_point(alpha = 0.4) + # create scatter plot of voteshare and
    difflog
  geom_smooth(method = "lm", formula = y ~ x) + # draw regression line
  labs(title = "Voteshare and Spending Differences",
       subtitle = "U.S. Congressional Races",
       x = "Spending Difference - Incumbent and Challenger",
       y = "Incumbent Voteshare") # create titles and labels for x and y
    axes

diffvote_plot # display plot
```



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

```
diffvote <- lm(voteshare ~ difflog , data = incumb) # regress voteshare on
    difflog
diffvote_resid <- resid(diffvote)  # save residuals in separate object
```

4. Write the prediction equation.

```
summary(diffvote) # summarise regression
```

The prediction equation is: voteshare = .579 + .042*difflog

Question 2

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`.

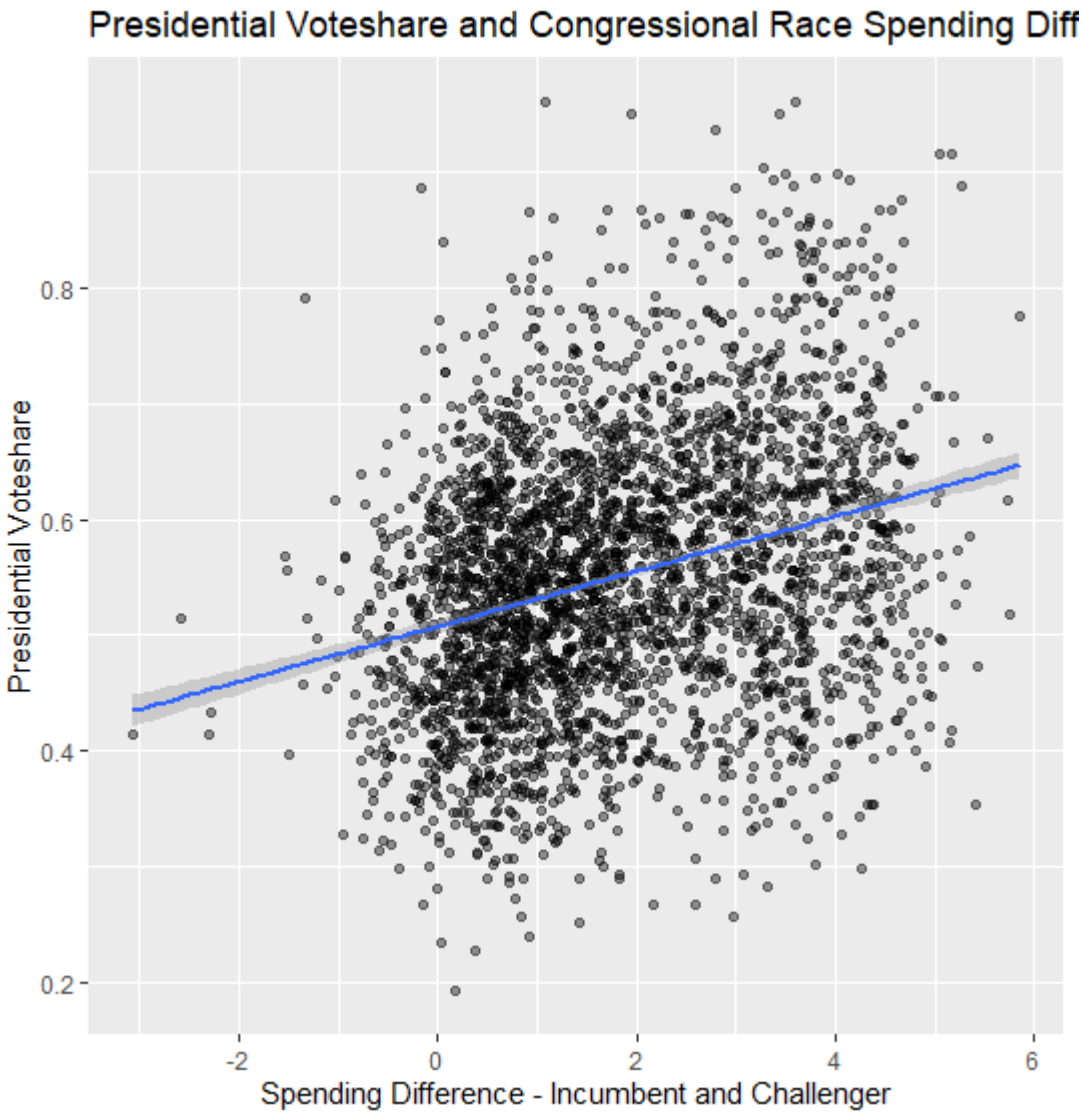
```
diffpres <- lm(presvote ~ difflog, data = incumb)  # regress presvote on
    difflog
summary(diffpres)  # summarise regression
```

difflog has a coefficient of .024, SE of .00136, t-value of 17.54, and p-value of <0.0000000000000002 (essentially 0). The intercept is .508, with a SE of .00316, t-value of 160.60, and p-value of <0.0000000000000002 (essentially 0).

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

