Problem Set 3

Stefan Keel
Applied Stats/Quant Methods 1

Due: November 19, 2023

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`.

   I run the regression with the inbuilt "lm" function whilst subsetting "voteshare" as Y (outcome variable) and "difflog" as X (explanatory variable). This is the code I used:

   ```
   model_1 <- lm(inc.sub$voteshare~inc.sub$difflog)
   summary_1 <- summary(model_1)
   ```

Table 1:

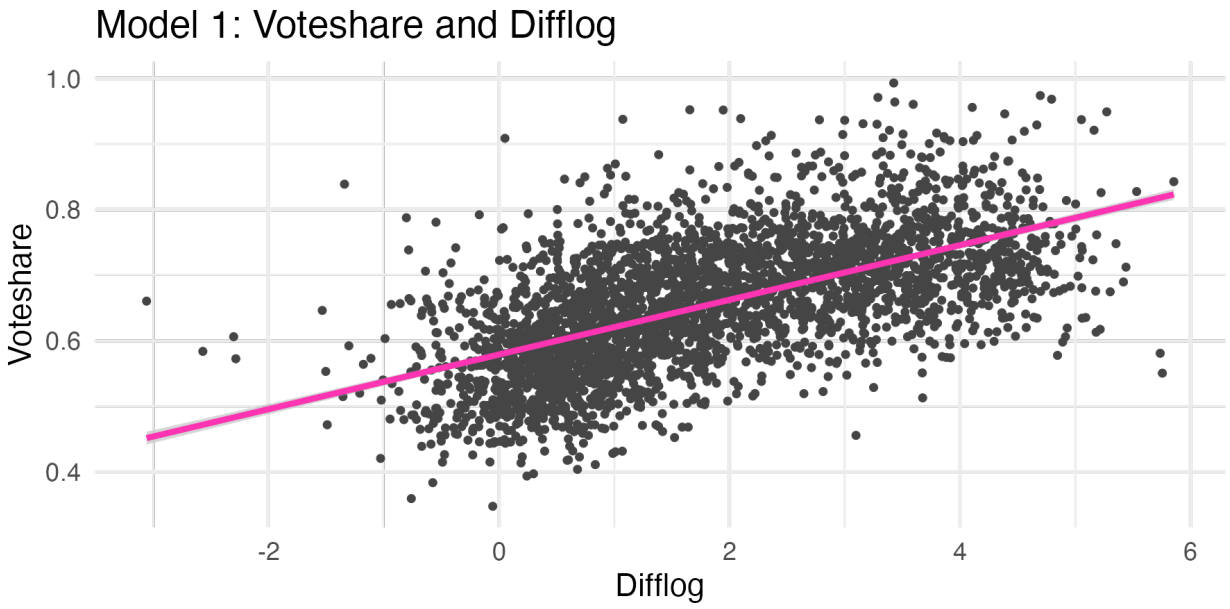
	Dependent variable:
	voteshare
difflog	0.042***
	(0.001)
Constant	0.579***
	(0.002)
Observations	3,193
R^2	0.367
Adjusted R^2	0.367
Residual Std. Error	0.079 (df = 3191)
F Statistic	1,852.791*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

   This is the code used to plot:

```
scatter_1 <- ggplot(data = inc.sub,
mapping = aes(x = difflog, y = voteshare)) +
labs(x = "Difflog", y = "Voteshare",
title = "Model 1: Voteshare and Difflog") +
theme_minimal() +
geom_point(color = "grey27", size = .8) +
geom_smooth(method = 'lm',col="maroon1")
```

Figure 1: Voteshare and Difflog



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

   I am saving the residuals of model 1 in the separate object . I used this code:

   ```
   residuals_1 <- model_1$residual
   ```

4. Write the prediction equation.

   Using the results from the summary (see solution 1.1) the prediction equation is: $\hat{y} = \mathbf{0.579} + \mathbf{0.042}\mathbf{X}_1$

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`.

