Problem Set 3 - Answers

Sara Cid

Due: November 19, 2022

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.
- This problem set is due before 23:59 on Sunday November 19, 2023. No late assignments will be accepted.

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`.

   Code for the regression:

```
# Creating model using lm (I have used lm a lot in the past):
model1 <- lm(voteshare~difflog , inc.sub)
```

   Reporting regression results:

Table 1: Model 1 Regression Results

	VoteSh
DiffLog	0.042***
	(0.001)
Constant	0.579***
	(0.002)
Observations	3,193
R^2	0.367
Adjusted R^2	0.367
Note:	*p<0.05; **p<0.01; ***p<0.001

   Brief interpretation: The results show that a one-unit increase in difflog is associated, on average, with a 0.042 unit increase in voteshare (this is 4.2% higher). Since $p < 0.001$, we can reject the null hypothesis that there is no association between difflog and voteshare, or that the slope of difflog in this model is zero.

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

   Code for the scatterplot:

```
# And creating the scatterplot using ggplot (geom_smooth does do the same
    as abline):
plot1 <- ggplot(inc.sub, aes(x = difflog , y = voteshare)) +
  geom_point(color = "darkslategray4", alpha = 0.7) + # Editing point
    color and transparency
  geom_smooth(method = "lm", se = FALSE, color = "grey") + # Adding
    regression line (no SE)
```

```
  theme_minimal() +  # Applying theme
  labs(x = "Difference in campaign spending between incumbent and
   challenger",
       y = "Incumbent vote Share",
       title = "Scatterplot with Regression Line, Model 1") + # Fixing
   labels
  theme(panel.grid = element_blank())  # Getting rid of the grid
```

Showing the scatterplot:

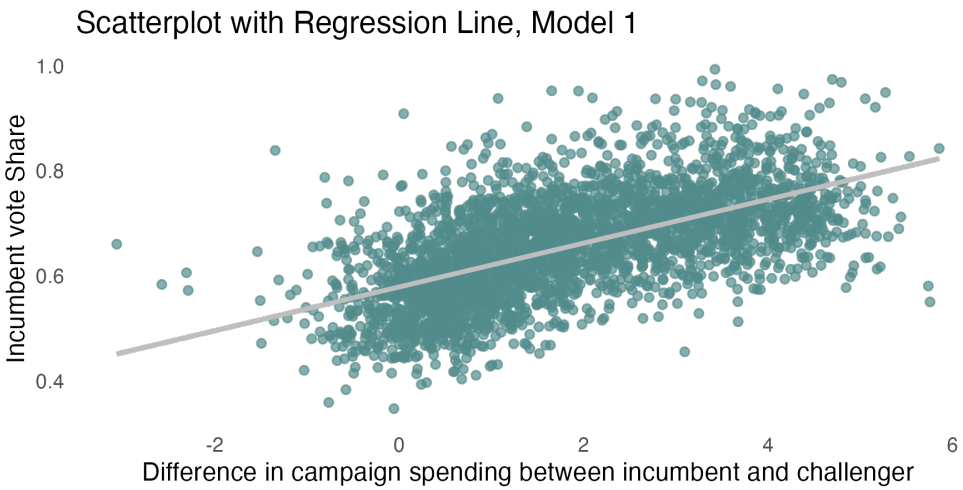

Figure 1: Model 1 Plot

Brief interpretation: The scatterplot is consistent with the results from the regression, showing a positive association between DiffLog and VoteShare. The shape that forms from the points is consistent with this as well, although the plot shows a significant amount of noise.

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

   Code for saving the residuals:

```
# Saving residuals separately:
resid1 <- model1$residuals
```

4. Write the prediction equation.

   $$\text{VoteSh} = 0.579 + 0.042 \cdot \text{DiffLog}$$

   where: $VoteSh$: Incumbent's vote share , and $DiffLog$: Logarithm of the difference between incumbent and challenger's spending

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`.

   Code for the regression:

```
# Creating and exploring model:
model2 <- lm(presvote~difflog, inc.sub)
summary(model2)
```

   Reporting regression results:

Table 2: Model 2 Regression Results

	PresVote
DiffLog	0.024***
	(0.001)
Constant	0.508***
	(0.003)
Observations	3,193
R^2	0.088
Adjusted R^2	0.088
Note:	*p<0.05; **p<0.01; ***p<0.001

   Brief interpretation: The results show that a one-unit increase in difflog is associated, on average, with a 0.024 unit increase in presvote (this is 2.4% higher). Since $p < 0.001$, we can reject the null hypothesis that there is no association between difflog and presvote, or that the slope of difflog in this model is zero.