```
diffpres_plot <- ggplot(aes(difflog, presvote), data = incumb) +
  geom_point(alpha = 0.4) +  # create scatter plot of presvote and
    difflog
  geom_smooth(method = "lm", formula = y ~ x) +  # draw regression line
  labs(title = "Presidential Voteshare and Congressional Race Spending
    Differences",
       x = "Spending Difference - Incumbent and Challenger",
       y = "Presidential Voteshare")  # create titles and labels for x
    and y axes

diffpres_plot  # display plot
```



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

```
diffpres <- lm(presvote ~ difflog, data = incumb)  # regress presvote on
    difflog
diffpres_resid <- resid(diffpres)  # save residuals in separate object
```

4. Write the prediction equation.

```
summary(diffpres) # summarise regression
```

The prediction equation is: presvote = .508 + .024*difflog

Question 3

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`.

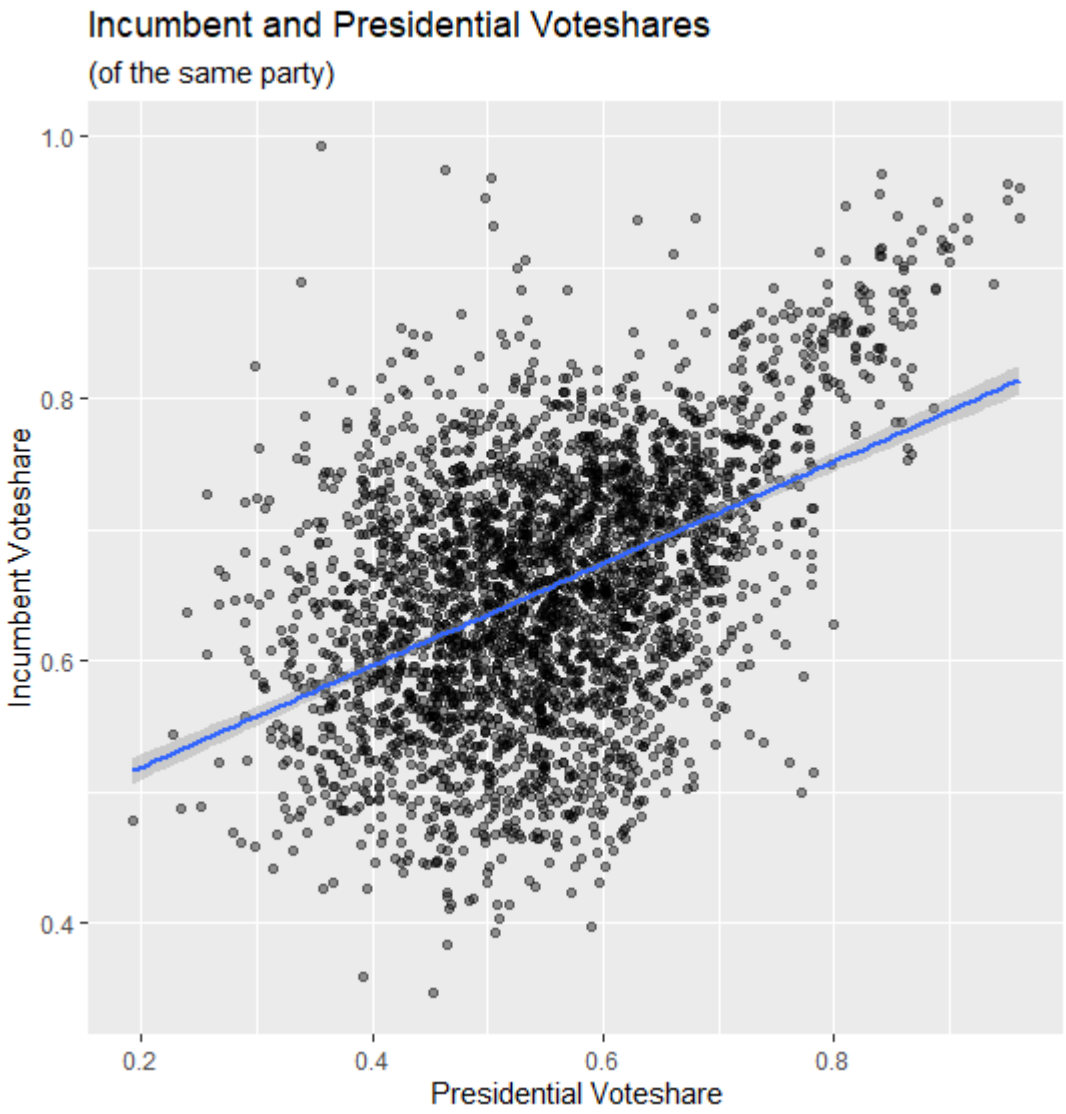
```
pres_voteshare <- lm(voteshare ~ presvote, data = incumb)  # regress
    voteshare on presvote
summary(pres_voteshare)  # summarise regression
```

   presvote has a coefficient of .388, SE of .01349, t-value of 28.76, and p-value of <0.0000000000000002 (essentially 0). The intercept is .441 with a SE of .00760, t-value of 58.08, and p-value of <0.0000000000000002 (essentially 0).

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

