Problem Set 3

Applied Stats/Quant Methods 1

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`.

   ### Answer.

   To run a regression where the outcome variable is voteshare and the explanatory variable is difflog the lm function in R can be used.

```
1       model1 <- lm(voteshare ~ difflog, data = inc.sub)
2
```

   Where voteshare is the outcome (Y axis) variable and difflog is the explanatory (X axis) variable. Use the summary function to see the outcome of the linear regression model. The summary of the linear regression model1 is displayed on the below table:

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.

Answer

To make a scatterplot of the linear model the following code can be used.

```
{plot(voteshare ~ difflog, data = inc.sub, col = 'black')+
  abline(model1, col = 'red', lwd = 3)+
  grid()+
  title(main = 'Scatter Plot Voteshare Vs. Difflog', y = "voteshare",
  x = "difflog")
  }

```

Which produces the following scatterplot:

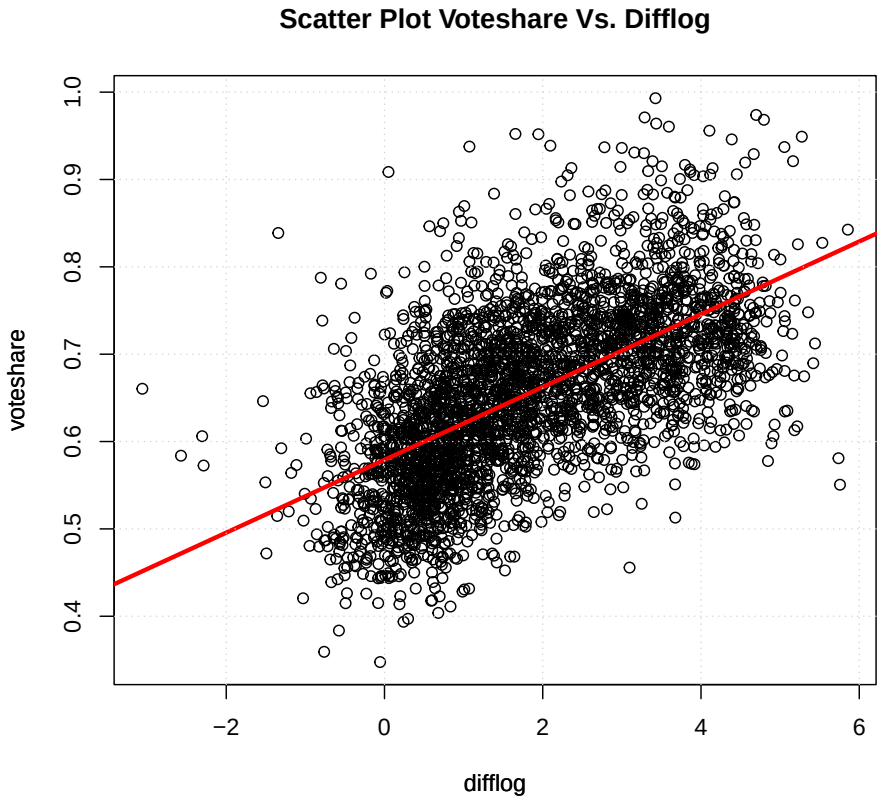

Figure 1: Model 1

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

Answer.

To save the residuals of the model, "model1", in a separate object the following code can be used.

```
model1_residuals <- resid(model1)

```

The model residuals are now saved in a separate object called:

```
"model1_residuals"
```

4. Write the prediction equation.

Answer.

The prediction equation represents the relationship between an outcome variable and one or more explanatory variables. It can be written as such:

$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 ...$

Where:

μ_y = is the predicted mean of the outcome variable.

β_0 = Is the Y intercept (the value of the predicted mean of the outcome variable when the predictor variable is equal to 0).

β_1 = Is the slope of the regression line (the coefficient).

x_1 = Is the value of the predictor variable.

In the case of the linear regression model1, the prediction equation can be written:

μ_y **= (0.57903) + (0.04167)*(**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`.

Answer.

To run a regression where the outcome variable is presvote and the explanatory variable is difflog the lm function in R can be used.

```
model1 <- lm(presvote ~ difflog, data = inc.sub)

```

Where presvote is the outcome (Y axis) variable and difflog is the explanatory (X axis) variable. Use the summary function to see the outcome of the linear regression model. The summary of the linear regression model1 is displayed on the below table:

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.

