Problem Set 3

Applied Stats/Quant Methods 1

Due: November 11, 2024

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 11, 2024. 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`.

   Before running the regression, we are interested in exploring these variables in order to understand their nature. From the name of our explanatory variable, `difflog`, we assume this variable has been transformed into a log-transformed one. This will be important later to correctly interpret our coefficients.

   By exploring our variables, we see that generally these two variables show a relatively symmetric distribution: variable `difflog` is slightly positively skewed, as the median is slightly lower than the mean. Furthermore, the difference between the third quartile and the maximum suggest that the high values pull the mean higher.

```r
summary(inc.sub$difflog)
summary(inc.sub$voteshare)
plot(inc.sub$difflog)
plot(inc.sub$voteshare)
hist(inc.sub$difflog)
hist(inc.sub$voteshare)
```

```
> summary(inc.sub$difflog)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.0601  0.6734  1.6106  1.8278  2.9857  5.8558
> summary(inc.sub$voteshare)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.3476  0.5846  0.6569  0.6552  0.7233  0.9930
```

Additionally, we also plot the variables to see what kind of relationship these two variables show. From the plot, we hypothesise that variable `difflog` probably has a positive effect on variable. We will confirm this first assumption when we run the regression. `voteshare`

```r
scatter <- ggplot(data = inc.sub,
                  mapping = aes (x = difflog,
                                 y = voteshare),
                                size = sec_enrol) +
  geom_point()  +
  labs(x = "Campaign Spending (log)",
       y = "Vote Share") +
  theme_classic() +
  theme(legend.box.background = element_rect(size = 0.1),
        legend.position = c(0.85, 0.85))
ggsave("scatter.png", width = 10, height = 6)
scatter
```

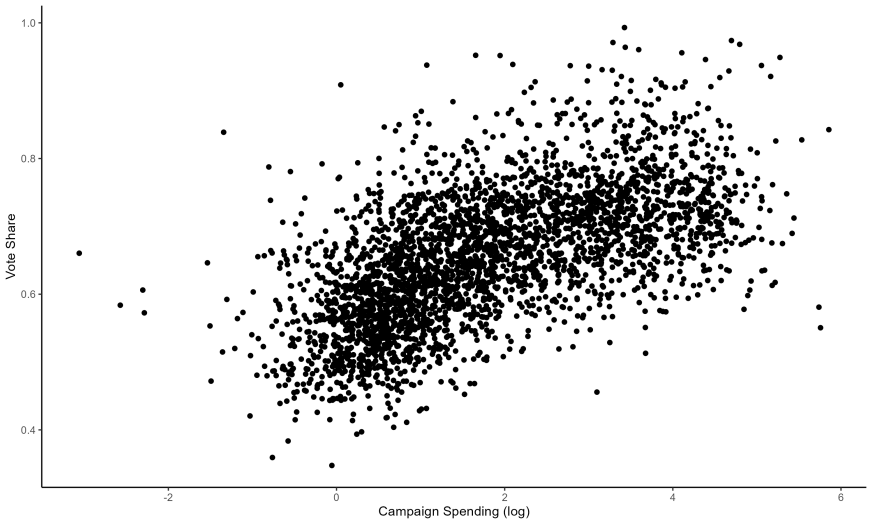

Figure 1: Effect of campaign spending on vote share

Before running the regression, we establish that our hypothesis are as follows:

H_0: There is no statistically relevant relationship between the difference in campaign spending between incumbent and challenger party and the vote share of the incumbent party.

H_1: There is a statistically relevant relationship between the difference in campaign spending between incumbent and challenger party and the vote share of the incumbent party.

We run the regression:

```
m1 <- lm(voteshare ~ difflog, data= inc.sub)
stargazer(m1, type = "latex")
```

First, we observe that our constant is 0.579. That is, that when the difference in campaign spending between the incumbent and the challenger party is 0 units, the incumbent's vote share is 0.579 units, or roughly 58%. However, when the explanatory variable experiences a one unit increase (which corresponds to around a 100% increase in the original "non log-transformed" variable) correlates with a 0.042 unit increase in the outcome variable. That is when the difference in campaign spending between incumbent and challenger increases one unit, the vote share for the incumbent party increases, on average, 0.042 units. These relationship is statistically relevant at the 1% level. That is, with these estimates, we have enough evidence to reject H_0: we do not have enough evidence to support that there is not a statistically relevant relation between the difference in campaign spending and the vote share of the incumbant party. The R-squared is 0.367. This represents that variability we observe in our explanatory variable explains 36,7% of the variability observed in our outcome variable.

