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`.

   Step 1: Read data,
   Step 2: build a voteshare difflog model,
   Step 3: and then inspect data through summary

```r
# Load the ggplot2 package for creating visualizations
library(ggplot2)
# read in data
inc.sub <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2023/main/datasets/incumbents_subset.csv")
```

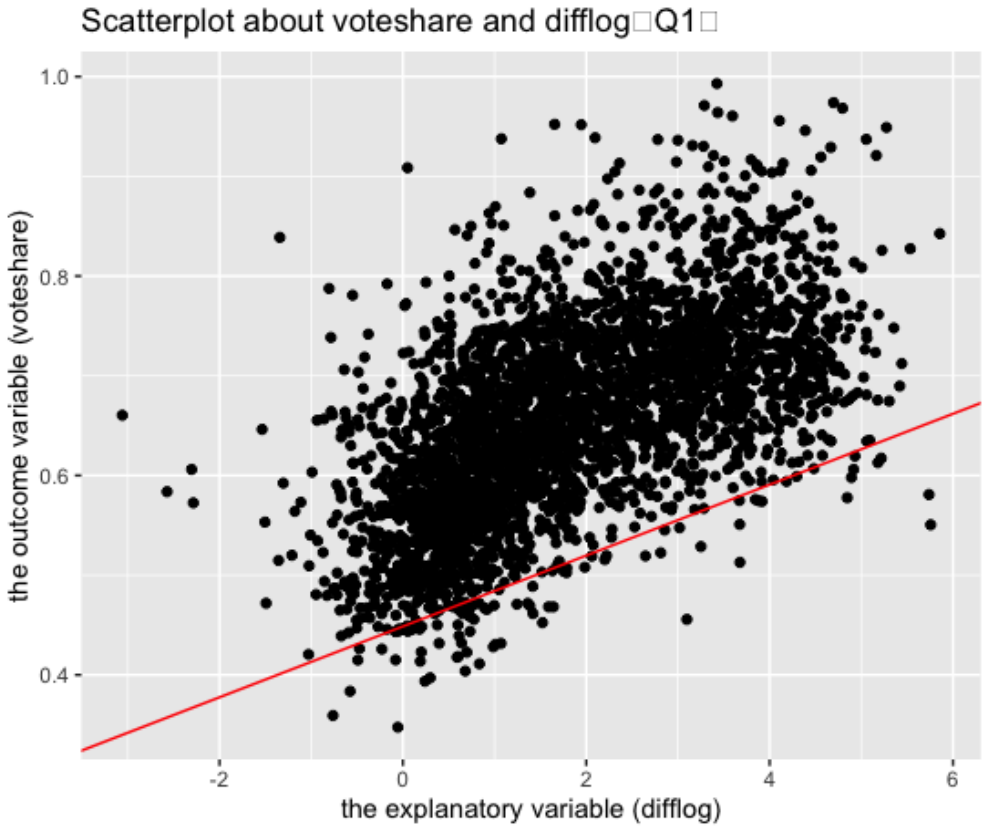
```r
# Build a voteshare ~ difflog model
# Formula: model <- lm(dependent_variable ~ independent_variable, data = dataset)
model <- lm(voteshare ~ difflog, data = inc.sub)
# inspect data through summary
# Summarize and display details of this model: including Regression coefficient, standard error, t statistic, p value, etc
summary(model)
# 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
```

The results of the regression model show that the difflog coefficient is 0.041666, and its p-value is very small (2.2e-16), much smaller than the commonly used significance level (0.05). This means that there is a significant positive correlation between voteshare and difflog. Therefore, we can conclude that increasing the campaign spending differential may have a positive effect on the incumbent president's vote share. This suggests that as the difference in campaign spending increases, the incumbent president's share of the vote correspondingly increases.

