Problem Set 3

Applied Stats/Quant Methods 1

Due: November 19, 2022///Wei Tang 23362496

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`.

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

Interpretation: I ran the regression and stored it as a variable model1.

```
>summary(model1)
Call:
lm(formula = voteshare ~ difflog, data = inc.sub)

Residuals:
```

```
    Min       1Q   Median       3Q      Max
-0.26832 -0.05345 -0.00377  0.04780  0.32749

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.579031   0.002251  257.19   <2e-16 ***
difflog     0.041666   0.000968   43.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07867 on 3191 degrees of freedom
Multiple R-squared:  0.3673,Adjusted R-squared:  0.3671
F-statistic:  1853 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```r
ggplot(inc.sub, aes(difflog, voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "voteshare ~ difflog", x = "difflog", y = "voteshare")
```

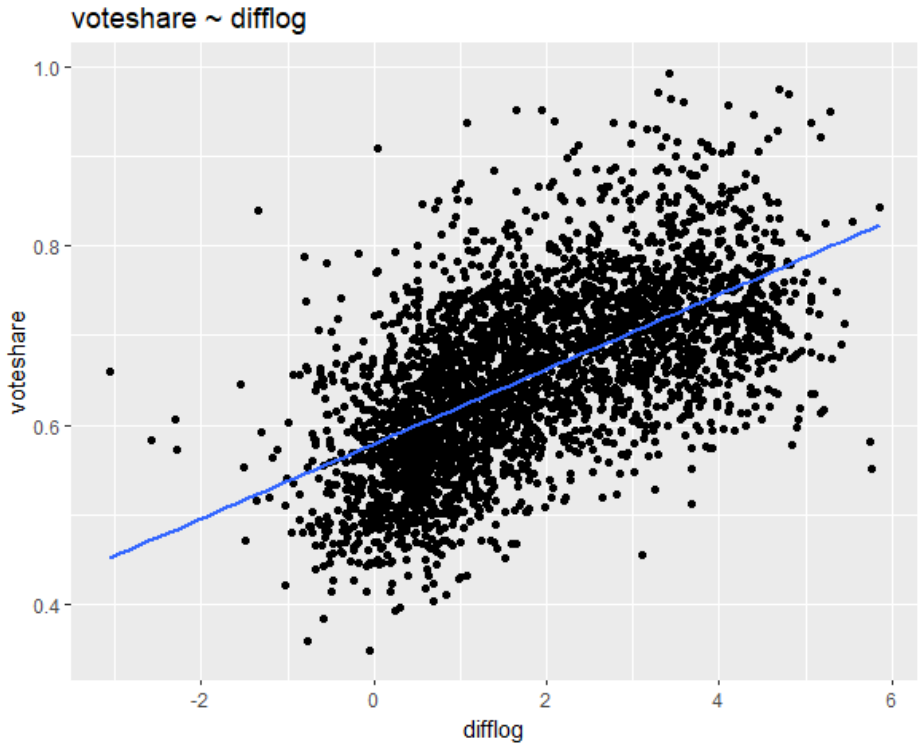

Figure 1: voteshare - difflog

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

```
residuals1 <- model1$residuals
```

4. Write the prediction equation.

```
coefficients1 <- coef(model1)
print(coefficients1)
prediction_equation1 <- paste("voteshare =", round(coefficients1[1], 5),
    "+", round(coefficients1[2], 5), "* difflog")
print(prediction_equation1)
```

The prediction equation is: `voteshare` = 0.57903 + 0.04167 * `difflog`
Interpretation:
For this prediction equation, the slope is 0.04167, the intercept is 0.57903, so I stored the equation as a string variable `prediction_equation1` showing the estimated relationship between these 2 variables. Because both of the p-values of the coefficients are smaller than 0.05, so they are all significant.
Slope is 0.04167 bigger than 0, so there is a positive relationship between `voteshare` and `difflog`.
Intercept is 0.57903, so when `difflog`=0, estimated `voteshare` is 0.57903.