Table 1:

	Dependent variable:
	voteshare
difflog	0.042***
	(0.001)
Constant	0.579***
	(0.002)
Observations	3,193
R^2	0.367
Adjusted R^2	0.367
Residual Std. Error	0.079 (df = 3191)
F Statistic	1,852.791*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

   As we have assessed before, the relationship that share the two variables is a positive one: `difflog` covariates positively with `voteshare`.

```
scatter2 <- ggplot(data = inc.sub,
                   mapping = aes(x = difflog,
                                 y = voteshare)) +
  geom_point(aes()) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Campaign Spending (log)",
       y = "Vote Share") +
  theme_classic() +
  theme(legend.box.background = element_rect(size = 0.1),
        legend.position = c(0.85, 0.85))
ggsave("scatter2.png", width = 10, height = 6)
scatter2
```

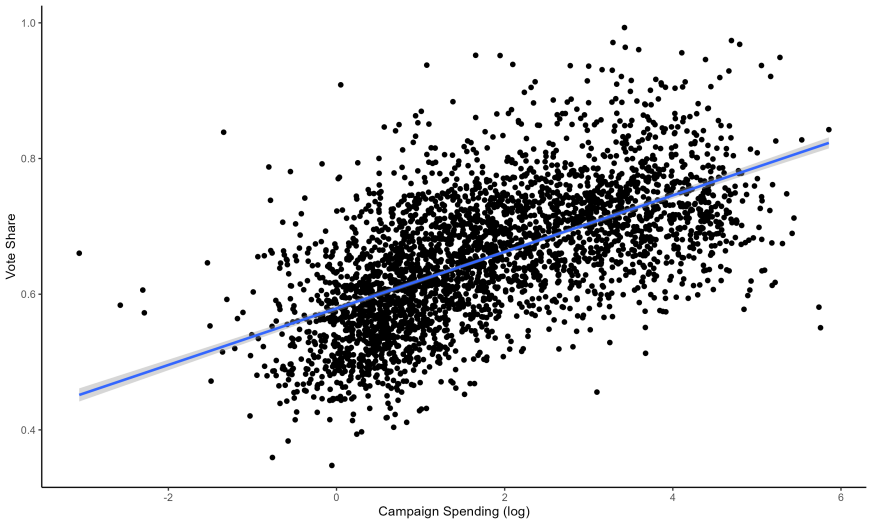

Figure 2: Effect of campaign spending on vote share

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

```
res1 <- residuals(m1)
res1
```

4. Write the prediction equation.

   The general equation for a bivariate regression is the following:

$$y = \beta_0 + \beta_1 \cdot x + e \tag{1}$$

   Here:

   - y represents the value we are trying to predict for the variable `voteshare`.

   - β_0 is the intercept, representing the value of `voteshare` when `difflog` is 0.

   - β_1 is the slope coefficient, which measures the change in y (i.e., `voteshare`) for each unit increase in x (i.e., `difflog`).

   - x represents our explanatory variable, `difflog`.

   - e is the error term.

   Applying the equation to our regression coefficients, we get:

$$y = 0.579 + 0.042 \cdot \text{difflog} + e \tag{2}$$

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`.