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

Make a scatterplot is constructed using the ggplot function,
where the X-axis represents the explanatory variable (difflog) and the Y-axis represents the outcome variable (voteshare).
Specifically:
geom point() adds scatter points where the x coordinate of each point is difflog and the y coordinate is voteshare.
geom abline() adds a regression line whose slope and intercept are derived from the coefficient of the linear regression model (model), respectively.
labs() is used to set the title and axis labels for the chart.

```r
ggplot(inc.sub, aes(x = difflog, y = voteshare)) +
  geom_point() +  # Add scatter
  geom_abline(slope = coef(model)[2], intercept = coef(model)[1], color =
    "red") +  # Add a regression line
  labs(title = "Scatterplot about voteshare and difflog Q1",
       x = "the explanatory variable (difflog)",
       y = "the outcome variable (voteshare)")
```



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

```
# Extract residuals from the current model and name them residuals_Q1
residuals_Q1 <- residuals(model)
head(residuals_Q1)
#           1             2             3             4             5
           6
#  -0.0004227622 -0.0316840149 -0.0045514943  0.0386688767  0.0355287965
    0.0322832521
```

Symbol of the residual: A positive value indicates that the actual vote share is higher than the model predicted value, while a negative value indicates that the actual vote share is lower than the model predicted value.
The models residuals are close to zero, with no clear pattern, suggesting that the model does a good job of explaining the relationship between voteshare and difflog.

4. Write the prediction equation.

```
# Extract coefficients from the current model
coefficients <- coef(model)

# coefficients[1] (intercept) are the intercepts and represent the
    estimated values of the dependent variable voteshare when the
    explanatory variable difflog equals zero. In interpretation, it
    represents the base level of voteshare at zero difflog.
# coefficients[2](the coefficients of difflog) are the coefficients of
    the explanatory variable difflog and indicate the rate of change of
    voteshare with respect to difflog.
# In this context, voteshare is the dependent variable and difflog is the
    explanatory variable
cat("Prediction Equation: voteshare =", coefficients[1], "+",
    coefficients[2], "* difflog\n")
# Prediction Equation: voteshare = 0.5790307 + 0.04166632 * difflog
```

The interpretation of the coefficients is as follows: 0.5790307 is the intercept, indicating the predicted voteshare when the difference in difflog is zero. 0.04166632 is the coefficient for difflog, indicating the expected change in voteshare for a one-unit increase in 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`.

   Step 1: build a presvote ~ difflog model,
   Step 2: and then inspect data through summary

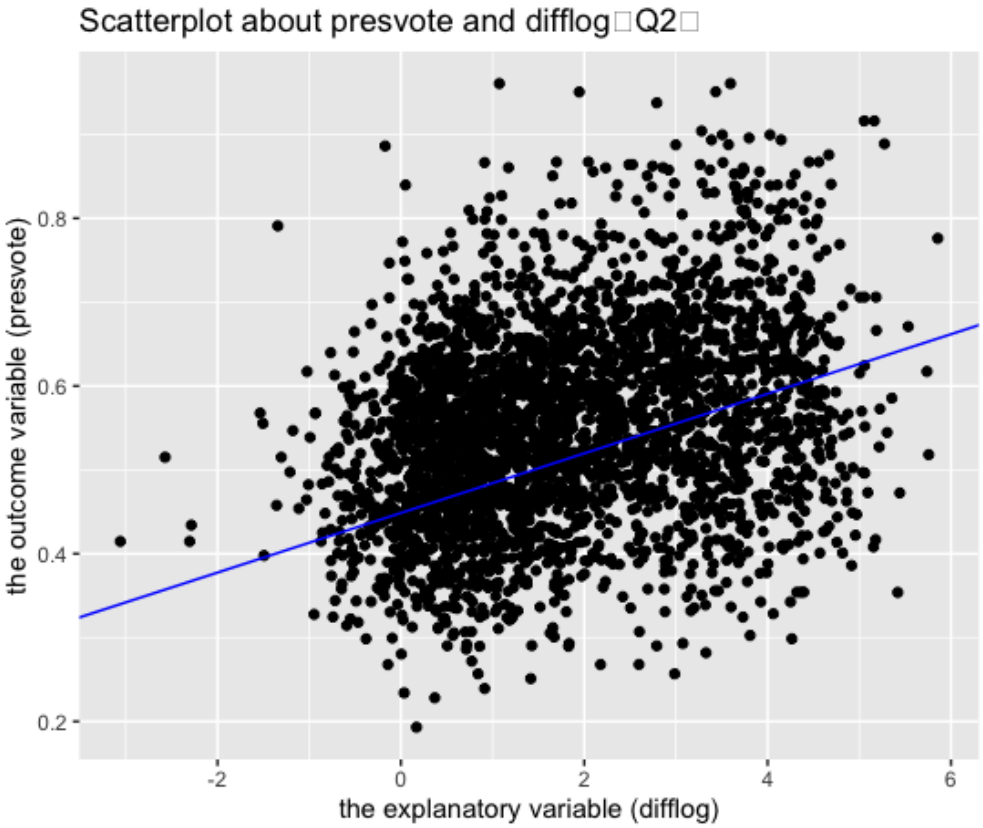
```r
# Build a presvote ~ difflog model
# Formula: model <- lm(dependent_variable ~ independent_variable, data = dataset)
model <- lm(presvote ~ difflog, data = inc.sub)
# inspect data through summary
# Summarize and display details of this model: including Regression coefficient, standard error, t statistic, p value, etc
summary(model)
# 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
```

