# Problem Set 3

Applied Stats/Quant Methods 1
**Maiia Skrypnyk 23371609**

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`.

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

Exploring the model's summary and using 'stargazer' function to format results:

```
summary(m1)
stargazer(m1, title = "Model 1: Incumbent's Vote Share vs Campaign
   Spending Difference", type = "latex")
```

¿

Table 1: Model 1: Campaign Spending Difference vs. Incumbent Vote Share

| | *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 |

On average, a one unit increase in *Difflog* (campaign spending difference between incumbent and challenger) is associated with 0.042 unit increase in *Voteshare* (incumbent's vote share). If $H_0 : \hat{\beta}_1 = 0$, given that our p-value is less than 0.01*, we can reject the null hypothesis that there is no association between *Difflog* and *Voteshare*.

*The estimated coefficient is statistically differentiable from 0 at the $\alpha$ level = 0.05.

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

```r
#Creating a scatterplot + saving as PDF
pdf("01.PS03_Skrypnyk_Plot1.pdf")
plot(inc.sub$difflog,
     inc.sub$voteshare,
     pch = 1,
     col = "coral",
     main = "Model 1: Incumbent's Vote Share vs Campaign Spending Difference",
     xlab = "Campaign spending difference between incumbent and challenger (difflog)",
     ylab = "Incumbent's vote share (voteshare)")
abline(m1, col = "black")
```

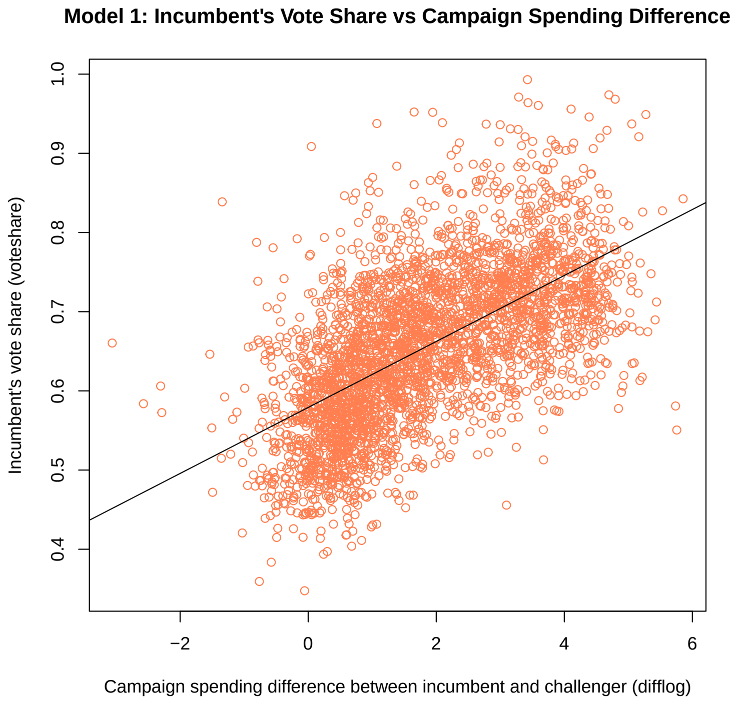

Figure 1: Scatterplot for Model 1

By analysing the scatterplot, we can see that there is a positive association between the variables (although with outliers).

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

```
m1res <- m1$residuals
```

4. Write the prediction equation.

   **Chatterjee and Hadi, 2012, p.41*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

$$\widehat{\boldsymbol{Voteshare}}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathit{Difflog}_i$$

- $\hat{y}_i = \widehat{\boldsymbol{Voteshare}}_i$ = predicted value of the response variable (incumbent's vote share)
- $\hat{\beta}_0$ = estimated intercept
- $\hat{\beta}_1$ = estimated slope coefficient

- $x_i = \mathit{Difflog}_i =$ (any chosen) value of the explanatory variable (difference in campaign spending between incumbent and challenger candidates)

$$\widehat{\boldsymbol{Voteshare}}_i = 0.579 + 0.042 \cdot \mathit{Difflog}_i$$

# 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`.

```r
m2 <- lm(presvote ~ difflog, data = inc.sub)
```

Exploring the model's summary and using 'stargazer' function to format results:

```r
summary(m2)
stargazer(m2, title = "Model 2: Presidential Candidate (Incumbent Party) Vote Share vs Campaign Spending Difference", type = "latex")
```

