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

Dataset is imported with:

```
df <- read.csv("incumbents_subset.csv")
```

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.

   The linear regression is done by the folowing code:

   ```
   lm1<-lm(formula=voteshare~difflog , data=df)
   ```

   which gives:

   ```
   > lm1
   Call:
   ```

```
lm(formula = voteshare ~ difflog, data = df)

Coefficients:
(Intercept)      difflog
    0.57903      0.04167
```

2. Make a scatterplot of the two variables and add the regression line. The scatter plot is made by the folowing code:

```
pdf(”lm1.pdf”)
plot(df$difflog,df$voteshare,
     xlab=”difflog”, ylab=”voteshare”)
abline(lm1$coefficients, col=2)
dev.off()
```

   Which gives the Fig.1.

3. Save the residuals of the model in a separate object.
   It is done as folows.

```
lm1_residuals<− lm1$residuals
```

   and plot with:

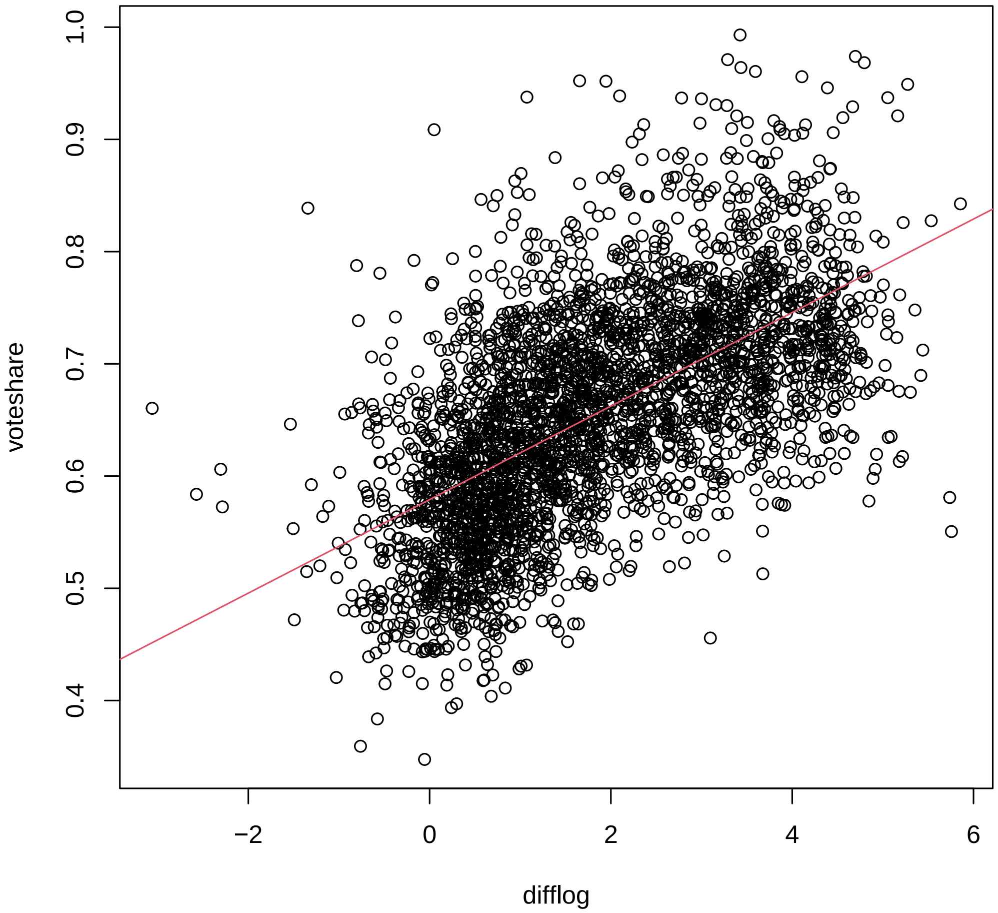
```
pdf(”res1.pdf”)
plot(df$difflog,lm1_residuals,
     xlab=”difflog”, ylab=”residuals model 1”)
abline(a=0,b=0)
dev.off()
```

   Fig. 2 shows residuals are randomly arranged around 0, which means that the linear regression is accurate here.