   The results of the regression model show that the difflog coefficient is 0.023837,
   and its p-value is very small (2.2e-16),
   much smaller than the commonly used significance level (0.05).
   This means that there is a significant positive correlation between presvote and difflog.
   Therefore, we can conclude that an increase in the difference between incumbent and challenger's spending is associated with a corresponding increase in the vote share of the presidential candidate of the incumbent's party.

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

Make a scatterplot is constructed using the ggplot function,
where the X-axis represents the explanatory variable (difflog) and the Y-axis represents the outcome variable (presvote).
Specifically:
geom point() adds scatter points where the x coordinate of each point is difflog and the y coordinate is presvote
geom abline() adds a regression line whose slope and intercept are derived from the coefficient of the linear regression model, respectively.
labs() is used to set the title and axis labels for the chart.

```
ggplot(inc.sub, aes(x = difflog, y = presvote)) +
  geom_point() +
  geom_abline(slope = coef(model)[2], intercept = coef(model)[1], color =
    "blue") +
  labs(title = "Scatterplot about presvote and difflog Q2",
       x = "the explanatory variable (difflog)",
       y = "the outcome variable (presvote)")
```



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

   Extract residuals from the current model and name them residuals Q2

```
residuals_Q2 <- residuals(model)
head(residuals_Q2)
#           1           2           3           4           5           6
# 0.005605594  0.037578519 -0.053134788 -0.052993694 -0.045842994
   0.074339701
```

   The models residuals are close to zero, with no clear pattern,
   suggesting that the model does a good job of explaining the relationship between presvote and difflog.

4. Write the prediction equation.

```
# Extract coefficients from the current model
coefficients <- coef(model)
# coefficients[1] (intercept) are the intercepts and represent the
   estimated values of the dependent variable presvote when the
   explanatory variable difflog equals zero. In interpretation, it
   represents the base level of presvote at zero difflog.
# coefficients[2](the coefficients of difflog) are the coefficients of
   the explanatory variable difflog and indicate the rate of change of
   presvote with respect to difflog.
# In this context, voteshare is the dependent variable and difflog is the
    explanatory variable

cat("Prediction Equation: presvote =", coefficients[1], "+", coefficients
   [2], "* difflog\n")
# Prediction Equation: presvote = 0.5075833 + 0.02383723 * difflog
```

   The interpretation of the coefficients is as follows:
   0.507583 is the intercept, indicating the predicted presvote when the difference in difflog is zero.
   0.023837 is the coefficient for difflog, indicating the expected change in voteshare for a one-unit increase in difflog.

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`.

   Step 1: build a voteshare presvote model,
   Step 2: and then inspect data through summary