Table 2: Model 2: Presidential Candidate (Incumbent Party) Vote Share vs Campaign Spending Difference

| | *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 |

On average, a one unit increase in *Difflog* (campaign spending difference between incumbent and challenger) is associated with 0.024 unit increase in *Presvote* (presidential candidate's from the incumbent party vote share). If $H_0 : \hat{\beta}_1 = 0$, given that our p-value is less than 0.01*, we can reject the null hypothesis that there is no association between *Difflog* and *Presvote*.

*The estimated coefficient is statistically differentiable from 0 at the $\alpha$ level = 0.05.

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

```r
#Creating a scatterplot + saving as PDF
pdf("02.PS03_Skrypnyk_Plot2.pdf")
plot(inc.sub$difflog,
     inc.sub$presvote,
     pch = 5,
     col = "green",
     main = "Model 2: Presidential Candidate (Incumbent Party) Vote Share
    vs Campaign Spending Difference",
     cex.main = 0.8,
     xlab = "Campaign spending difference between incumbent and
   challenger (difflog)",
     ylab = "Presidential candidate (incumbent party) vote share (
   presvote)")
abline(m2, col = "black")
```

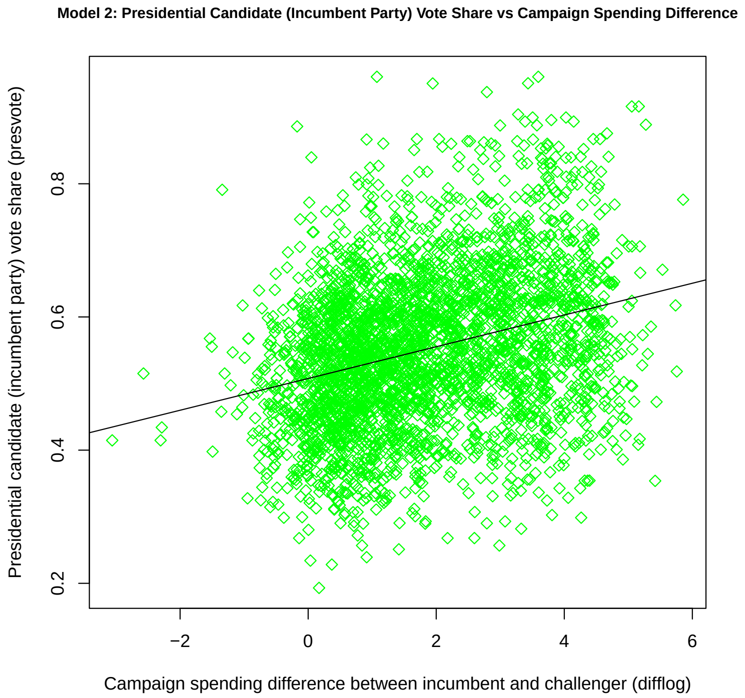

Figure 2: Scatterplot for Model 2

Analysing the scatterplot, we can see (similarly to the previous one) that there is a positive association between the variables (although with outliers).

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

```r
m2res <- m2$residuals
```

4. Write the prediction equation.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

$$\widehat{\boldsymbol{Presvote}}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathit{Difflog}_i$$

- $\hat{y}_i = \widehat{\boldsymbol{Presvote}}_i$ = predicted value of the response variable (vote share of the presidential candidate from the incumbent party)
- $\hat{\beta}_0$ = estimated intercept
- $\hat{\beta}_1$ = estimated slope coefficient
- $x_i = \mathit{Difflog}_i$ = (any chosen) value of the explanatory variable (difference in campaign spending between incumbent and challenger candidates)

$$\widehat{\boldsymbol{Presvote}}_i = 0.508 + 0.024 \cdot \mathit{Difflog}_i$$

## 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
m3 <- lm(voteshare ~ presvote, data = inc.sub)
```

Exploring the model's summary and using 'stargazer' function to format results:

```r
summary(m3)
stargazer(m3, title = "Model 3: Incumbent's Vote Share vs. Presidential
    Candidate (Incumbent Party) Vote Share", type = "latex")
```

Table 3: Model 3: Incumbent's Vote Share vs. Presidential Candidate (Incumbent Party) Vote Share

| | *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 |

On average, a one unit increase in *Presvote* (presidential candidate's from the incumbent party vote share) is associated with 0.388 unit increase in *Voteshare* (incumbent's vote share). It should be noted that this is large value for a slope coefficient, especially compared to the previous two models. If $H_0 : \hat{\beta}_1 = 0$, given that our p-value is less than 0.01*, we can reject the null hypothesis that there is no association between *Presvote* and *Voteshare*.

*The estimated coefficient is statistically differentiable from 0 at the $\alpha$ level = 0.05.

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

```r
#Creating a scatterplot + saving as PDF
pdf("03.PS03_Skrypnyk_Plot3.pdf")
par(mar = c(4, 4, 4, 4))
plot(inc.sub$presvote,
     inc.sub$voteshare,
     pch = 16,
     col = "cyan",
     main = "Model 3: Incumbent's Vote Share vs. Presidential Candidate (Incumbent Party) Vote Share",
     cex.main = 0.7,
     xlab = "Presidential candidate (incumbent party) vote share (presvote)",
     ylab = "Incumbent's vote share (voteshare)")
abline(m3, col = "black")
```

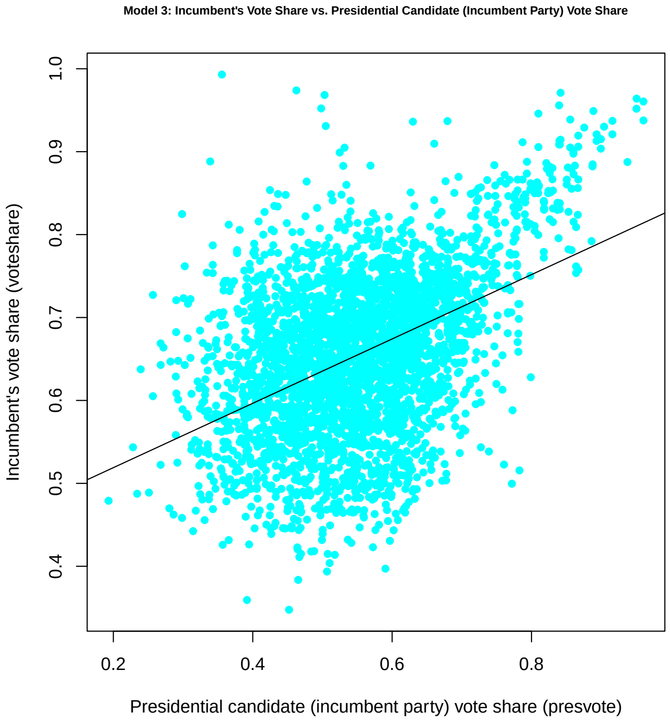

Figure 3: Scatterplot for Model 3

Analysing the scatterplot, we can see (similarly to the previous ones) that there is a positive association between the variables (although with outliers).

3. Write the prediction equation.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

$$\hat{\boldsymbol{Voteshare}}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot Presvote_i$$

- $\hat{y}_i = \hat{\boldsymbol{Voteshare}}_i$ = predicted value of the response variable (estimated incumbent's vote share)
- $\hat{\beta}_0$ = estimated intercept
- $\hat{\beta}_1$ = estimated slope coefficient
- $x_i = Presvote_i$ = (any chosen) value of the explanatory variable (vote share of the presidential candidate from the incumbent party)

$$\hat{\boldsymbol{Voteshare}}_i = 0.441 + 0.388 \cdot Presvote_i$$

# 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.

   ```r
   m4 <- lm(m1res ~ m2res)
   ```

   Exploring the model's summary and using 'stargazer' function to format results:

   ```r
   summary(m4)
   stargazer(m4, title = "Model 4: Model 1 Residuals vs. Model 2 Residuals",
       type = "latex")
   ```

   Table 4: Model 4: Model 1 Residuals vs. Model 2 Residuals

   | | *Dependent variable:* |
   |---|---|
   | | m1res |
   | m2res | 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 |

   On average, a one unit increase in *Model 2 Residuals* is associated with 0.257 unit increase in *Model 1 Residuals*. It should be noted that this is also a large value for a slope coefficient. If $H_0 : \hat{\beta}_1 = 0$, given that our p-value is less than 0.01*), we can reject the null hypothesis that there is no association between *Model 1 Residuals* and *Model 2 Residuals*.

   *The estimated coefficient is statistically differentiable from 0 at the $\alpha$ level = 0.05.

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