4. Write the prediction equation.
   Considering $\hat{Y}$ the predictor of `voteshare`,
   $\hat{\beta}_1$ the calculated coefficient,
   $\hat{\beta}_0$ the calculated intercept,
   X the `difflog`
   and the margin of error $\epsilon \rightsquigarrow \mathcal{N}(\mu, \sigma), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,
   the linear regression model gives:

$$\begin{aligned}\hat{Y} &= \hat{\beta}_1 X + \hat{\beta}_0 + \epsilon \\ &= 0.04167X + 0.57903 + \epsilon\end{aligned}$$

Figure 1: Regression between voteshare and difflog



Question 2

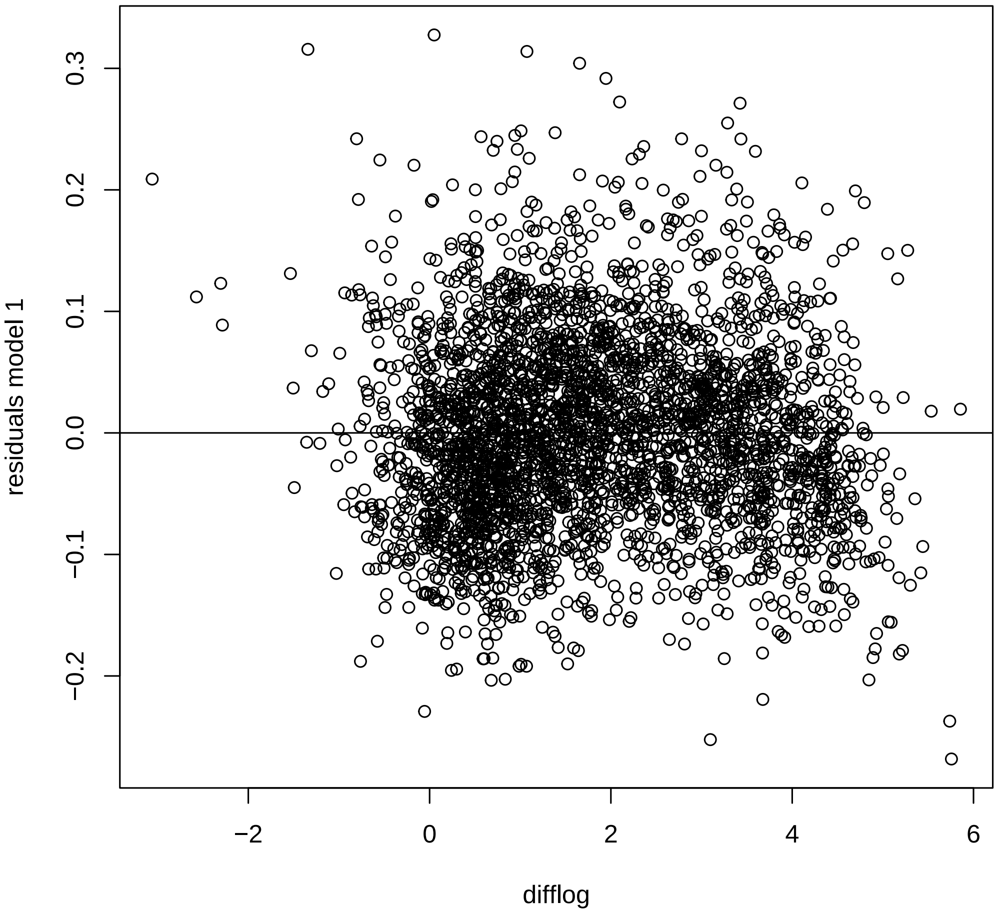
We are interested in knowing how the difference between incumbent and challenger's spending and the vote share of the presidential candidate of the incumbent's party are related.

1. Run a regression where the outcome variable is `presvote` and the explanatory variable is `difflog`.
   The linear regression is done by the folowing code:

```
lm2<-lm(formula=presvote~difflog , data=df)
```

   which gives:

Figure 2: Residuals of the linear regression between voteshare and difflog



```
> lm2
Call:
lm(formula = presvote ~ difflog, data = df)

Coefficients:
(Intercept)      difflog
    0.50758      0.02384
```