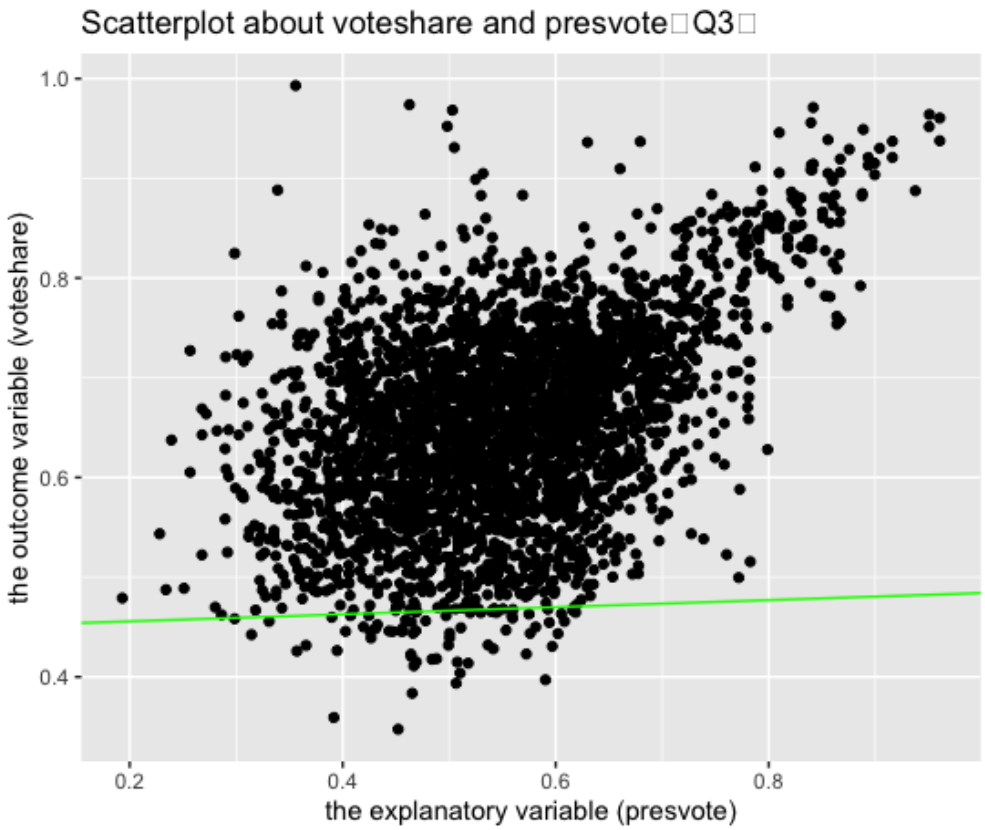
```r
# Build a voteshare ~ presvote model
# Formula: model <- lm(dependent_variable ~ independent_variable, data = dataset)
model <- lm(voteshare ~ presvote, data = inc.sub)
# inspect data through summary
# Summarize and display details of this model: including Regression coefficient, standard error, t statistic, p value, etc
summary(model)
# 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
```

   The results of the regression model show that the presvote coefficient is 0.388018, and its p-value is very small (2.2e-16),
   much smaller than the commonly used significance level (0.05).
   This means that there is a significant positive correlation between voteshare and presvote.
   Therefore, we can conclude that there is a positive association between the vote share of the presidential candidate (voteshare) and the electoral success of the incumbent (presvote).
   An increase in the incumbent"s electoral success is likely to have a positive impact on the vote share of the presidential candidate.

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

Make a scatterplot is constructed using the ggplot function,
where the X-axis represents the explanatory variable (presvote) and the Y-axis represents the outcome variable (voteshare).
Specifically:
geom point() adds scatter points where the x coordinate of each point is presvote and the y coordinate is voteshare.
geom abline() adds a regression line whose slope and intercept are derived from the coefficient of the linear regression model, respectively.
labs() is used to set the title and axis labels for the chart.

```
ggplot(inc.sub, aes(x = presvote, y = voteshare)) +
  geom_point() +
  geom_abline(slope = coef(model)[2], intercept = coef(model)[1], color =
    "green") +
  labs(title = "Scatterplot about voteshare and presvote Q3 ",
       x = "the explanatory variable (presvote)",
       y = "the outcome variable (voteshare)")
```



3. Write the prediction equation.

```r
# Extract coefficients from the current model
coefficients <- coef(model)
# coefficients[1] (intercept) are the intercepts and represent the estimated values of the dependent variable voteshare when the explanatory variable presvote equals zero. In interpretation, it represents the base level of voteshare at zero presvote.
# coefficients[2](the coefficients of presvote) are the coefficients of the explanatory variable presvote and indicate the rate of change of voteshare with respect to presvote.
# In this context, voteshare is the dependent variable and presvote is the explanatory variable

cat("Prediction Equation: voteshare =", coefficients[1], "+", coefficients[2], "* presvote\n")
# Prediction Equation: voteshare = 0.4413299 + 0.3880184 * presvote
```

The interpretation of the coefficients is as follows:
0.4413299 is the intercept, indicating the predicted voteshare when the difference in presvote is zero.
0.3880184 is the coefficient for presvote, indicating the expected change in voteshare for a one-unit increase in presvote.

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.