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

   Code for the scatterplot:

```
# And creating the scatterplot using ggplot:
plot2 <- ggplot(inc.sub, aes(x = difflog, y = presvote)) +
  geom_point(color = "darkseagreen4", alpha = 0.7) + # Editing point
    color and transparency
```

```
  geom_smooth(method = "lm", se = FALSE, color = "grey") +  # Adding
    regression line (no SE)
  theme_minimal() +  # Applying theme
  labs(x = "Difference in campaign spending between incumbent and
    challenger",
       y = "Presidential candidate vote \n share (incumbent's party)",
       title = "Scatterplot with Regression Line, Model 2") + # Fixing
    labels
  theme(panel.grid = element_blank())  # Getting rid of the grid
```

Showing the scatterplot:

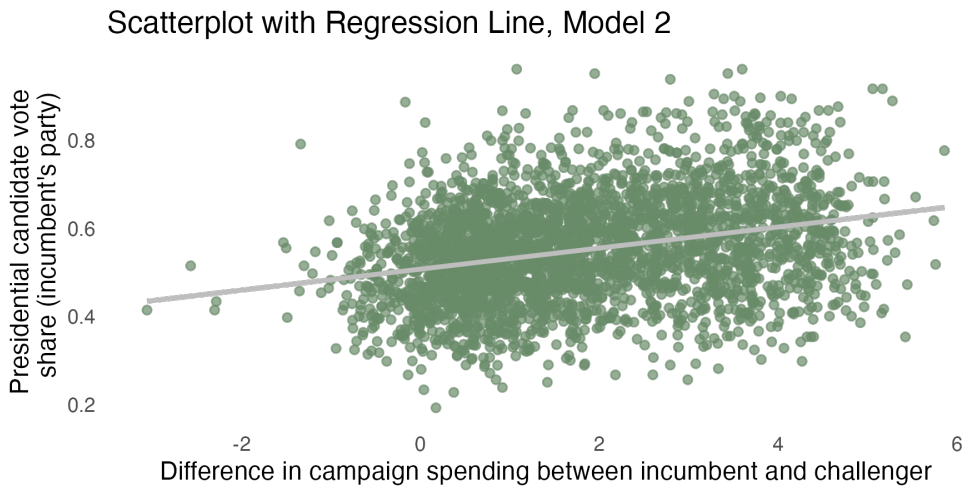

Figure 2: Model 2 Plot

Brief interpretation: Again, the scatterplot is consistent with the results from the regression, this time showing a positive association between DiffLog and PresVote. The shape that forms from the points is consistent with this as well, although this plot shows a considerable amount of noise (more than in the first one).

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

   Code for saving the residuals:

```
# Saving residuals separately:
resid2 <- model2$residuals
```

4. Write the prediction equation.

   $$\text{PresVote} = 0.508 + 0.024 \cdot \text{DiffLog}$$

   where $PresSh$: Vote share of the presidential candidate of the incumbent's party, and $DiffLog$: Logarithm of the difference between incumbent and challenger's spending

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`.

   Code for the regression:

```
# Creating and exploring model:
model3 <- lm(voteshare~presvote, inc.sub)
summary(model3)
```

   Reporting regression results:

Table 3: Model 3 Regression Results

	VoteSh
PresVote	0.388***
	(0.013)
Constant	0.441***
	(0.008)
Observations	3,193
R^2	0.206
Adjusted R^2	0.206
Note:	*p<0.05; **p<0.01; ***p<0.001

   Brief interpretation: The results show that a one-unit increase in presvote is associated, on average, with a 0.388 unit increase in voteshare (this is 38.8% higher). The coefficient is very large in magnitude. Since $p < 0.001$, we can reject the null hypothesis that there is no association between presvote and voteshare, or that the slope of presvote in this model is zero.

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

   Code for the scatterplot:

```
# Creating the scatterplot using ggplot:
plot3 <- ggplot(inc.sub, aes(x = presvote, y = voteshare)) +
  geom_point(color = "lightpink4", alpha = 0.7) +  # Editing point color
    and transparency
```

```
  geom_smooth(method = "lm", se = FALSE, color = "grey") +  # Adding
    regression line (no SE)
  theme_minimal() +  # Applying theme
  labs(x = "Presidential candidate vote share (incumbent's party)",
       y = "Incumbent's vote share",
       title = "Scatterplot with Regression Line, Model 3") + # Fixing
    labels
  theme(panel.grid = element_blank())  # Getting rid of the grid
```

Showing the scatterplot:

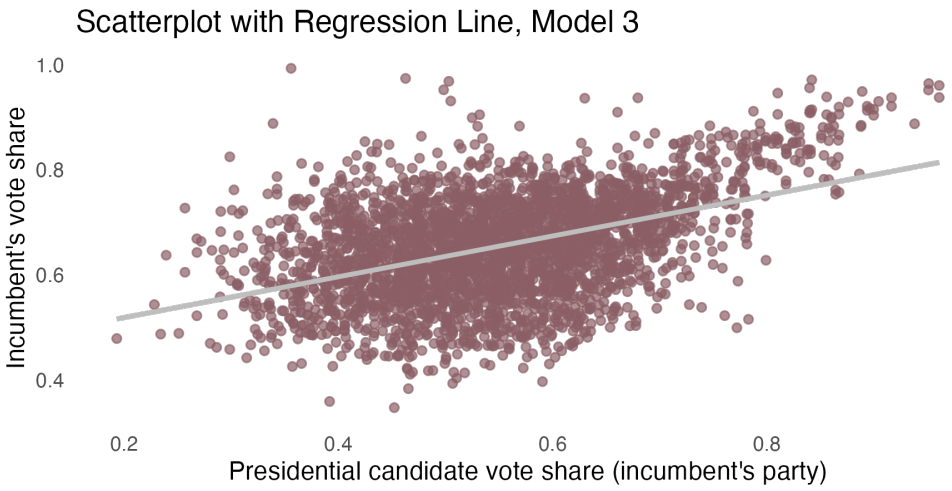

Figure 3: Model 3 Plot

Brief interpretation: Again, the scatterplot is consistent with the results from the regression, this time showing a positive association between PresVote and VoteShare. The shape that forms from the points is consistent with this as well, although there is a significant amount of noise once again.

3. Write the prediction equation.

$$\text{VoteSh} = 0.441 + 0.388 \cdot \text{PresVote}$$

where:

- $VoteSh$: Incumbent's vote share
- $PresVote$: Vote share of the presidential candidate of the incumbent's party

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.

   Code for the regression:

```
# Creating and exploring model:
model4 <- lm(resid1~resid2, inc.sub)
summary(model4)
```

   Reporting the regression results:

Table 4: Model 4 Regression Results

	Model1Res
Model2Res	0.257***
	(0.012)
Constant	−0.000
	(0.001)
Observations	3,193
R^2	0.130
Adjusted R^2	0.130
Note:	*p<0.05; **p<0.01; ***p<0.001

   Brief interpretation: The results show that a one-unit increase in Model2Res is associated, on average, with a 0.257 unit increase in Model1Res. Since $p < 0.001$, we can reject the null hypothesis that there is no association between Model2Res and Model1Res, or that the slope of Model2Res in this model is zero.

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

   Code for the scatterplot:

```
# Creating the scatterplot using ggplot:
plot4 <- ggplot(inc.sub, aes(x = resid2, y = resid1)) +
  geom_point(color = "thistle3", alpha = 0.7) +  # Editing point color
    and transparency
```

```
  geom_smooth(method = "lm", se = FALSE, color = "grey") +  # Adding
    regression line (no SE)
  theme_minimal() +  # Applying theme
  labs(x = "Residuals from Q2: Variation in PresVote not explained by
    difference in spending",
       y = "Residuals from Q1: Variation in VoteShare \n not explained by
     difference in spending",
       title = "Scatterplot with Regression Line, Model 4") + # Fixing
    labels
  theme(panel.grid = element_blank())  # Getting rid of the grid
```

Showing the scatterplot:

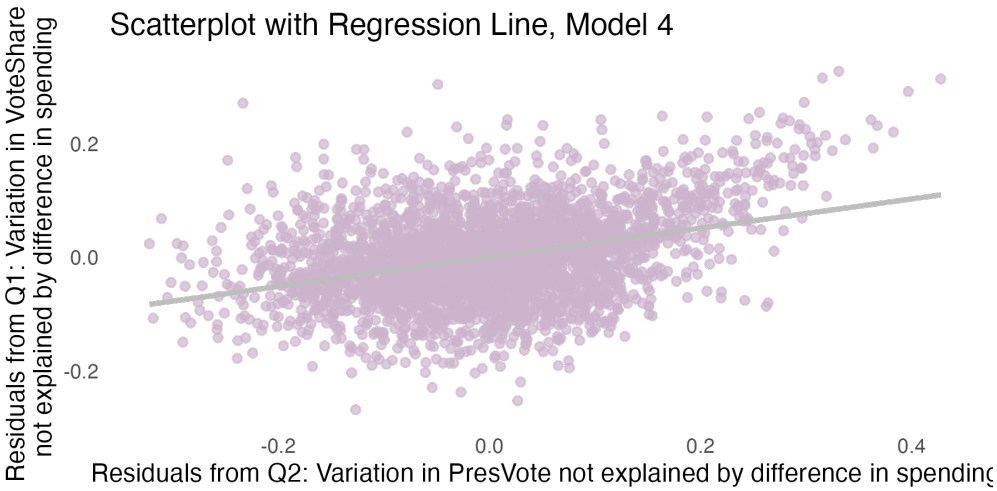

Figure 4: Model 4 Plot

Brief interpretation: Once more, the scatterplot is consistent with the results from the regression, this time showing a positive association between the residuals from Model2 and those of Model1. The shape that forms from the points is consistent with this as well, although there is a significant amount of noise.