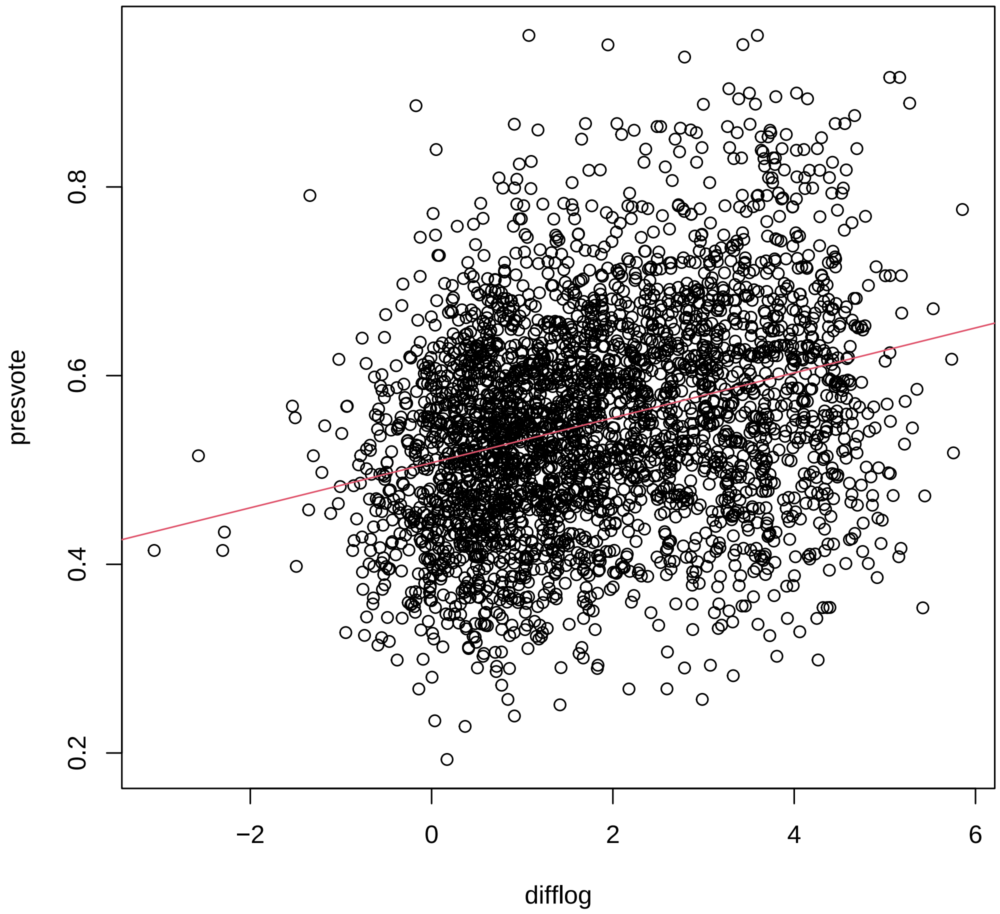
2. Make a scatterplot of the two variables and add the regression line.
   The scatter plot is made by the folowing code:

```
pdf("lm2.pdf")
```

```
plot(df$difflog,df$presvote,
     xlab=”difflog”, ylab=”presvote”)
abline(lm2$coefficients, col=2)
dev.off()
```

Which gives the Fig.3.

Figure 3: Regression between presvote and difflog



3. Save the residuals of the model in a separate object.
   It is done as folows.

```
lm2_residuals<- lm2$residuals
```

and plot with:

```
pdf("res2.pdf")
plot(df$difflog,lm2_residuals,
     xlab="difflog", ylab="residuals model 2")
abline(a=0,b=0)
dev.off()
```