Answer

To make a scatterplot of the linear model the following code can be used.

```
{plot(presvote ~ difflog, data = inc.sub, col = 'black')+
  abline(model2, col = 'red', lwd = 3)+
  grid()+
  title(main = 'Scatter Plot Presvote Vs. Difflog', y = "presvote",
   x = "difflog")
}

```

Which produces the following scatterplot:

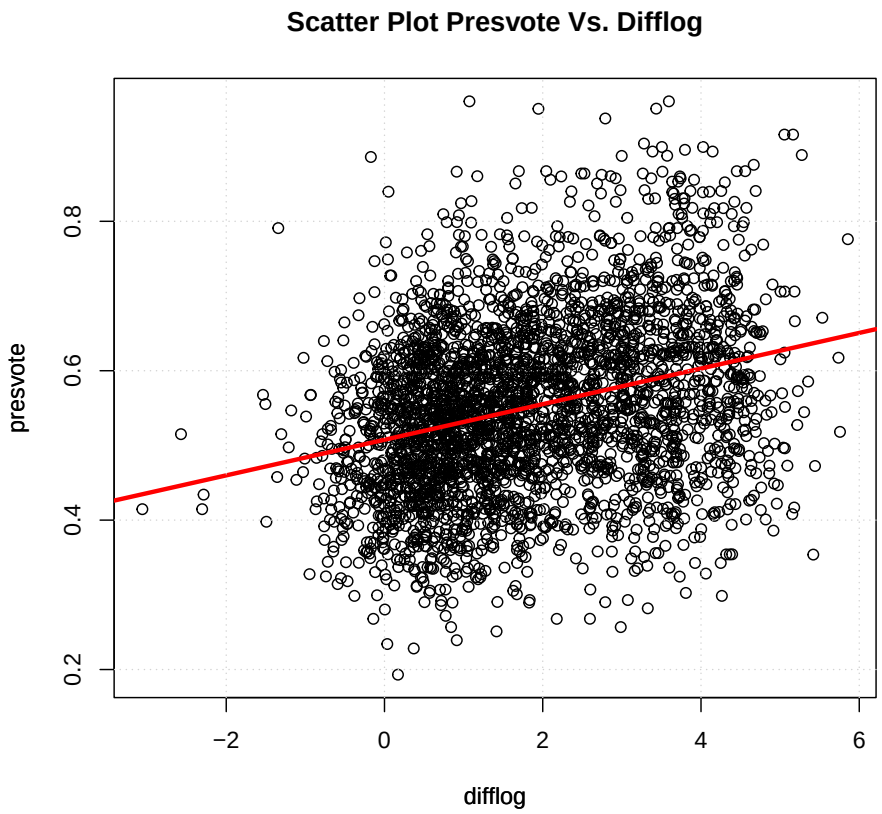

Figure 2: Model 2

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

Answer.

To save the residuals of the model, "model2", in a separate object the following code can be used.

```
model2_residuals <- resid(model2)

```

The model residuals are now saved in a separate object called:

"model2_residuals"

4. Write the prediction equation.

The prediction equation for model2 can be written as:[1]

$\mu_y = \mathbf{(0.50758)} + \mathbf{(0.02384)}*(x_1)$

[1]Refer to Question 1, part 4 for an explanation of the prediction equation.

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`.

Answer.

To run a regression where the outcome variable is voteshare and the explanatory variable is presvote the lm function in R can be used.

```
model3 <- lm(voteshare ~ presvote, data = inc.sub)

```

Where voteshare is the outcome (Y axis) variable and presvote is the explanatory (X axis) variable. The summary of the linear regression model3 is displayed on the below table:

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.

Answer

To make a scatterplot of the linear model the following code can be used.

```
{plot(voteshare ~ presvote, data = inc.sub, col = 'black')+
  abline(model3, col = 'red', lwd = 3)+
  grid()+
  title(main = 'Scatter Plot Voteshare Vs. Presvote', y = "voteshare",
  x = "presvote")
}

```

Which produces the following scatterplot:

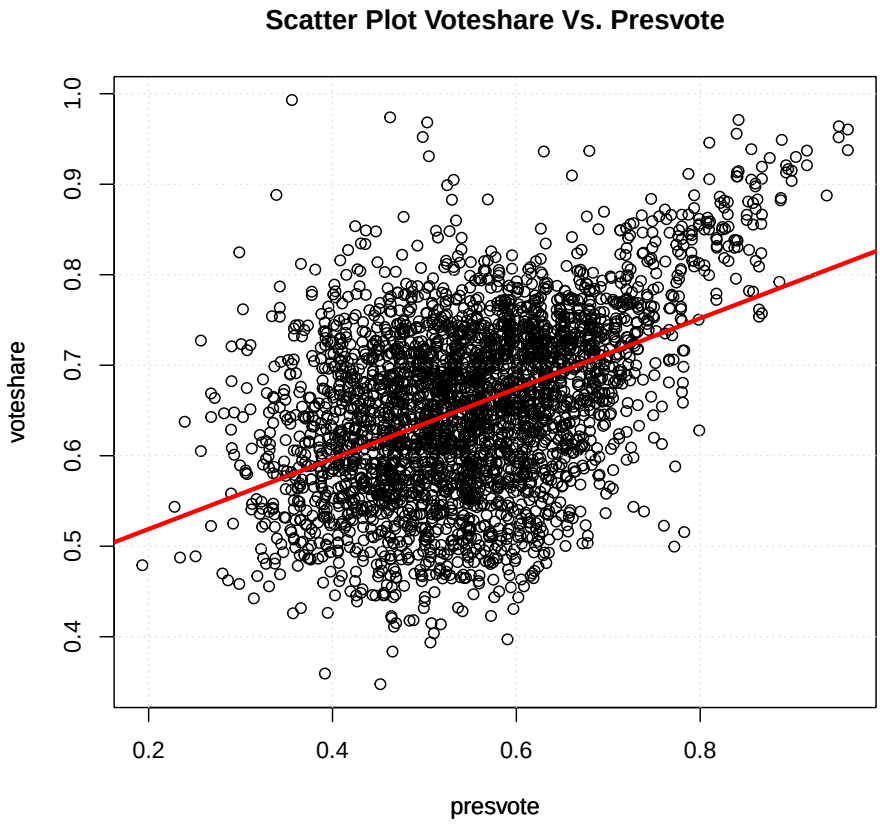

Figure 3: Model 3

3. Write the prediction equation.

Answer.

The prediction equation for model3 can be written as:[2]

μ_y = **(0.4413)** + **(0.3880)*** (x_1)

[2]Refer to Question 1, part 4 for an explanation of the prediction equation.

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.

   ## Answer.

   To run a regression where the outcome variable is the residuals from Question 1 and the explanatory variable is the residuals from Question 2 the residuals must first be saved in a dataframe together, which can be done with the following code.

```
{df1 <- as.data.frame(model1_residuals)
  df2 <- as.data.frame(model2_residuals)
  list_df3 <- c(df1, df2)
  df3 <- as.data.frame(list_df3)
}

```

   Now the residuals from both Questions are saved in a common dataframe called 'df3'. A linear regression can be run on the residuals from Questions 1 and 2 with the lm function.

```
model4 <- lm(model1_residuals ~ model2_residuals, data = df3)

```

   Where model1 residuals is the outcome (Y axis) variable and model2 residuals is the explanatory (X axis) variable. The summary of the linear regression model3 is displayed on the below table:

Table 4:

	Dependent variable:
	model1_residuals
model2_residuals	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.

Answer.

To make a scatterplot of the linear regression with a regression line the following code can be used:

```
{plot(model1_residuals ~ model2_residuals, data = df3, col = 'black')
+
   abline(model4, col = 'red', lwd = 3)+
   grid()+
   title(main = 'Scatter Plot model1_residuals Vs. model2_residuals',
y = "model1_residuals", x = "model2_residuals")
}

```

Which produces this scatterplot:

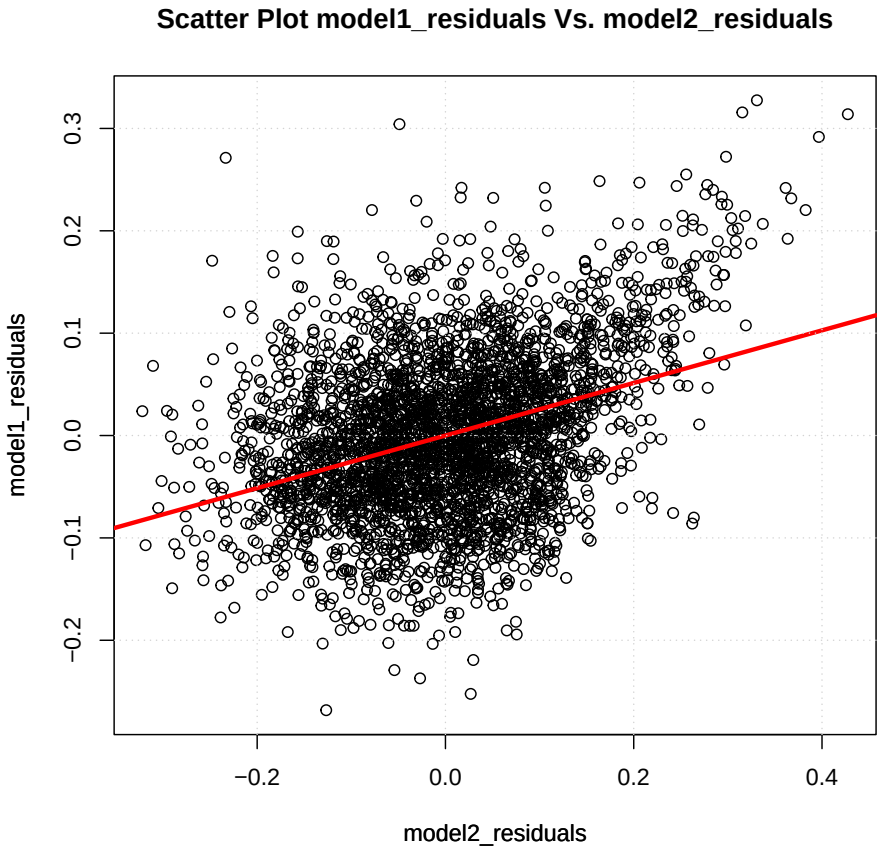

Figure 4: Model 4

3. Write the prediction equation.

Answer.

The prediction equation for model4 can be written as:[3]

μ_y = **(-5.934e-18)** + **(2.569e-01)***(x_1)

[3]Refer to Question 1, part 4 for an explanation of the prediction equation.

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`.

Answer.

To run a regression where the outcome variable is the incumbent's voteshare and the explanatory variables are difflog and presvote the lm function can be used in R.

```
model5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)

```

A summary of the linear regression model5 can be viewed on the table below.

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.

Answer.

The prediction equation for model5 can be written as: [4]

$\mu_y = \mathbf{(0.44864)} + \mathbf{(0.03554)}*(x_1) + \mathbf{(0.25688)}*(x_2)$

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

Answer.

The standard residual error in both model4 and model5 are identical, it is equal to 0.073 in both cases.

A standardized residual in a linear regression model is the difference between a predicted value and an observed value in the model. The standard residual error is the value assigned to the error of the model as a whole, it is used to quantify the variability of the observed data points around a fitted regression line.

When model4 and model5 are printed in R, it seems as though the intercept value assigned to model2-residuals (in model4) and presvote (in model5) is very similar, this can also be viewed in tables 4 and 5, where the value is 0.257 for both model2-residuals (in model4) and presvote (in model5). The correlation of these values can be calculated in R with the cor function.

```
1       cor(model2_residuals, inc.sub$presvote)
2
```

Which returns a value of 0.9550126, suggesting they are very highly correlated.

Model1 is the regression of voteshare vs difflog, where voteshare is the outcome variable. The residuals of this model tell how much of the variation in voteshare is not explained by difflog.

Model2 is the regression of presvote vs difflog, where presvote is the outcome variable. The residuals of this model tell how much of the variation in presvote is not explained by difflog (the difference in spending between the incumbent and the challenger).

Model4 is a regression of the residuals of both these models, model1-residuals vs model2-residuals, where model1-residuals is the outcome variable. The residuals from

[4] Refer to Question 1, part 4 for an explanation of the prediction equation.

this model tell how much of of the variation in model1-residuals is not explained by model2-residuals.

Model5 is a regression of voteshare vs difflog + presvote. Where voteshare is the outcome variable. The residuals from this model tell how much of the variation in voteshare is not explained by both difflog + presvote.

In the regression model1-residuals vs model2-residuals, 0.257 is the slope of the their line of best fit. In the regression voteshare vs presvote + difflog 0.257 is the slope of the line of best fit between presvote and voteshare. Both regressions contain the same variables interacting in similar ways, the variable difflog cancels out, this is why they have a very highly correlated standard residual error.

Bibliography

- Dr. Jeffery Ziegler's lecture slides.
- Hannah Frank's tutorial material.
- Zach (2021) 'How to Interpret Residual Standard Error', Statology, 11 May. Available at: https://www.statology.org/how-to-interpret-residual-standard-error/ (Accessed: 19 November 2023).
- Residual Standard Error (RSE) (2018). Available at: https://www.youtube.com/watch?v=rLwn7OK (Accessed: 19 November 2023).
- Zach (2020) 'What Are Standardized Residuals?', Statology, 22 December. Available at: https://www.statology.org/standardized-residuals/ (Accessed: 19 November 2023).