   Step 1: build a residuals Q1 ~ residuals Q2 model,
   Step 2: and then inspect data through summary

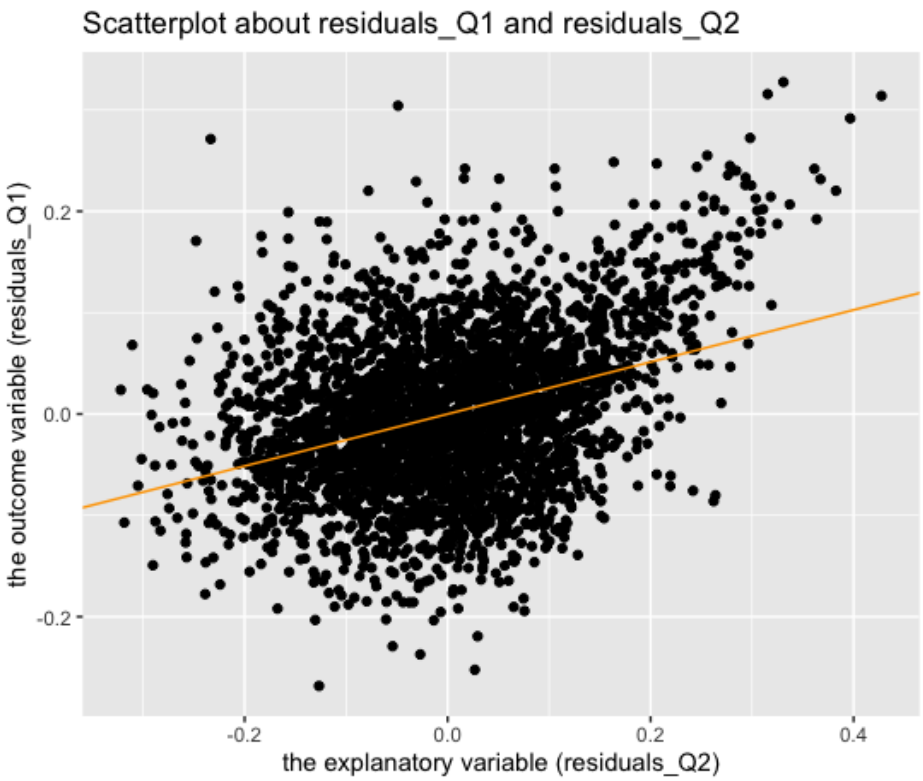
```
# Build a residuals_Q1 ~ residuals_Q2 model
# Formula: model <- lm(dependent_variable ~ independent_variable, data =
    dataset)
model_res <- lm(residuals_Q1 ~ residuals_Q2, data = inc.sub)
summary(model_res)
# Call:
#   lm(formula = residuals_Q1 ~ residuals_Q2, data = inc.sub)

# 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) -5.934e-18  1.299e-03    0.00        1
# residuals_Q2  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
```

   The results of the regression model show that the residuals Q2 coefficient is 2.569e-01, and its p-value is very small (2.2e-16),
   much smaller than the commonly used significance level (0.05).
   This means that there is a significant positive correlation between residuals Q1 and residuals Q2.
   Therefore, we can conclude that the increase in unexplained variation in voteshare is positively correlated with the increase in unexplained variation in presvote, indicating a significant relationship.
   This suggests that factors contributing to the unexplained variation in one area may also contribute to the unexplained variation in the other.

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

Make a scatterplot is constructed using the ggplot function,
where the X-axis represents the explanatory variable (residuals Q2) and the Y-axis represents the outcome variable (residuals Q1).
Specifically:
geom point() adds scatter points where the x coordinate of each point is residuals Q2 and the y coordinate is residuals Q1.
geom abline() adds a regression line whose slope and intercept are derived from the coefficient of the linear regression model, respectively.
labs() is used to set the title and axis labels for the chart. Build a data set for residuals Q1 and residuals Q2

```
residuals_df <- data.frame(Residuals_Q2 = residuals_Q2, Residuals_Q1 =
    residuals_Q1)
# Use ggplot to create a scatter plot and add regression lines
ggplot(residuals_df, aes(x = Residuals_Q2, y = Residuals_Q1)) +
  geom_point() +
  geom_abline(slope = coef(model_res)[2], intercept = coef(model_res)[1],
     color = "orange") +
  labs(title = "Scatterplot about residuals_Q1 and residuals_Q2",
       x = "the explanatory variable (residuals_Q2)",
       y = "the outcome variable (residuals_Q1)")
```



