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`.

   Below is the R code:

```r
####### Question 1 #######

##### (1) #####

q1_reg_model <- lm(voteshare ~ difflog , data=inc.sub)
summary(q1_reg_model)
```

Table 1: Abstract of Regression Model in Question 1

	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
```

We can conclude from the results:

- F test p-value for the entire regression model is 2.2e-16, which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis.. So at least one slope in regression model is not 0.
- T test p-value for the intercept and slope is both 2e-16,which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis.. So the estimate of intercept and slope is not 0.
- The adjusted R-squared is 0.3671, which means that the fit of this model is average.

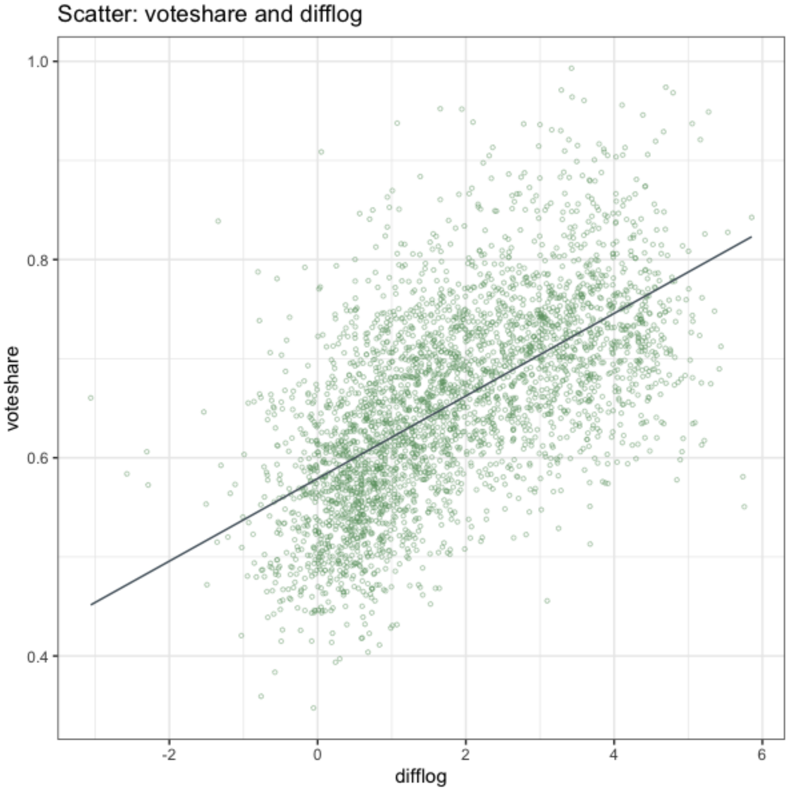
2. Make a scatterplot of the two variables and add the regression line.

   Below is the R code:

```r
##### (2) #####

q1_scatter <- ggplot(inc.sub, aes(x=difflog, y=voteshare)) +
  geom_point(shape=1, size=0.8, color="#609966", alpha=0.4) +
  geom_smooth(method="lm", se=FALSE, color="#52616B", size=0.5) +
  ggtitle("Scatter: voteshare and difflog") +
  theme_bw()
q1_scatter
```

Figure 1: Scatter - Difflog and Voteshare



- Generally speaking, difflog and voteshare are positively correlated.

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

   In regression model, residuals suggests the differences between the predict values and the true values.

   The formulation of residuals is: $e_i = y_i - \hat{y}_i$

   In this formulation:
   e_i is residuals, y_i is the reality value, and $\hat{y}_i$ is the value of the regression model.

   Below is the R code:

```r
##### (3) #####

q1_residuals <- residuals(q1_reg_model)
str(q1_residuals)
summary(q1_residuals)
```

   Below is the output in R studio:

```
> str(q1_residuals) Named num [1:3193] -0.000423 -0.031684 -0.004551
 0.038669 0.035529 ...
- attr(*, "names")= chr [1:3193] "1" "2" "3" "4" ...
> summary(q1_residuals)      Min.    1st Qu.    Median     Mean    3rd Qu.    Max.
-0.268319 -0.053454 -0.003769  0.000000  0.047798  0.327488
```

4. Write the prediction equation.

   Because regression model in R is a list composed of coefficients, residuals, effects and many other elements.

   For coefficients, the first element is the intercept, and the second element is the slope.

   Below is the R code:

```r
##### (4) #####

q1_cof_vec <- coef(q1_reg_model)

q1_intercept <- q1_cof_vec[1]
q1_slope <- q1_cof_vec[2]

q1_pre_equation <- paste(
  "voteshare =", round(q1_intercept, 2), "+", round(q1_slope, 2), "* difflog")
q1_pre_equation
```

   Below is the output in R studio:

```
[1] "voteshare = 0.58 + 0.04 * 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`.