```
pres_voteshare_plot <- ggplot(aes(presvote, voteshare), data = incumb) +
  geom_point(alpha = 0.4) +  # create scatter plot of presvote and
    voteshare
  geom_smooth(method = "lm", formula = y ~ x) +  # draw regression line
  labs(title = "Incumbent and Presidential Voteshares",
       subtitle = "(of the same party)",
       x = "Presidential Voteshare",
       y = "Incumbent Voteshare")  # create titles and labels for x and y
     axes

pres_voteshare_plot  # display plot
```



3. Write the prediction equation.

```
summary(pres_voteshare) # summarise regression
```

The prediction equation is: voteshare = .441 + .388*presvote

Question 4

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.

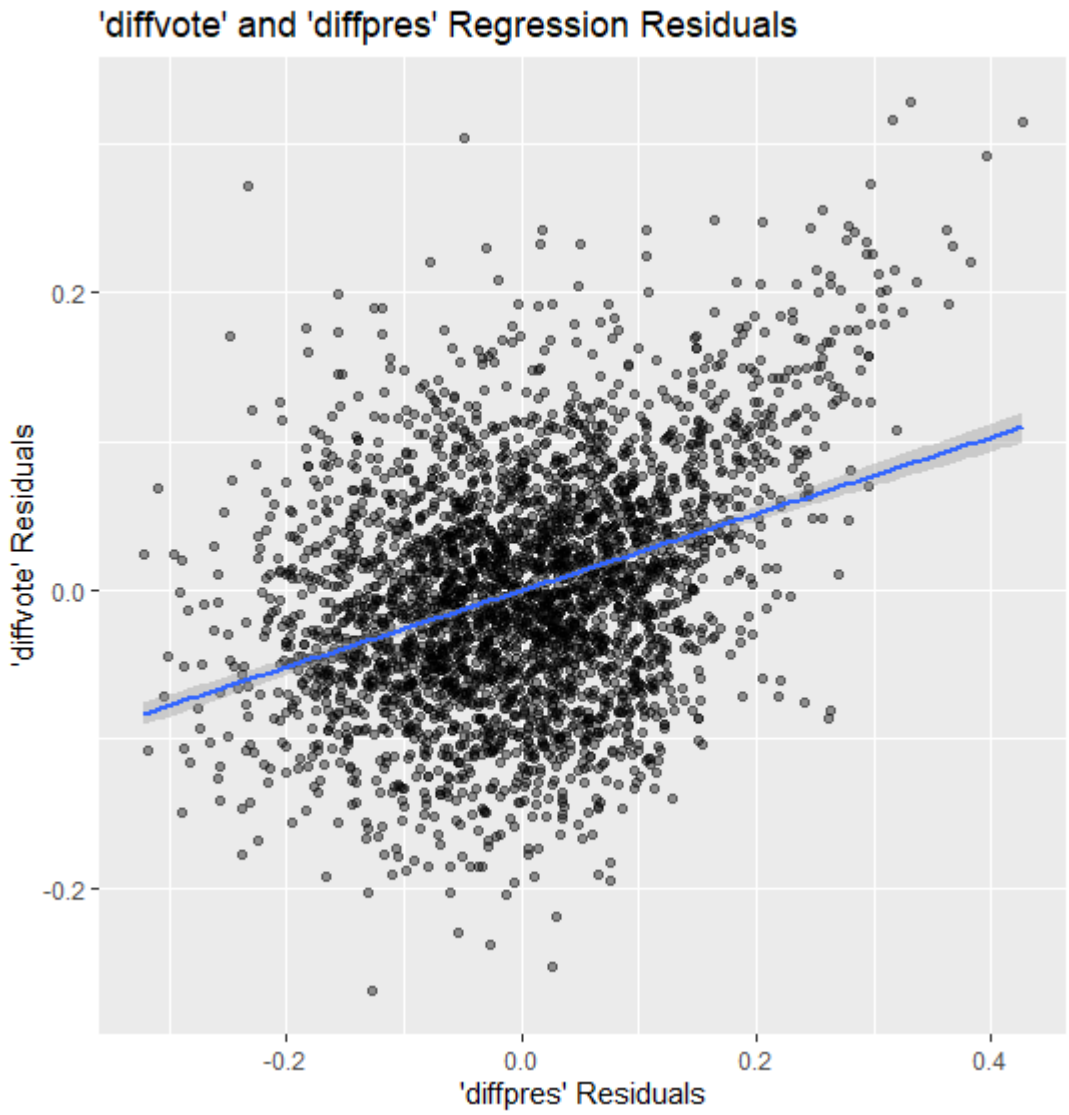
```
resid_rg <- lm(diffvote_resid ~ diffpres_resid)  # regress diffvote_resid
    on diffpres_resid
summary(resid_rg)  # summarise regression
```

   diffpres_resid has a coefficient of .257, SE of .01176, t-value of 21.84, and p-value of <0.0000000000000002 (essentially 0). The intercept is essentially 0, with a SE of .00130, t-value of 0, and p-value of 1 (i.e. there is no constant).

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