```r
#Creating a scatterplot + saving as PDF
pdf("04.PS03_Skrypnyk_Plot4.pdf")
par(mar = c(6, 6, 6, 6))
plot (m2res,
      m1res,
      pch = 8,
      col = "grey",
      main = "Model 4: \n Model 1 Residuals (variation in voteshare not
    explained by the difference in spending between incumbent and
    challenger) vs. \n Model 2 Residuals (variation in presvote not
    explained by the difference in spending between incumbent and
    challenger)",
      cex.main = 0.7,
      xlab = "Model 2 Residuals",
      ylab = "Model 1 Residuals")
abline(m4, col = "black")
```

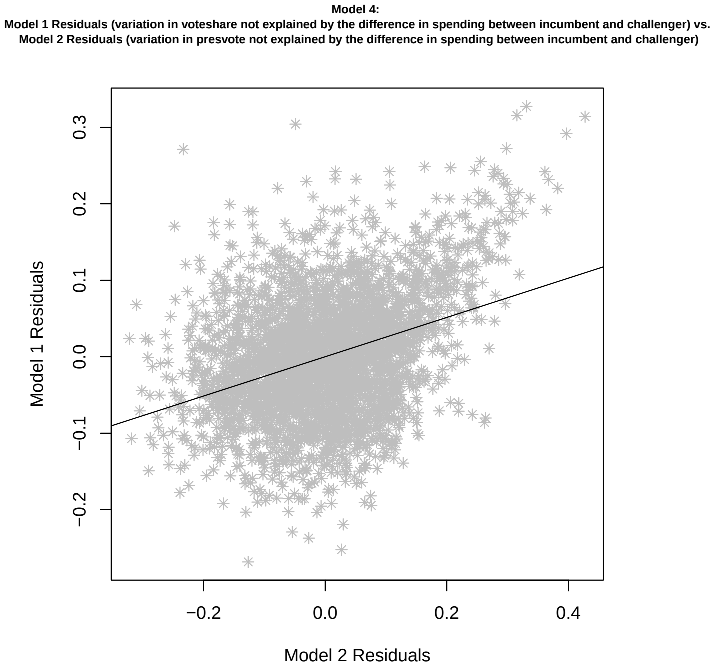

Figure 4: Scatterplot for Model 4

Analysing the scatterplot, we can see (similarly to the previous ones) that there is a positive association between the variables (although with outliers).

3. Write the prediction equation.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

$$\hat{\boldsymbol{m1res}}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot m2res_i$$

- $\hat{y}_i = \hat{\boldsymbol{m1res}}_i$ = predicted value of the response variable – Model 1 Residuals (tell us how much of the variation in `voteshare` is *not* explained by the difference in spending between incumbent and challenger)
- $\hat{\beta}_0$ = estimated intercept
- $\hat{\beta}_1$ = estimated slope coefficient
- $x_i = m2res_i$ = (any chosen) value of the explanatory variable – Model 2 Residuals (tell us how much of the variation in `presvote` is *not* explained by the difference in spending between incumbent and challenger in the district)

$$\hat{\boldsymbol{m1res}}_i = 0.257 \cdot m2res_i$$

## 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`.