Figure 4: Residuals of the linear regression between presvote and difflog

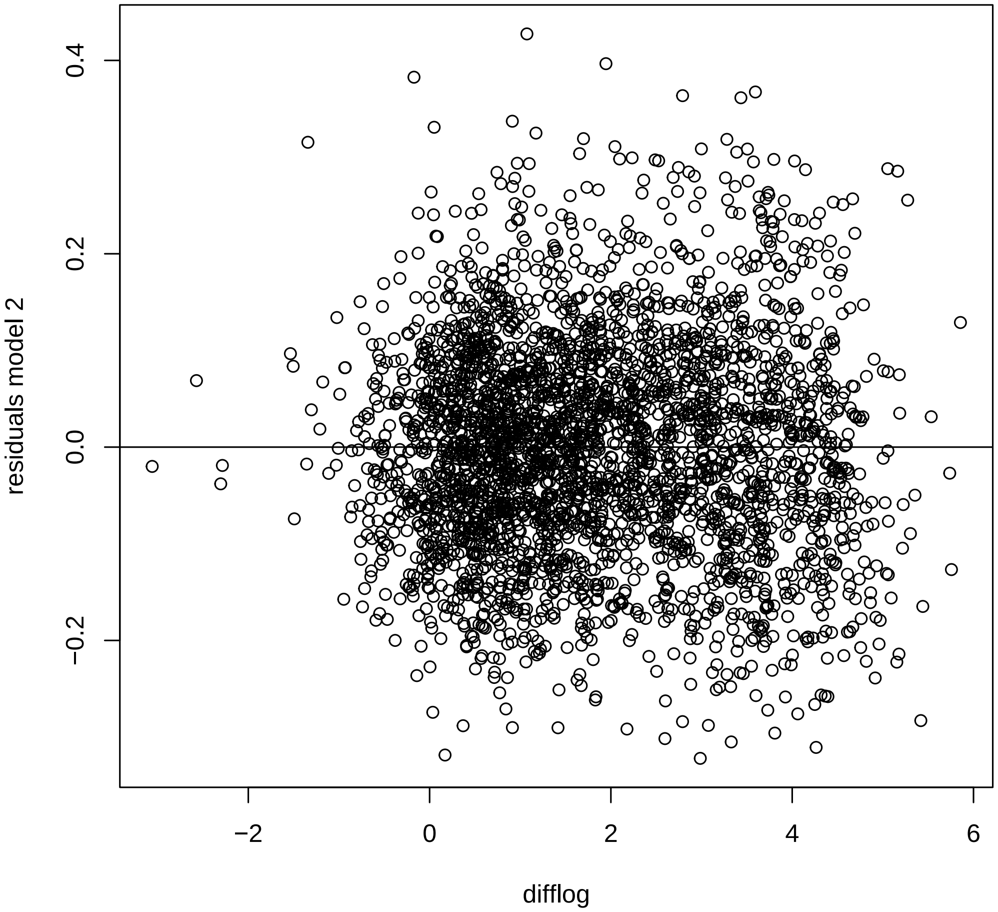


Fig. 4 shows residuals are randomly arranged around 0, which means that the linear regression is accurate here.

4. Write the prediction equation. Considering $\hat{Y}$ the predictor of `presvote`, $\hat{\beta}_1$ the calculated coefficient,

$\hat{\beta}_0$ the calculated intercept,
X the `difflog`
and the margin of error $\epsilon \rightsquigarrow \mathcal{N}(\mu, \sigma), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,
the linear regression model gives:

$$\begin{aligned}\hat{Y} &= \hat{\beta}_1 X + \hat{\beta}_0 + \epsilon \\ &= 0.02384X + 0.50758 + \epsilon\end{aligned}$$

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`.
   The linear regression is done by the folowing code:

```
lm3<-lm(formula=voteshare~presvote, data=df)
```

   which gives:

```
> lm3

Call:
lm(formula = voteshare ~ presvote, data = df)

Coefficients:
(Intercept)     presvote
     0.4413       0.3880
```

2. Make a scatterplot of the two variables and add the regression line.
   The scatter plot is made by the folowing code:

```
pdf(”lm3.pdf”)
plot(df$presvote,df$voteshare,
     xlab=”presvote”, ylab=”voteshare”)
abline(lm3$coefficients, col=2)
dev.off()
```

   Which gives the Fig.5.

   The residuals of the model are given by:

```
lm3_residuals<- lm3$residuals
```

   and plotted with:

```
pdf(”res3.pdf”)
plot(df$difflog,lm3_residuals,
     xlab=”presvote”, ylab=”residuals model 3”)
abline(a=0,b=0)
dev.off()
```