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`.

```
model2 <- lm(presvote ~ difflog, data = inc.sub)
summary(model2)
```

```
> summary(model2)

Call:
lm(formula = presvote ~ difflog, data = inc.sub)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32196 -0.07407 -0.00102  0.07151  0.42743

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.507583   0.003161  160.60   <2e-16 ***
```

```
difflog      0.023837   0.001359   17.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1104 on 3191 degrees of freedom
Multiple R-squared:  0.08795,Adjusted R-squared:  0.08767
F-statistic: 307.7 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```
ggplot(inc.sub, aes(difflog,presvote)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "presvote ~ difflog", x = "difflog", y = "presvote")
```

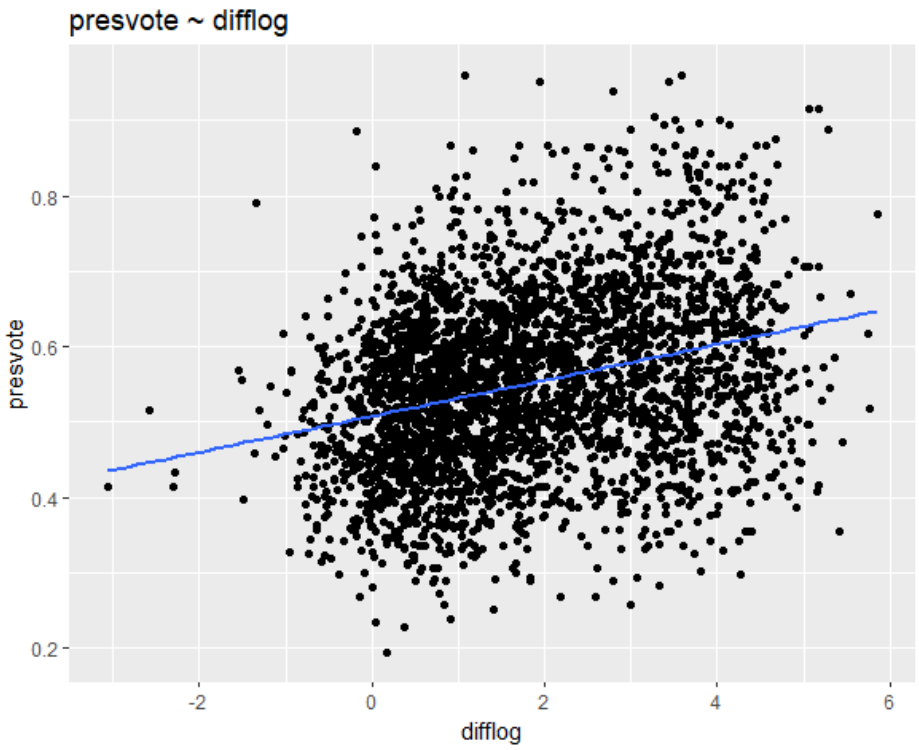

Figure 2: presvote - difflog

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

```
residuals2 <- model2$residuals
print(residuals2)
```

4. Write the prediction equation.

```r
coefficients2 <- coef(model2)
print(coefficients2)
prediction_equation2 <- paste("presvote =", round(coefficients2[1], 5), "
   +", round(coefficients2[2], 5), "* difflog")
print(prediction_equation2)
```

The prediction equation is: `presvote` = 0.50758 + 0.02384 * `difflog`
Interpretation:
For this prediction equation, the slope is 0.02384, the intercept is 0.50758, so I stored the equation as a string variable `prediction_equation2` showing the estimated relationship between these 2 variables. Because both of the p-values of the coefficients are smaller than 0.05, so they are all significant.
Slope is 0.02384 bigger than 0, so there is a positive relationship between `presvote` and `difflog`.
Intercept is 0.50758, so when `difflog`=0, estimated `presvote` is 0.50758.

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`.

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

```
> summary(model3)

Call:
lm(formula = voteshare ~ presvote, data = inc.sub)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27330 -0.05888  0.00394  0.06148  0.41365

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.441330   0.007599   58.08   <2e-16 ***
presvote    0.388018   0.013493   28.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08815 on 3191 degrees of freedom
```

```
Multiple R-squared:  0.2058,Adjusted R-squared:  0.2056
F-statistic:   827 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```r
ggplot(inc.sub, aes(presvote,voteshare)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "voteshare~presvote", x = "presvote", y = "voteshare")
```