```
resid_rg <- ggplot(aes(diffpres_resid, diffvote_resid), data = NULL) +
  geom_point(alpha = 0.4) +  # create scatter plot of diffpres_resid and
    diffvote_resid
  geom_smooth(method = "lm", formula = y ~ x) +  # draw regression line
  labs(title = "Plot of 'diffvote' and 'diffpres' Regression Residuals",
       x = "'diffpres' Residuals",
       y = "'diffvote' Residuals")  # create titles and labels for x and
    y axes

resid_rg  # display plot
```



3. Write the prediction equation.

```
summary(resid_rg) # summarise regression
```

The prediction equation is: diffvote_resid = .257*diffpres_resid

Question 5

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`.

```
diff_pres_voteshare <- lm(voteshare ~ difflog + presvote, data = incumb)
    # regress voteshare on difflog and presvote
summary(diff_pres_voteshare)  # summarise regression
```

   difflog has a coefficient of .036, SE of .00095, t-value of 37.59, and p-value of <0.0000000000000002 (essentially 0). presvote has a coefficient of .257, SE of .01176, t-value of 21.84, and p-value of <0.0000000000000002 (essentially 0). The intercept is .449, with a SE of .00633, t-value of 70.88, and p-value of <0.0000000000000002 (essentially 0).

2. Write the prediction equation.

```
summary(diff_pres_voteshare) # summarise regression
```

   The prediction equation is: voteshare = .449 + .036*difflog + .257*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 coefficient, SE, t-value, and p-value for presvote (with voteshare as the response variable) in Question 5 are identical to those for diffpres_resid (with diffvote_resid as the response variable) in Question 4. They both have a coefficient of .257, SE of .01176, t-value of 21.84, and p-value of <0.0000000000000002 (essentially 0). This is the case because of how these variables are associated to one another. For Question 4, diffvote_resid is the component of variation within voteshare that cannot be explained by difflog. Meanwhile, diffpres_resid is the component of variation within presvote that also cannot be explained by difflog.

   In the Question 5 equation (voteshare = .449 + .036*difflog + .257*presvote), the Beta-1 term (.036*difflog) is accounting for the variation within voteshare that *is* associated with difflog. The model attributes the remaining variation to presvote. In the Question 4 model, both variables have removed the variation explained by difflog. In the Question 5 model, the control variable difflog similarly allows us to

see the variation in voteshare that is not attributable to difflog (the Beta-2 term). Both prediction equations can therefore look at the relationship between voteshare and presvote without difflog. Therefore, the coefficients, SEs, t-values, and p-values are the same for diffpres_resid and presvote.