Figure 5: Regression between voteshare and presvote

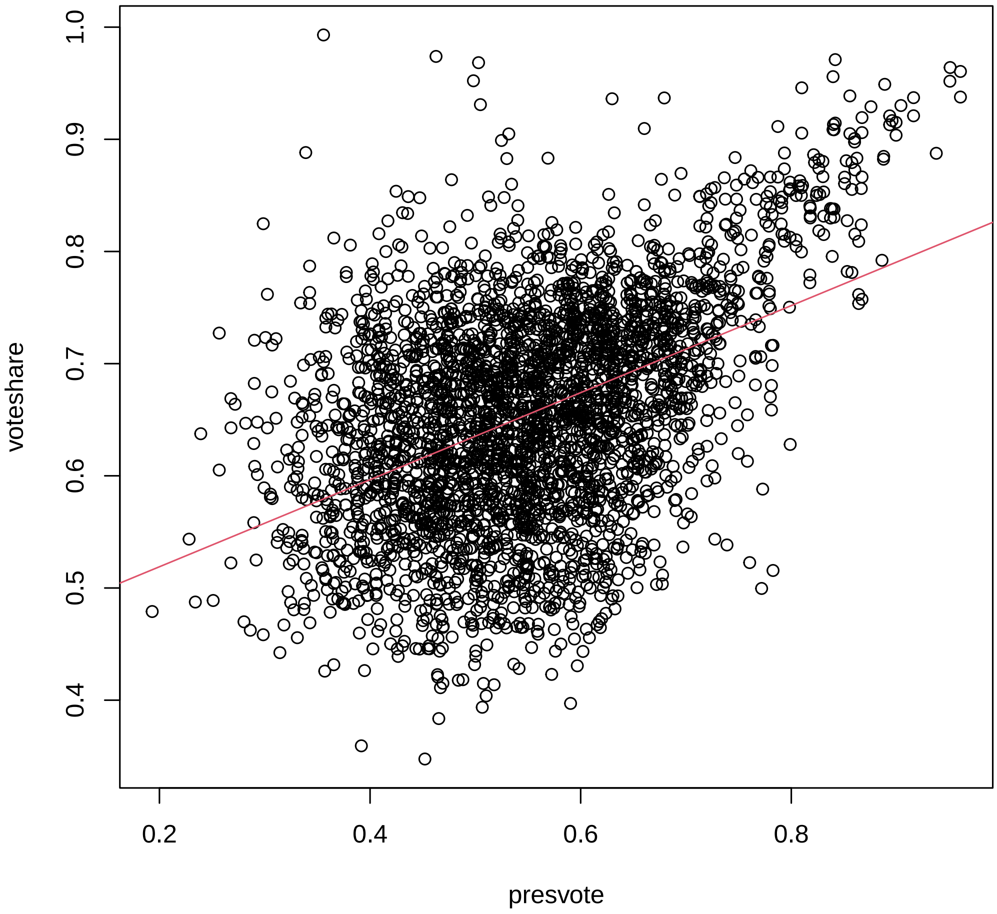
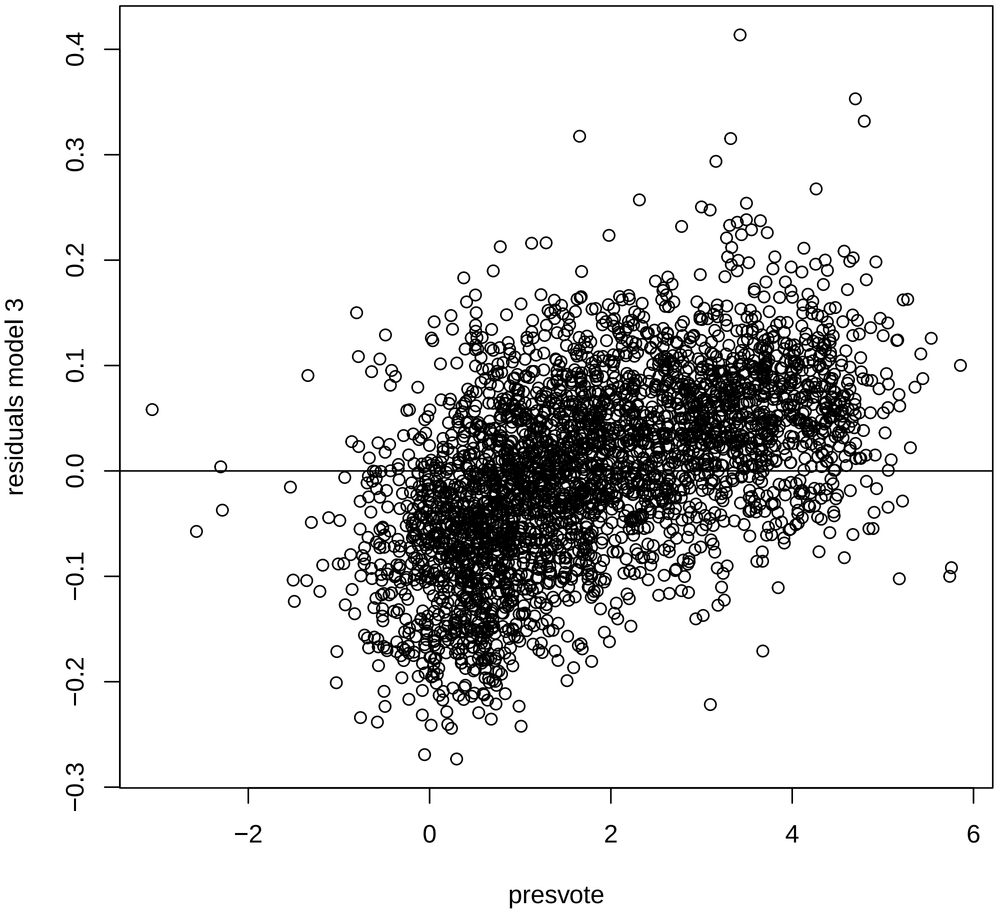


Fig. 6 shows residuals are not randomly arranged around 0, which means that the linear regression is not accurate here. A squared or logarithmic relation are prehaps to expect.