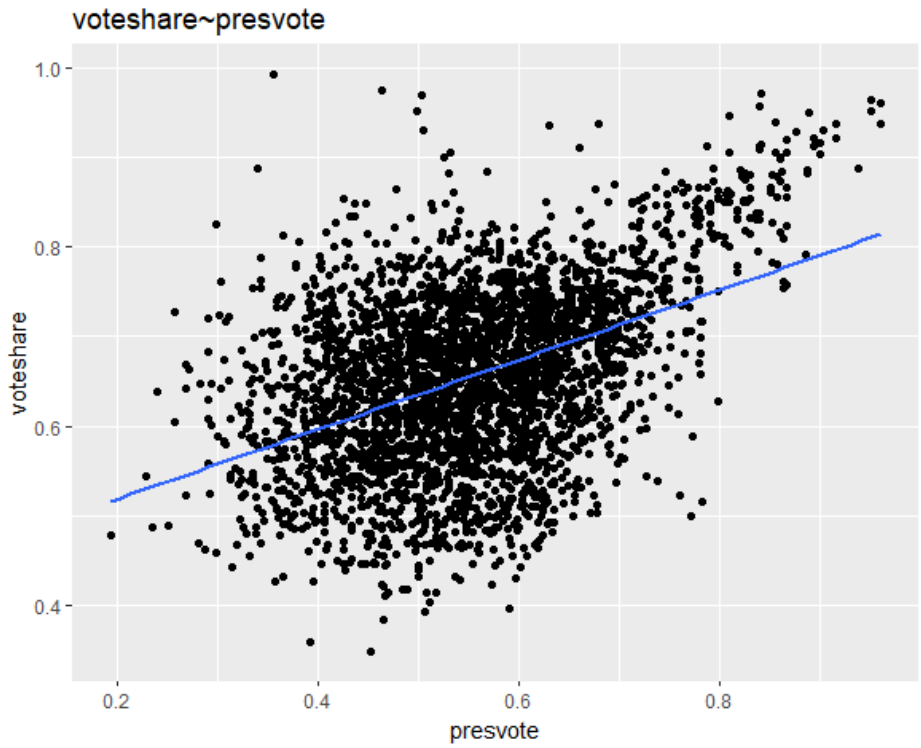

Figure 3: voteshare - presvote

3. Write the prediction equation.

```r
coefficients3 <- coef(model3)
print(coefficients3)
prediction_equation3 <- paste("voteshare =", round(coefficients3[1], 5),
    "+", round(coefficients3[2], 5), "* presvote")
print(prediction_equation3)
```

The prediction equation is: `voteshare` = 0.44133 + 0.38802 * `presvote`
Interpretation:
For this prediction equation, the slope is 0.38802, the intercept is 0.44133, so I stored

the equation as a string variable `prediction_equation3` showing the estimated relationship between these 2 variables. Because both of the p-values of the coefficients are smaller than 0.05, so they are all significant.
Slope is 0.38802 bigger than 0, so there is a positive relationship between `voteshare` and `presvote`.
Intercept is 0.44133, so when `presvote`=0, estimated `voteshare` is 0.44133.

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.

```
model4 <- lm(residuals1~residuals2)
summary(model4)
```

```
> summary(model4)

Call:
lm(formula = residuals1 ~ residuals2)

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

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.934e-18  1.299e-03    0.00        1
residuals2   2.569e-01  1.176e-02   21.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07338 on 3191 degrees of freedom
Multiple R-squared:   0.13,Adjusted R-squared:  0.1298
F-statistic:   477 on 1 and 3191 DF,  p-value: < 2.2e-16
```

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

```r
ggplot(inc.sub, aes(residuals2,residuals1)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "residuals1~residuals2", x = "residuals2", y = "residuals1
    ")
```