3. Write the prediction equation.

$$\text{Resid1} = 0 + 0.257 \cdot \text{Resid2}$$

where:

- *Resid*1: Residuals from first model (variation in incumbent's vote share not explained by difference in spending)
- *Resid*2: Residuals from second model (variation in presidential candidate's vote share not explained by difference in spending)

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`.

   Code for the regression:

```
# Creating and exploring model:
model5 <- lm(voteshare~difflog+presvote, inc.sub)
summary(model5)
```

   Reporting regression results:

Table 5: Model 5 Regression Results

	VoteSh
DiffLog	0.036***
	(0.001)
PresVote	0.257***
	(0.012)
Constant	0.449***
	(0.006)
Observations	3,193
R^2	0.450
Adjusted R^2	0.449
Note:	*p<0.05; **p<0.01; ***p<0.001

   Brief interpretation: The results show that a one-unit increase in difflog is associated, on average and holding presvote constant, with a 0.036 unit increase in voteshare. Since $p < 0.001$, we can reject the null hypothesis that there is no association between difflog and voteshare, or that the slope of difflog in this model is zero. For presvote, the results show that a one-unit increase in presvote is associated, on average and holding difflog constant, with a 0.257 unit increase in voteshare. Since $p < 0.001$, we can reject the null hypothesis that there is no association between presvote and voteshare, or that the slope of difflog in this model is zero. In comparison with Model 1 and Model 2, where we tested the association of difflog and voteshare, and then presvote and voteshare separately, the coefficients in this last model are slightly smaller. This

is explained by the fact that presvote and difflog share some variability among each other, which we also proved in Model 3.

2. Write the prediction equation.

   $$\text{VoteSh} = 0.449 + 0.036 \cdot \text{DiffLog} + 0.257 \cdot \text{PresVote}$$

   where:

   - $VoteSh$: Incumbent's vote share
   - $DiffLog$: Logarithm of the difference between incumbent and challenger's spending
   - $PresVote$: Vote share of the presidential candidate of the incumbent's party

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

   Code to compare residuals:

```r
# Comparing Model 4 and Model 5:
# I think the residuals between the two models are going to be the same
# I will check for this.

# Naming my residuals from both models:
resid4 <- model4$residuals
resid5 <- model5$residuals

# I see in statisticsglobe (link provided below in R script) that I use identical()
# to compare
identical(resid4,resid5) # Not identical

# But perhaps they are not identical to the very last decimal point,
# however if I give a certain tolerance level (for example 0.001)
# the residuals will be the same.
# I see in stackoverflow (link provided below in R script)
# how to create the following function to compare each element
# in my residuals vectors, allowing for a small difference between them

# And I just have to sum the output
sum(mapply(function(element1, element2) abs(element1 - element2) <= 0.001,
           resid4, resid5))

# Double checking residuals length -> this means all residuals in M4 and M5 are the same
length(resid4)
length(resid5)
```

First, the coefficient for presvote in Model 5 is the same as that of Resid2 in Model 4. This is because both coefficients describe the relationship between the unexplained variation in presvote and the unexplained variation in voteshare after accounting for difflog.

In Model 4, the coefficient of Resid2 describes the relationship between the unexplained variation in presvote after accounting for difflog (Resid2) and the unexplained variation in voteshare after accounting for difflog (Resid1) because we are running a regression of Resid1 over Resid2.

In Model 5, the coefficient of presvote captures essentially the same thing: how presvote varies with share, after accounting for difflog, which is the other independent variable in the multivariate model.

Also, the residuals in Models 4 and 5 are the same, because since our outcome in Model 4 is the unexplained variation in voteshare after including difflog in our regression, and since our explanatory variable in Model 4 is the unexplained variation in presvote after including difflog in our regression, the residuals of this model are exactly the same thing as the residuals in Model 5, where we run a multivariate regression of voteshare over difflog and presvote.

That is, the residuals in both cases consist of the unexplained variation in voteshare after accounting for both difflog and presvote.