   Again, we first explore variable `presvote` to assess its nature. We see that the distribution of this variable can be considered as normally distributed.

```
summary(inc.sub$presvote)
h2 <- hist(inc.sub$presvote)
```

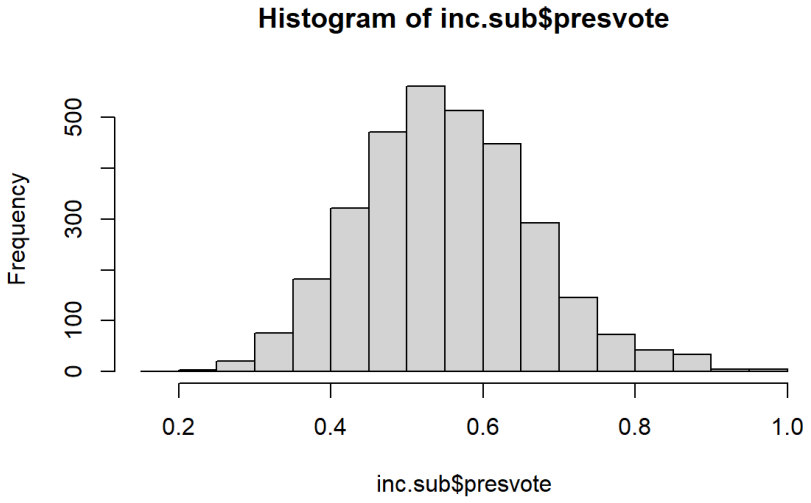

Figure 3: Distribution of the variable "presvote"

```
summary(inc.sub$presvote)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.1931  0.4695  0.5478  0.5512  0.6240  0.9606
```

Before we run the regression, we stipulate our hypothesis:

H_0: There is no statistically relevant relation between the difference between incumbent and challenger's campaign spending and the vote share of the presidential candidate of the incumbent's party.

H_1: There is a statistically relevant relation between the difference between incumbent and challenger's campaign spending and the vote share of the presidential candidate of the incumbent's party.

We run the regression:

```
m2 <- lm(presvote ~ difflog, data = inc.sub)
stargazer(m2, type="latex")
```

Table 2:

	Dependent variable:
	presvote
difflog	0.024^{***}
	(0.001)
Constant	0.508^{***}
	(0.003)
Observations	3,193
R^2	0.088
Adjusted R^2	0.088
Residual Std. Error	0.110 (df = 3191)
F Statistic	307.715^{***} (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

First, we observe that our constant is 0.508. That indicates us that when the difference in campaign spending between the incumbent and the challenger party is 0, the presidential vote share is 0.508. However, when the difference in campaign spending increases in one unit, the vote share of the presidential candidate increases, on average, 0.024 units. Both coefficients are statistically relevant at the 1% level. That is, with these estimates, we have enough evidence to reject H_0: we do not have enough evidence to support that there is not a statistically relevant relation between the difference in campaign spending and the vote share of the presidential candidate.

The R-squared is 0.088. This represents that variability we observe in our explanatory variable explains 8,8% of the variability observed in our outcome variable.

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

```
scatter3 <- ggplot(data = inc.sub,
```

```
                    mapping = aes(x = difflog,
                                  y = presvote)) +
  geom_point(aes()) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Campaign Spending (log)",
       y = "Vote Share of Pres. Cand.") +
  theme_classic() +
  theme(legend.box.background = element_rect(size = 0.1),
        legend.position = c(0.85, 0.85))
ggsave("scatter3.png", width = 10, height = 6)
scatter3
```

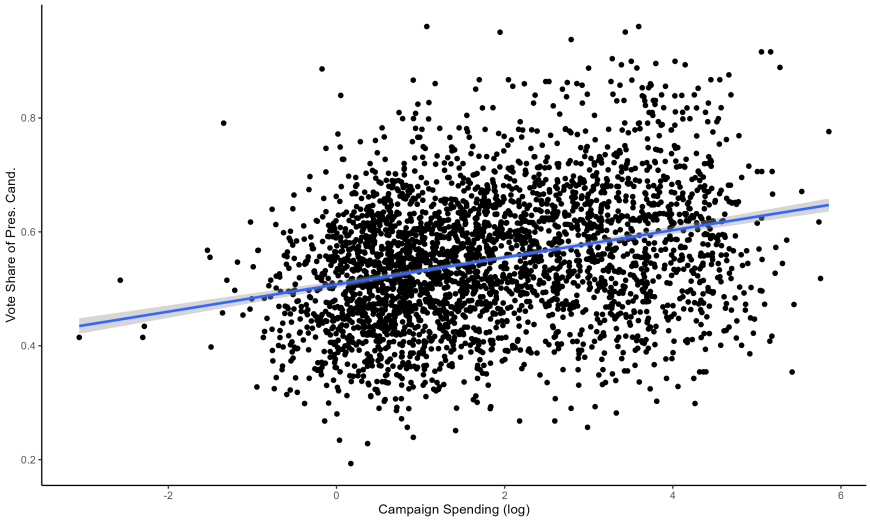

Figure 4: Plotted relation between "presvote" and "difflog"