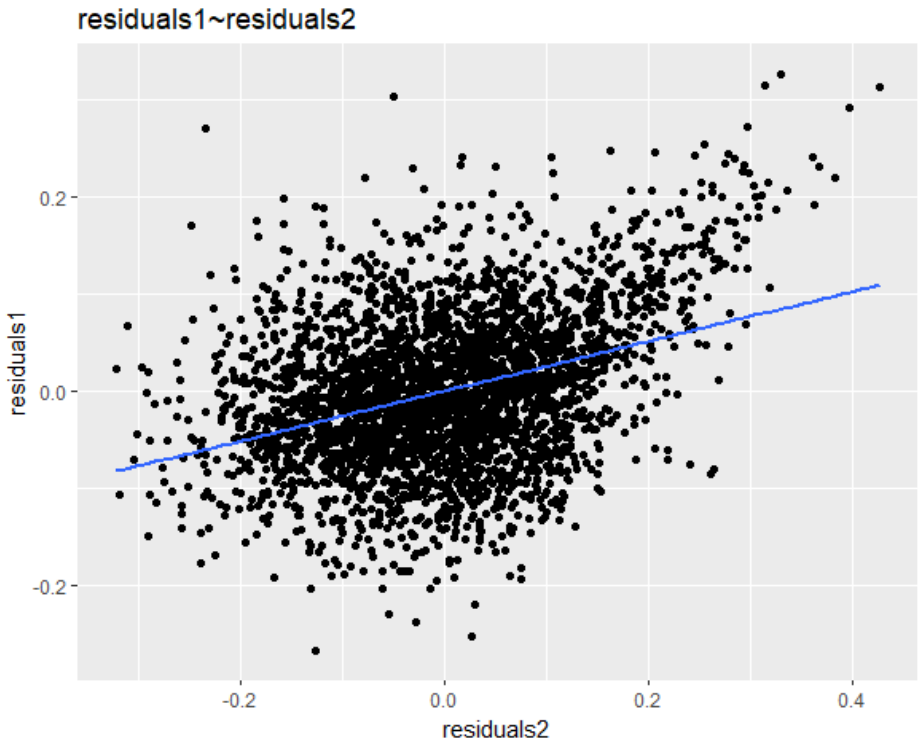

Figure 4: residuals1 - residuals2

3. Write the prediction equation.

```r
coefficients4 <- coef(model4)
print(coefficients4)
prediction_equation4 <- paste("residuals1 =", round(coefficients4[1], 5),
    "+", round(coefficients4[2], 5), "* residuals2")
print(prediction_equation4)
```

The prediction equation is: `residuals1` = 0 + 0.25688 * `residuals2`
Interpretation:
For this prediction equation, the slope is 0.25688, the intercept is 0, so I stored the equation as a string variable `prediction_equation4` showing the estimated relationship between these 2 variables. Because of that the p-values of the slope is smaller than 0.05 and the p-value of the intercept is 1, so the slope is significant but the intercept is not, which means the intercept puts nearly no effect on the model(because every number plus 0 is still the same number).

Slope is 0.25688 bigger than 0, so there is a positive relationship between `residuals1` and `residuals2`.
Intercept is 0, so when `residuals2`=0, estimated `residuals1` is 0.

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`.

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

```
> summary(model5)

Call:
lm(formula = voteshare ~ difflog + presvote, data = inc.sub)

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.4486442  0.0063297   70.88   <2e-16 ***
difflog     0.0355431  0.0009455   37.59   <2e-16 ***
presvote    0.2568770  0.0117637   21.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07339 on 3190 degrees of freedom
Multiple R-squared:  0.4496,Adjusted R-squared:  0.4493
F-statistic:  1303 on 2 and 3190 DF,  p-value: < 2.2e-16
```

2. Write the prediction equation.

```
coefficients5 <- coef(model5)
print(coefficients5)
prediction_equation5 <- paste("voteshare =", round(coefficients5[1], 5),
    "+", round(coefficients5[2], 5), "* difflog", "+", round(coefficients4
    [2], 5), "* presvote")
```

The prediction equation is: voteshare = 0.44864 + 0.03554 * difflog + 0.25688*presvote
Interpretation:
For this prediction equation, the coefficient of difflog is 0.03554, the coefficient of presvote is 0.25688, the intercept is 0.44864, so I stored the equation as a string variable prediction_equation5 showing the estimated relationship between these 2 explanatory variables and the dependent variable. Because the p-values of coefficients are smaller than 0.05, so they are all significant.
Coefficient of difflog is 0.03554 bigger than 0, so there is a positive relationship between voteshare and residuals2.
Coefficient of presvote is 0.25688 bigger than 0, so there is a positive relationship between voteshare and residuals2.
Intercept is 0.44864, so when difflog and presvote are both 0, estimated voteshare is 0.44864.

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

```
print((model4$coefficients))
print((model5$coefficients))
```

```
model4$coefficients:
(Intercept)    residuals2
-5.934078e-18  2.568770e-01

model5$coefficients
(Intercept)     difflog    presvote
 0.44864422  0.03554309  0.25687701
```

I found that the coefficient of residuals2 in model4 is identical to the coefficient of presvote in model5.
Interpretation: I think that because that voteshare is lineal correlated with difflog, and presvote is also correlated with difflog, sopresvote can be considered as (k*difflog+b), so it is clear that the $\frac{d\,\text{residuals1}}{d\,\text{residuals2}}$ is the same as $\frac{\partial \text{voteshare}}{\partial \text{presvote}}$(they have the same ratio), this is the reason why the coefficient of residuals2 in model4 is identical to the coefficient of presvote in model5.