3. Write the prediction equation. Considering $\hat{Y}$ the predictor of `voteshare`,
$\hat{\beta_1}$ the calculated coefficient,
$\hat{\beta_0}$ the calculated intercept,
X the `presvote`
and the margin of error $\epsilon \rightsquigarrow \mathcal{N}(\mu, \sigma), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,
the linear regression model gives:

Figure 6: Residuals of the linear regression between voteshare and presvote



$$\begin{aligned}\hat{Y} &= \hat{\beta}_1 X + \hat{\beta}_0 + \epsilon \\ &= 0.3880X + 0.4413 + \epsilon\end{aligned}$$

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.
   The linear regression is done by the folowing code:

```r
lm4<-lm(formula=lm1_residuals~lm2_residuals, data=df)
```

   which gives:

```
> lm4

Call:
lm(formula = lm1_residuals ~ lm2_residuals, data = df)

Coefficients:
  (Intercept)  lm2_residuals
   -5.934e-18      2.569e-01
```

2. Make a scatterplot of the two residuals and add the regression line.
   The scatter plot is made by the folowing code:

```r
pdf("lm4.pdf")
plot(lm2_residuals,lm1_residuals,
     xlab="residuals model 2", ylab="residuals model 1")
abline(lm4$coefficients, col=2)
dev.off()
```

   Which gives the Fig.7.

   The residuals of the model are given by:

```r
lm4_residuals<- lm4$residuals
```

   and plotted with:

```r
pdf("res4.pdf")
plot(lm2_residuals,lm4_residuals,
     xlab="residuals model 2", ylab="residuals model 1")
abline(a=0,b=0)
dev.off()
```