   I run the regression with the inbuilt "lm" function whilst subsetting "presvote" as Y (outcome variable) and "difflog" as X (explanatory variable). This is the code I used:

```
model_2 <- lm(inc.sub$presvote~inc.sub$difflog)
summary_2 <- summary(model_2)
```

Table 2:

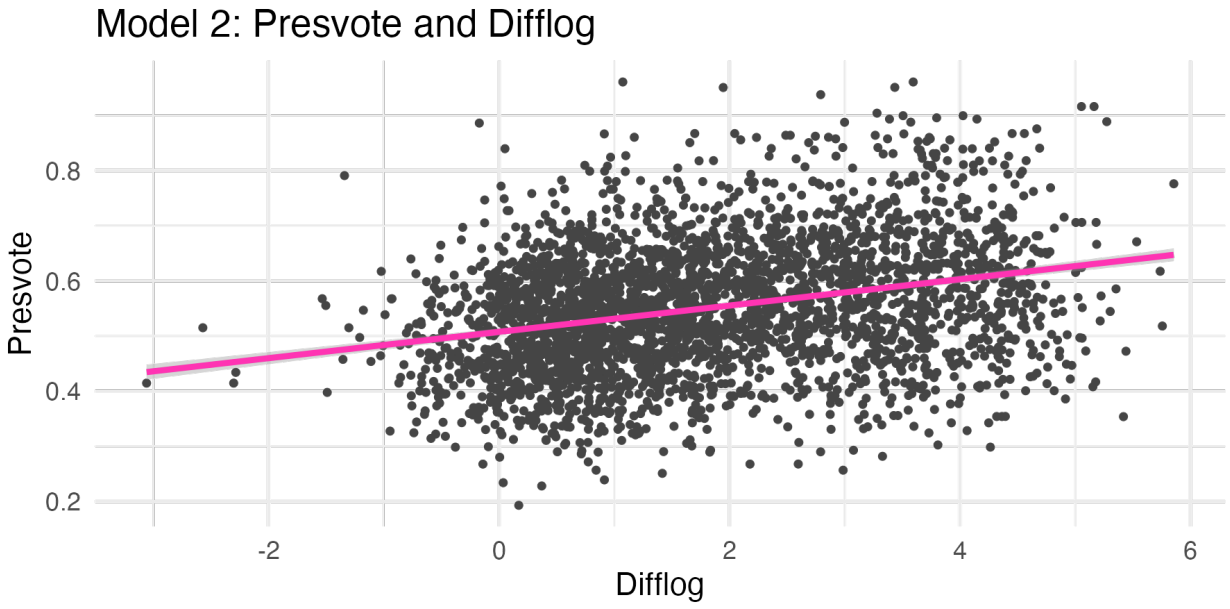
	Dependent variable:
	presvote
difflog	0.024^{***}
	(0.001)
Constant	0.508^{***}
	(0.003)
Observations	3,193
R^2	0.088
Adjusted R^2	0.088
Residual Std. Error	0.110 (df = 3191)
F Statistic	307.715^{***} (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

   This is the code used to plot:

```
scatter_2 <- ggplot(data = inc.sub, mapping = aes(x = difflog,
y = presvote)) +
labs(x = "Difflog", y = "Presvote",
title = "Model 2: Presvote and Difflog") +
theme_minimal() +   geom_point(color = "grey27", size = .8) +
geom_smooth(method = 'lm',col="maroon1")
```

Figure 2: Presvote and Difflog



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

   I am saving the residuals of model 1 in the separate object . I used this code:

   ```
   residuals_2 <- model_2$residual
   ```

4. Write the prediction equation.

   Using the data from the summary (see solution 2.1) the prediction equation is:
   $\hat{y} = \mathbf{0.508 + 0.024X}_1$

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`.

   I run the regression with the inbuilt "lm" function whilst subsetting "voteshare" as Y (outcome variable) and "presvote" as X (explanatory variable). This is the code I used:

```
model_3 <- lm(inc.sub$voteshare~inc.sub$presvote)
summary_3 <- summary(model_3)
```

Table 3:

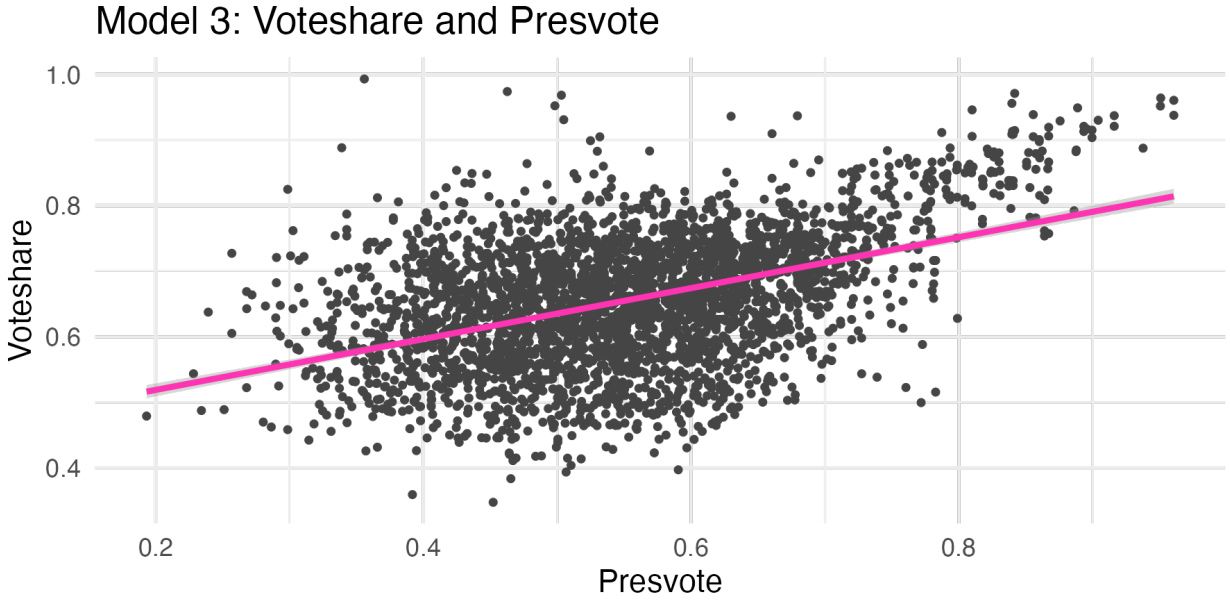
	Dependent variable:
	voteshare
presvote	0.388***
	(0.013)
Constant	0.441***
	(0.008)
Observations	3,193
R^2	0.206
Adjusted R^2	0.206
Residual Std. Error	0.088 (df = 3191)
F Statistic	826.950*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

   This is the code used to plot:

```
scatter_3 <- ggplot(data = inc.sub, mapping = aes(x = presvote,
y = voteshare)) +   labs(x = "Presvote",       y = "Voteshare",
title = "Model 3: Voteshare and Presvote") +
theme_minimal() +   geom_point(color = "grey27", size = .8) +
geom_smooth(method = 'lm',col="maroon1")
```

Figure 3: Voteshare and Presvote



3. Write the prediction equation.

   Using the data from the summary (see solution 3.1) the prediction equation is:
   $\hat{y} = \mathbf{0.441} + \mathbf{0.388X}_1$

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.

   I run the regression with the inbuilt "lm" function with "Residuals 1" as Y (outcome variable) and "Residuals 2" as X (explanatory variable). This is the code I used:

```
model_4 <- lm(residuals_1~residuals_2)
summary_4 <- summary(model_4)
```

Table 4:

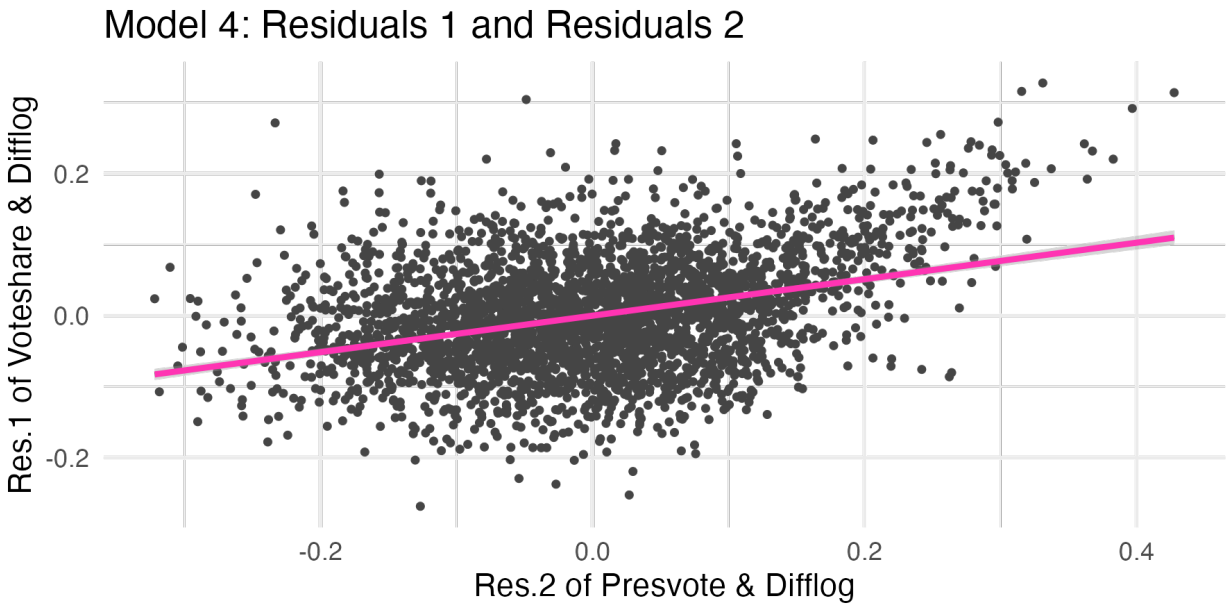
	Dependent variable:
	residuals_1
residuals_2	0.257***
	(0.012)
Constant	−0.000
	(0.001)
Observations	3,193
R^2	0.130
Adjusted R^2	0.130
Residual Std. Error	0.073 (df = 3191)
F Statistic	476.975*** (df = 1; 3191)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

   This is the code used to plot:

```
scatter_4 <- ggplot(data = inc.sub,
 mapping = aes(x = residuals_2, y = residuals_1)) +
 labs(x = "Res.2 of Presvote & Difflog",  y = "Res.1 of Voteshare & Difflog",
 title = "Model 4: Residuals 1 and Residuals 2") +
theme_minimal() +
 geom_point(color = "grey27", size = .8) +
 geom_smooth(method = 'lm',col="maroon1")
```

Figure 4: Residuals 1 and Residuals 2



3. Write the prediction equation.

   Using the intercepts from the summary (see solution 4.1) the prediction equation is:
   $\hat{y}$ = **-0.000 + 0.257X**$_1$
   or the same in scientific numbers: $\hat{y}$ = -1.942e-18 + 2.569e-01X$_1$

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`.

   I run the multivariate regression with the inbuilt "lm" function with "Voteshare" as Y (outcome variable) and "Difflog" and "Presvote as X (explanatory variables). This is the code I used:

```
model_5 <- lm(inc.sub$voteshare~inc.sub$difflog + inc.sub$presvote)
summary_5 <- summary(model_5)
```

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.

   Using the data from the summary (see solution 5.1) the prediction equation is:
   $\hat{y} = \mathbf{0.449} + \mathbf{0.036}X_1 + \mathbf{0.257}X_2$

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 the above summary tables were gathered with the "stargazer" package in R, the residuals have not been transferred to this pdf file. Nevertheless we can observe that the **"Residual Std. Error"** for Q4 and Q5 is the same for both models with **0.073**.

   For a closer look - Here are the Residuals from Q4 and Q5 directly extracted from R Summary for the two models:

```
Q4 Residuals:
Min -0.25928
1Q -0.04737
Median -0.00121
3Q 0.04618
Max 0.33126

Q5 Residuals:
Min -0.25928
1Q -0.04737
Median -0.00121
3Q 0.04618
Max 0.33126
```

Residuals describe the distance between the observed and the value that is predicted by the model for y. Furthermore the model in Q4 takes in the residuals from the same explanatory variables that are used in Q5.
- Residuals 1: Voteshare and Difflog
- Residuals 2: Presvote and Difflog
- Model in Q5: Voteshare and Difflog + Presvote

Furthermore, the models are done on the same datasource so this also attributes to this result. As both models are accounting for residuals on the same variables the residuals are the same. The R^2 and $R^2 adjusted$ in Model No. 5 are almost the same, which supports my theory, that the explanatory variables do not improve the outcome.

Bonus thought: Such a similar R^2 and $R^2 adjusted$ in model no. 5 (multivariate regression) would in - my knowledge - be an indicator for multicolinearity. But as the regression models from question 1 and 2 do not indicate strong correlation I cannot make a solid conclusion of this - but it strikes me odd (overfitting).