The relationship both these variables share is a positive one. As indicated by our regression coefficients, the variable "difflog" and the variable "presvote" covary in a positive way: when the variable "difflog" increases in one unit, the variable "presvote", on average, also increases. Specifically, it increases 0.508 units.

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

```
res2 <- residuals(m2)
res2
```

4. Write the prediction equation.

   The general equation for a bivariate regression is the following:

(a) Write the prediction equation.

$$y = \beta_0 + \beta_1 \cdot x + e \tag{3}$$

Here:

- y represents the value we are trying to predict for the variable `presvote`.
- β_0 is the intercept, representing the value of `presvote` when `difflog` is 0.
- β_1 is the slope coefficient, which measures the change in y (i.e., `presvote`) for each unit increase in x (i.e., `difflog`).
- x represents our explanatory variable, `difflog`.
- e is the error term.

Applying the equation to our regression coefficients, we get:

$$y = 0.579 + 0.042 \cdot \text{difflog} + e \tag{4}$$

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`.

   In this case, we do not need to have a first exploration of these variables, as we have done so previously, and we have already assessed that these variables present a relatively normal distribution.

   Before running the regression, we establish that our hypothesis are as follows:

   H_0: There is no statistically relevant relationship between the vote share of the presidential candidate of the incumbent's party and the incumbent's electoral success.

   H_1: There is a statistically relevant relationship between the vote share of the presidential candidate of the incumbent's party and the incumbent's electoral success.

   We proceed to run the bivariate regression:

```
m3 <- lm(voteshare ~ presvote, data=inc.sub)
stargazer(m3, type="latex")
```

Table 3:

	Dependent variable:
	voteshare
presvote	0.388***
	(0.013)
Constant	0.441***
	(0.008)
Observations	3,193
R^2	0.206
Adjusted R^2	0.206
Residual Std. Error	0.088 (df = 3191)
F Statistic	826.950*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

   In this case, we see that our constant is 0.441. That is, when variable `presvote` is 0, the outcome variable is situated at 0.441. If this variable is expressed in percentage points, we could say when the average vote share for the presidential candidate is 0, the vote

share for the incumbent is 44.1%. The coefficient of variable `presvote` is 0.388. As we can observe, it is positive and statistically relevant at the 1% confidence level. That is, for every one unit increase in our explanatory variable `presvote`, the outcome variable `voteshare` increases, on average, 0.388 units. Due to the fact that this coefficient is statistically relevant, we are able to rule out our initial null hypothesis: we have enough evidence to reject that there is no statistically relevant association between the vote share of the presidential candidate of the incumbent's party and the incumbent's electoral success. The R-squared also denotes that around 20,6% of the variability we observe in our outcome variable is explained by our explanatory variable.

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

```
scatter4 <- ggplot(data = inc.sub,
                   mapping = aes(x = presvote,
                                 y = voteshare)) +
  geom_point(aes()) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Vote Share of Pres. Cand.",
       y = "Vote Share Incument Party") +
  theme_classic() +
  theme(legend.box.background = element_rect(size = 0.1),
        legend.position = c(0.85, 0.85))
ggsave("scatter4.png", width = 10, height = 6)
scatter4
```