3. Write the prediction equation.

```r
# Extract coefficients from the current model
coefficients_res <- coef(model_res)
# coefficients[1] (intercept) are the intercepts and represent the estimated values of the dependent variable residuals_Q1 when the explanatory variable residuals_Q2 equals zero. In interpretation, it represents the base level of residuals_Q1 at zero residuals_Q2
# coefficients[2](the coefficients of difflog) are the coefficients of the explanatory variable difflog and indicate the rate of change of voteshare with respect to difflog.
# In this context, residuals_Q1 is the dependent variable and residuals_Q2 is the explanatory variable

cat("Prediction Equation: residuals_Q1 =", coefficients_res[1], "+", coefficients_res[2], "* residuals_Q2\n")
# Prediction Equation: residuals_Q1 = -5.934078e-18 + 0.256877 * residuals_Q2
```

The interpretation of the coefficients is as follows:
-5.934078e-18 is the intercept, indicating the predicted residuals Q1 when the difference in residuals Q2 is zero.
0.256877 is the coefficient for residuals Q2, indicating the expected change in residuals Q1 for a one-unit increase in residuals Q2.

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`.

   Step 1: build a voteshare difflog + presvote model,
   Step 2: and then inspect data through summary

```
# Build a voteshare ~ difflog + presvote model
# Formula: model <- lm(dependent_variable ~ independent_variable, data =
    dataset)
model <- lm(voteshare ~ difflog + presvote, data = inc.sub)
summary(model)
# 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
```

   The results of the regression model show that the (difflog + presvote) coefficient is 0.041666,
   and its p-value is very small (2.2e-16),
   much smaller than the commonly used significance level (0.05).
   This means that there is a significant positive correlation between voteshare and (difflog + presvote).

2. Write the prediction equation.

```r
# Extract coefficients from the current model
coefficients_model <- coef(model)

# coefficients[1] (intercept) This is the predicted value of voteshare
    when both (difflog + presvote) are zero. In this context, it
    represents the baseline level of voteshare when all other explanatory
    variables are held constant.
# coefficients[2](the coefficients of difflog) indicating how sensitive
    voteshare is to changes in difflog. Specifically, it represents the
    expected change in voteshare for a one-unit increase in difflog,
    holding all other variables constant.
# coefficients[3](the coefficients of presvote) indicating how sensitive
    voteshare is to changes in presvote. It represents the expected change
     in voteshare for a one-unit increase in presvote, holding all other
    variables constant.
# In this context, voteshare is the dependent variable and difflog is the
     explanatory variable


cat("Prediction Equation: voteshare =", coefficients_model[1], "+",
    coefficients_model[2], "* difflog +", coefficients_model[3], "*
    presvote\n")
# Prediction Equation: voteshare = 0.4486442 + 0.03554309 * difflog +
    0.256877 * presvote
```

The interpretation of the coefficients is as follows:
0.03554309 (Intercept): This represents the predicted value of voteshare when all explanatory variables, including both (difflog + presvote), are zero. In this context, it represents the baseline level of voteshare when all other explanatory variables are held constant.
0.256877 (Coefficient for difflog + presvote): This coefficient signifies the expected change in voteshare for a one-unit increase in the combined effect of (difflog + presvote). It reflects the impact on voteshare when both the difference in spending and incumbent's vote share increase by one unit.
0.4486442 (Coefficient for presvote): This represents the expected change in voteshare for a one-unit increase in the incumbent's vote share (presvote), holding the difference in spending (difflog) constant.
It quantifies the influence of the incumbent's vote share on the overall vote share.'

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 p-values for the coefficients in both Question 4 and Question 5 are extremely small (2.2e-16), indicating highly significant statistical findings.

The model in Question 5 essentially encompasses and extends the analysis from Question 4:

```
Q5: voteshare = β0 + β1 * difflog + β2 * presvote + ϵ
Q4: residuals_Q1 = α0 + α1 * residuals_Q2 + ϵ
residuals_Q1 (voteshare and difflog)
residuals_Q2 (presvote and difflog).
Thus, in Q4: (voteshare and difflog) = α0 + α1 * (presvote and difflog) + ϵ
```

I think that:
When the residuals of a simple linear regression are equivalent to the residuals of a multiple linear regression with two independent variables,
it suggests the presence of collinearity between difflog and presvote.
This implies a certain degree of strong linear relationship between these variables.

In particular, if two or more independent variables exhibit almost perfect linear relationships, they are considered collinear.