```
m5 <- lm(voteshare ~ difflog + presvote, data = inc.sub)
```

Exploring the model's summary and using 'stargazer' function to format results:

```
summary(m5)
stargazer(m5, title = "Model 5: Incumbent's Vote Share vs. Campaign
    Spending Difference + Presidential Candidate (Incumbent Party) Vote
    Share", type = "latex")
```

Table 5: Model 5: Incumbent's Vote Share vs. Campaign Spending Difference + Presidential Candidate (Incumbent Party) Vote Share

| | *Dependent variable:* |
|---|---|
| | voteshare |
| difflog | 0.036$^{***}$ |
| | (0.001) |
| presvote | 0.257$^{***}$ |
| | (0.012) |
| Constant | 0.449$^{***}$ |
| | (0.006) |
| Observations | 3,193 |
| $\mathrm{R}^2$ | 0.450 |
| Adjusted $\mathrm{R}^2$ | 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 |

- Holding *Presvote* constant, a one unit increase in *Difflog* (campaign spending difference between incumbent and challenger) is, on average, associated with a 0.036 unit increase in *Voteshare* (the incumbent's vote share). If $H_0 : \hat{\beta}_1 = 0$, given that our p-value is less than 0.01*, we can reject the null hypothesis that there is no association between *Difflog* and *Voteshare* (controlling for other variables).
- Holding *Difflog* constant, a one unit increase in *Presvote* (vote share of the presidential candidate from the incumbent party) is, on average, associated with a 0.0257 unit increase in *Voteshare* (the incumbent's vote share). If $H_0 : \hat{\beta}_2 = 0$, given that our p-value is less than 0.01*, we can reject the null hypothesis that there is no association between *Presvote* and *Voteshare* (controlling for other variables).

*The estimated coefficient is statistically differentiable from 0 at the $\alpha$ level = 0.05.

2. Write the prediction equation.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{1i} + \hat{\beta}_2 \cdot x_{2i}$$

$$\hat{\boldsymbol{Voteshare}}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \textit{Difflog}_i + \hat{\beta}_2 \cdot \textit{Presvote}_i$$

- $\hat{y}_i = \hat{\boldsymbol{Voteshare}}_i$ = predicted value of the response variable (incumbent's vote share)
- $\hat{\beta}_0$ = estimated intercept
- $\hat{\beta}_1$ = estimated slope coefficient for *Difflog*
- $x_1 = \textit{Difflog}_i$ = (any chosen) value of the first explanatory variable (difference in campaign spending between incumbent and challenger candidates)
- $\hat{\beta}_2$ = estimated slope coefficient for *Presvote*
- $x_1 = \textit{Presvote}_i$ = (any chosen) value of the second explanatory variable (vote share of the presidential candidate from the incumbent party)

$$\hat{\boldsymbol{Voteshare}}_i = 0.449 + 0.036 \cdot \textit{Difflog}_i + 0.257 \cdot \textit{Presvote}_i$$

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

   From the result tables, we can see that a slope coefficient $\hat{\beta}_1$ for *m2res_i* (Table 4) and $\hat{\beta}_1$ for *Presvote* over *Voteshare* (*Difflog* controlled for)(Table 5) are identical = 0.257.

   In Model 4, our outcome variable is *Model 1 Residuals* that tells us how much of the variation in *Voteshare* is **not** explained by *Difflog* and our explanatory variable is *Model 2 Residuals* that tells us how much of the variation in *Presvote* is **not** explained by Difflog.

   Multiple linear regression analysis allows us to estimate the association between an outcome variable and an explanatory variable **holding all other variables constant.** In Model 5, we run a multivariate regression of *Voteshare* over both *Presvote* and *Difflog*. Therefore, $\hat{\beta}_1$ for *Presvote* (*Difflog* controlled for) represents the variance for *Voteshare* explained by *Presvote* and **not** explained by *Difflog* – similarly to Model 4.

   We can also see that the Residual Standard Errors of both models = 0.073, so let's also check whether the residuals of these two models are identical (as they should be):

```
#Storing residuals from Model 4 and Model 5 in separate objects
m4res <- m4$residuals
m5res <- m5$residuals

#Using 'all.equal' function to compare the residuals:
all.equal(m4res, m5res) #TRUE
```