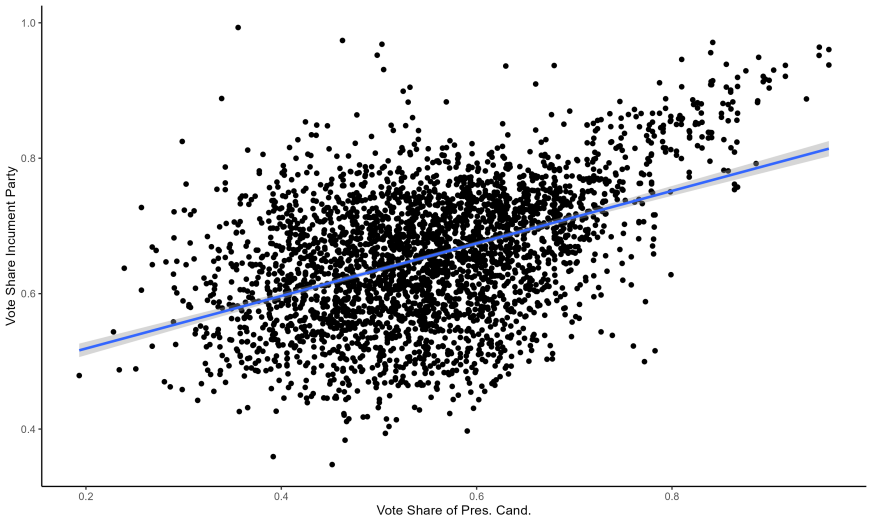

Figure 5: Plotted relation between "presvote" and "voteshare"

3. Write the prediction equation.

$$y = \beta_0 + \beta_1 \cdot x + e \tag{5}$$

Here:

- y represents the value we are trying to predict for the variable `voteshare`.

- β_0 is the intercept, representing the value of `voteshare` when `presvote` is 0.

- β_1 is the slope coefficient, which measures the change in y (i.e., `voteshare`) for each unit increase in x (i.e., `presvote`).

- x represents our explanatory variable, `presvote`.

- e is the error term.

Applying the equation to our regression coefficients, we get:

$$y = 0.441 + 0.388 \cdot \text{difflog} + e \tag{6}$$

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.

   Before we run the regression, we stipulate our hypothesis:

   H_0: There is no statistically relevant association between the difference between residuals of our first model (m1) and the residuals of our second model (m2).

   H_1: There is a statistically relevant association between the difference between residuals of our first model (m1) and the residuals of our second model (m2).

```
m5 <- lm(res1 ~ res2)
stargazer(m5, type = "latex")
```

Table 4:

	Dependent variable:
	res1
res2	0.257***
	(0.012)
Constant	−0.000
	(0.001)
Observations	3,193
R^2	0.130
Adjusted R^2	0.130
Residual Std. Error	0.073 (df = 3191)
F Statistic	476.975*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

As we can appreciate, the intercept of the constant of this regression is -0.000. This is to be expected, as in residuals are usually centred around zero. The coefficient of the residuals of our second model are positive and statistically relevant. Therefore, we have enough evidence to reject our null-hypothesis: we have enough proof to reject that there isn't a statistically relevant association between the residuals of our first model and the residuals of our second model. This could mean that residuals from our

second model (res2) explain some of the variation we observe in our first model (res1). This could be understood as that there might be an underlying structure or factors not accounted or not included, such as omitted variables, in our models. If we take a look at our R-squared, we also observe that around 13% of the variation we observe in the residuals of our first model (res1) are explained by the residuals of our second model (res2).

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

```
scatter5 <- ggplot( mapping = aes(x = res2,
                                  y = res1)) +
  geom_point(aes()) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Residuals Q2",
       y = "Residuals Q1") +
  theme_classic() +
  theme(legend.box.background = element_rect(size = 0.1),
        legend.position = c(0.85, 0.85))
ggsave("scatter5.png", width = 10, height = 6)
scatter5
```