Figure 7: Regression between residuals of two first models

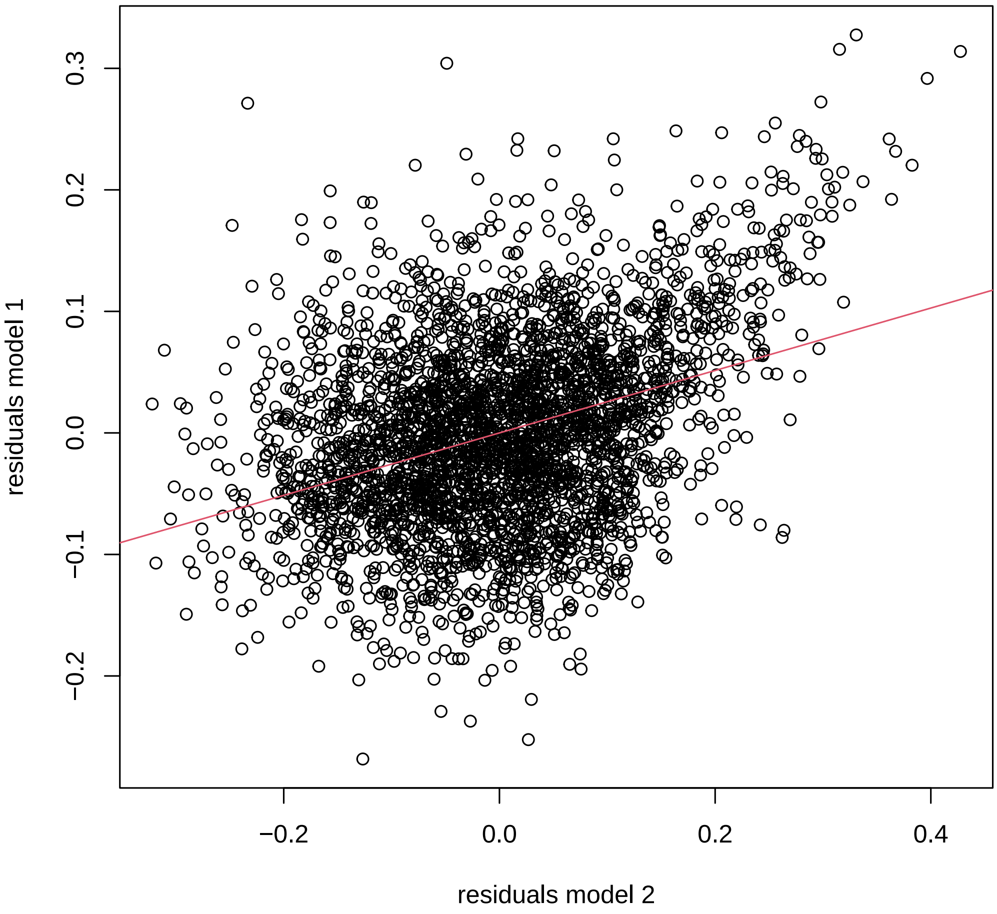
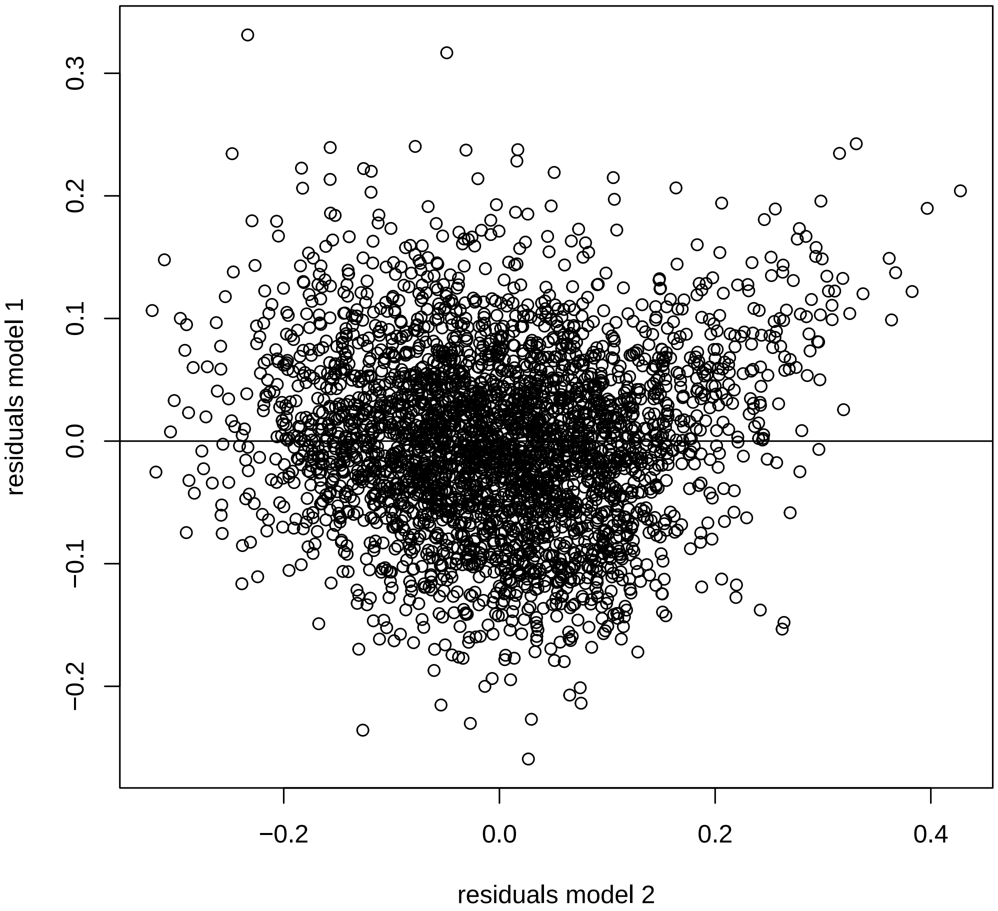


Fig. 8shows residuals are randomly arranged around 0, which means that the linear regression is accurate here.

3. Write the prediction equation.
Considering $\hat{Y}$ the predictor of `lm1_residuals`,
$\hat{\beta}_0$ the calculated coefficient,
$\hat{\beta}_1$ the calculated intercept,
X the `lm2_residuals`
and the margin of error $\epsilon \rightsquigarrow \mathcal{N}(\mu, \sigma), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,
the linear regression model gives:

Figure 8: Residuals of the linear regression between voteshare and presvote



$$\begin{aligned}\hat{Y} &= \hat{\beta}_1 X + \hat{\beta}_0 + \epsilon \\ &= 0.2569X - 5.934 * 10^{-18} + \epsilon\end{aligned}$$

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`.
   The linear regression is done by the folowing code:

   ```
   lm5<-lm(formula=voteshare~difflog +  presvote, data=df)
   ```

   which gives:

   ```
   > lm5

   Call:
   lm(formula = voteshare ~ difflog + presvote, data = df)

   Coefficients:
   (Intercept)      difflog     presvote
       0.44864      0.03554      0.25688
   ```

   The scatter plot is made by the folowing code:

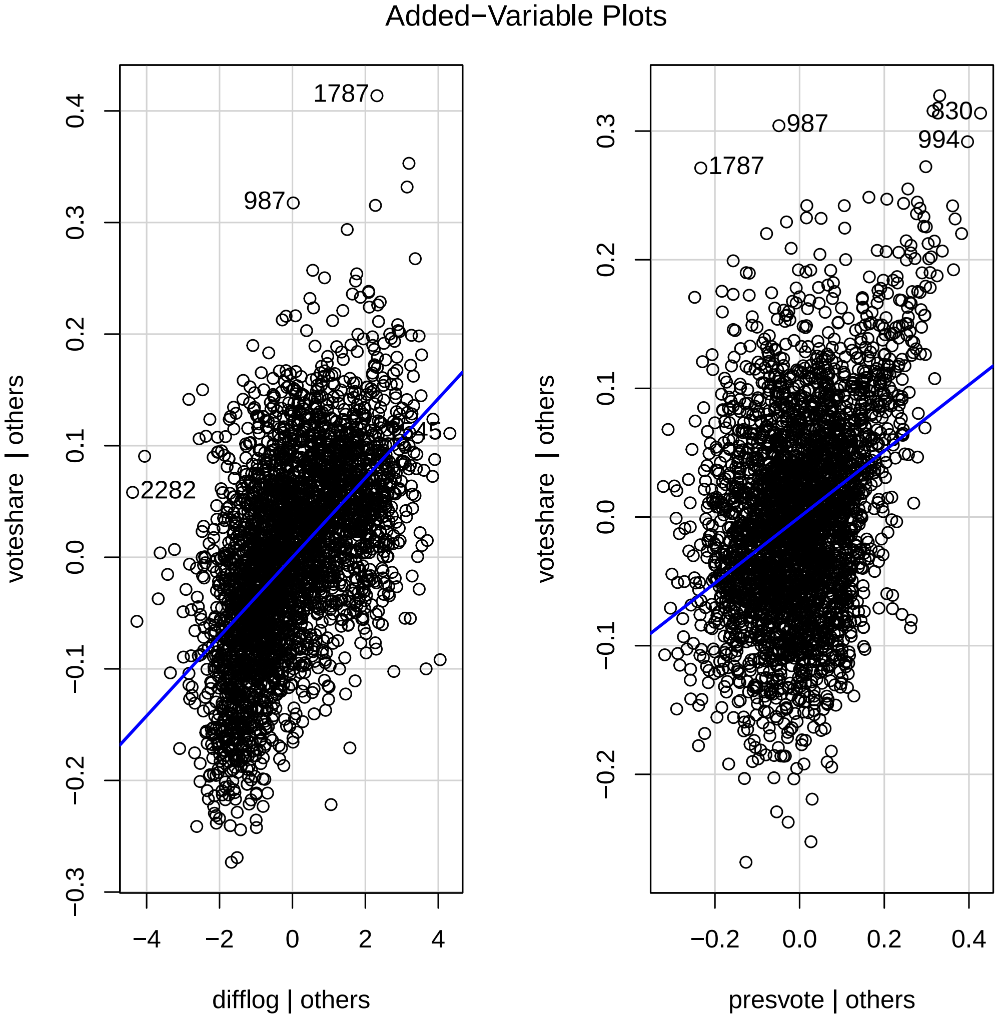
   ```
   library(car)
   library(tidyverse)

   pdf(”lm5.pdf”)
   avPlots(lm5, layout= c(1,2))
   dev.off()
   ```

   Which gives the Fig.9.

2. Write the prediction equation.
   Considering $\hat{Y}$ the predictor of `voteshare`,
   $\hat{\beta}_2$ the slope along `difflog` ,
   $\hat{\beta}_1$ the slope along `presvote` ,
   $\hat{\beta}_0$ the calculated intercept,
   X_2 the `difflog`
   X_1 the `presvote`
   and the margin of error $\epsilon \rightsquigarrow \mathcal{N}(\mu, \sigma), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,
   the linear regression model gives:

$$\begin{aligned}\hat{Y} &= \hat{\beta}_2 X_2 + \hat{\beta}_1 X_1 + \hat{\beta}_0 + \epsilon \\ &= 0.03554 X_2 + 0.25688 X_1 + 0.44864 + \epsilon\end{aligned}$$

Figure 9: Regression between `voteshare` and variables `difflog` and `presvote`



3. What is it in this output that is identical to the output in Question 4? Why do you think this is the case?
   The slope of the model of Question 4 is the same is equal to the slope $\hat{\beta}_1$ of this model. That would mean that the variability of `voteshare` not explained by `difflog` is due almost entirely to `presvote`. Question 4 aimed to guess if there were a relationship between two unmodelised variations (in models of Questions 1 and 2).