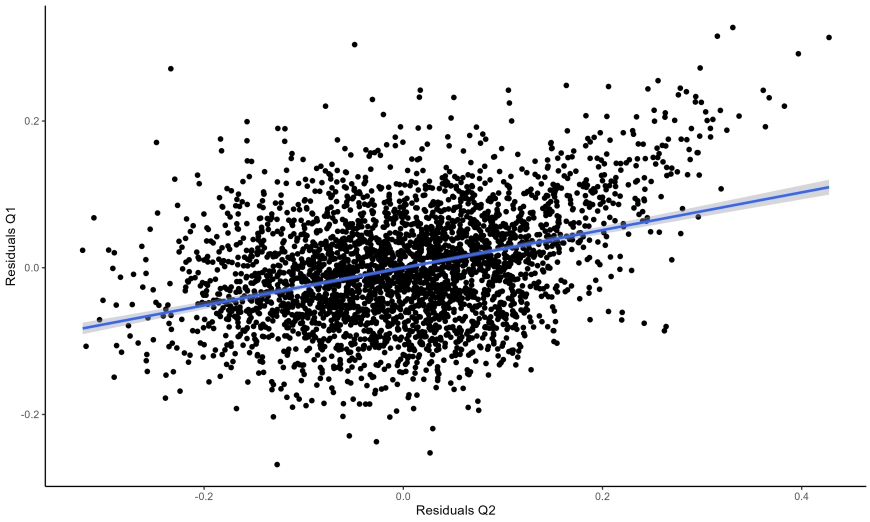

Figure 6: Plotted relation between "res1" and "res2"

3. Write the prediction equation.

$$res1 = 0.257 \cdot res2 - 0.000 + \epsilon \tag{7}$$

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`.

   Before I run the multivariate regression, I state the null and alternative hypothesis, which differ slightly from the hypothesis we state in bivariate regressions:

   H_0: There is no statistically relevant association between the explanatory variables, both `difflog` and `presvote`, and the outcome variable, `voteshare`. That is, there is no statistically relevant association between the difference in either campaign spending or the presidential candidate's vote share, and the vote share of the incumbent party.

   H_1: There is a statistically relevant association between the explanatory variables, both `difflog` and `presvote`, and the outcome variable, `voteshare`. That is, there is a statistically relevant association between the difference in either campaign spending or the presidential candidate's vote share, and the vote share of the incumbent party.

```
m6 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
stargazer (m6, type = "latex")
```

   From this last regression, we observe that the constant is positive and statistically relevant, situated at 0.449. That means that when `difflog` and `presvote` is 0, variable `voteshare` is 0.449 units. That is, when the difference in campaing spending and the presidential candidate's vote share are both 0, the vote share for the incumbent party is 0.449. If we observe the coefficient of variable `difflog`, we also notice that it is positive and statistically relevant at the 1% level. If we hold the other variables constant, i.e., `presvote`, we could affirm that for one unit increase in variable `difflog`, variable `voteshare` experiences an increase of 0.036 units, on average. In variable `presvote`, we also see that the coefficient is positive and statistically relevant at the 1% level. Holding `difflog` constant, we could affirm that for every unit increase in variable `presvote`, variable `voteshare` is expected to increase, on average, 0.257. Considering that our coefficients turn out to be statistically relevant, we can reject the null-hypothesis we previously stated. We have enough evidence to reject that both our explanatory variables do not have any statistically relevant association with the outcome variable. Considering our R-squared, we can state that about 45% of the variation we observe in variable `voteshare` is explained by the outcome variables.

Table 5:

	Dependent variable:
	voteshare
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
Residual Std. Error	0.073 (df = 3190)
F Statistic	1,302.947*** (df = 2; 3190)
Note:	*p<0.1; **p<0.05; ***p<0.01

2. Write the prediction equation.

$$voteshare = 0.449 + 0.036 \cdot difflog + 0.257 \cdot presvote + \epsilon \tag{8}$$

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

   As we can observe, the coefficient of variable `presvote` in our fifth model is the same as the coefficient presented by the residuals of model 2 we have regressed in our fourth model (Table 4). As we discussed earlier, in Table 4, the unexplained variability of model 2 (res2) was positively correlated with the unexplained variation in `voteshare`. The fact that in Table 5 variable `presvote` presents the same coefficient suggests that the unexplained variation observered in `voteshare` is closely related to the variation observed in `presvote`. Furthermore, the addition on variable `difflog` does not seem to alter the relationship between variable `voteshare` and variable `presvote`. In sum, the similarity in coefficients suggests that the part of `presvote` that is not explained

by campaign spending differences (model 4) is still statistically relevant in predicting the incumbent's vote share (model 5).