   Below is the R code:

```
####### Question 2 #######

##### (1) #####

q2_reg_model <- lm(presvote ~ difflog, data=inc.sub)
summary(q2_reg_model)
```

Table 2: Abstract of Regression Model in Question 2

	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
```

We can conclude from the results:

- F test p-value for the entire regression model is 2.2e-16, which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis.. So at least one slope in regression model is not 0.
- T test p-value for the intercept and slope is both 2e-16,which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis.. So the estimate of intercept and slope is not 0.
- The adjusted R-squared is 0.08767, which means that the fit of this model is very bad.

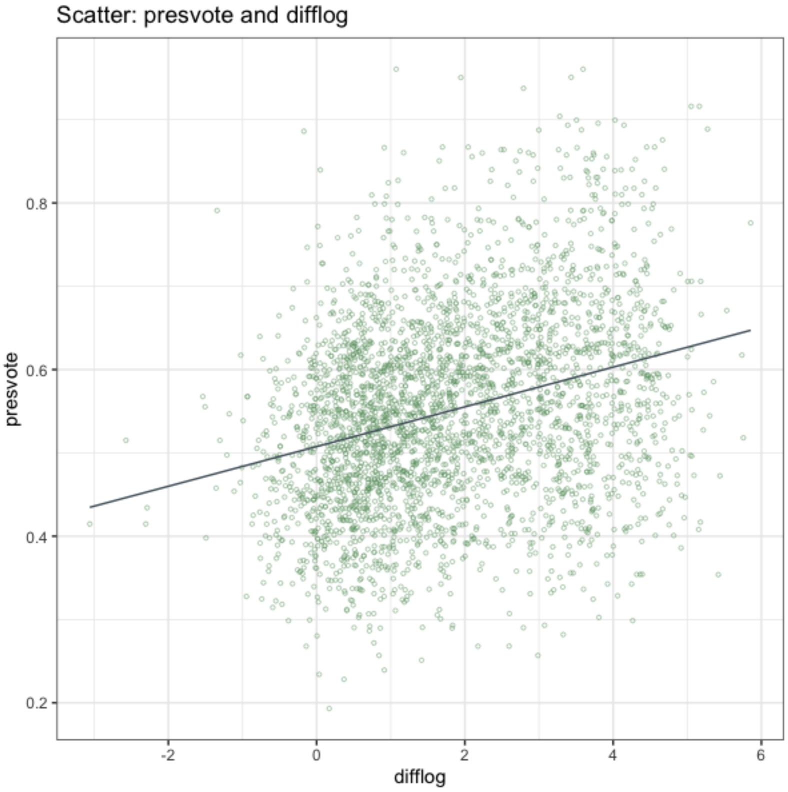
2. Make a scatterplot of the two variables and add the regression line.

   Below is the R code:

```r
##### (2) #####

q2_scatter <- ggplot(inc.sub, aes(x=difflog, y=presvote)) +
  geom_point(shape=1, size=0.8, color="#609966", alpha=0.4) +
  geom_smooth(method="lm", se=FALSE, color="#52616B", size=0.5) +
  ggtitle("Scatter: presvote and difflog") +
  theme_bw()
q2_scatter
```

Figure 2: Scatter - Difflog and Presvote



- Generally speaking, difflog and voteshare are positively correlated.

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

   In regression model, residuals suggests the differences between the predict values and the true values.

   The formulation of residuals is: $e_i = y_i - \hat{y}_i$

   In this formulation:
   e_i is residuals, y_i is the reality value, and $\hat{y}_i$ is the value of the regression model.

   Below is the R code:

```r
##### (3) #####

q2_residuals <- residuals(q2_reg_model)
str(q2_residuals)
summary(q2_residuals)
```

   Below is the output in R studio:

```
> str(q2_residuals) Named num [1:3193] 0.00561 0.03758 -0.05313
-0.05299 -0.04584 ...
- attr(*, "names")= chr [1:3193] "1" "2" "3" "4" ...
> summary(q2_residuals)   Min.  1st Qu.   Median  Mean  3rd Qu.  Max.
-0.321965 -0.074069 -0.001018  0.000000  0.071507  0.427435
```

4. Write the prediction equation.

   Because regression model in R is a list composed of coefficients, residuals, effects and many other elements.

   For coefficients, the first element is the intercept, and the second element is the slope.

   Below is the R code:

```r
##### (4) #####

q2_cof_vec <- coef(q2_reg_model)

q2_intercept <- q2_cof_vec[1]
q2_slope <- q2_cof_vec[2]

q2_pre_equation <- paste(
  "presvote =", round(q2_intercept, 2), "+", round(q2_slope, 2), "* difflog")
q2_pre_equation
```

   Below is the output in R studio:

```
[1] "presvote = 0.51 + 0.02 * 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`.

   Below is the R code:

```r
##### (1) #####

q3_reg_model <- lm(voteshare ~ presvote, data=inc.sub)
summary(q3_reg_model)
```

Table 3: Abstract of Regression Model in Question 3

	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
```

We can conclude from the results:

- F test p-value for the entire regression model is 2.2e-16, which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis. So at least one slope in regression model is not 0.

- T test p-value for the intercept and slope is both 2e-16,which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis. So the estimate of intercept and slope is not 0.

- The adjusted R-squared is 0.08815, which means that the fit of this model is very bad.

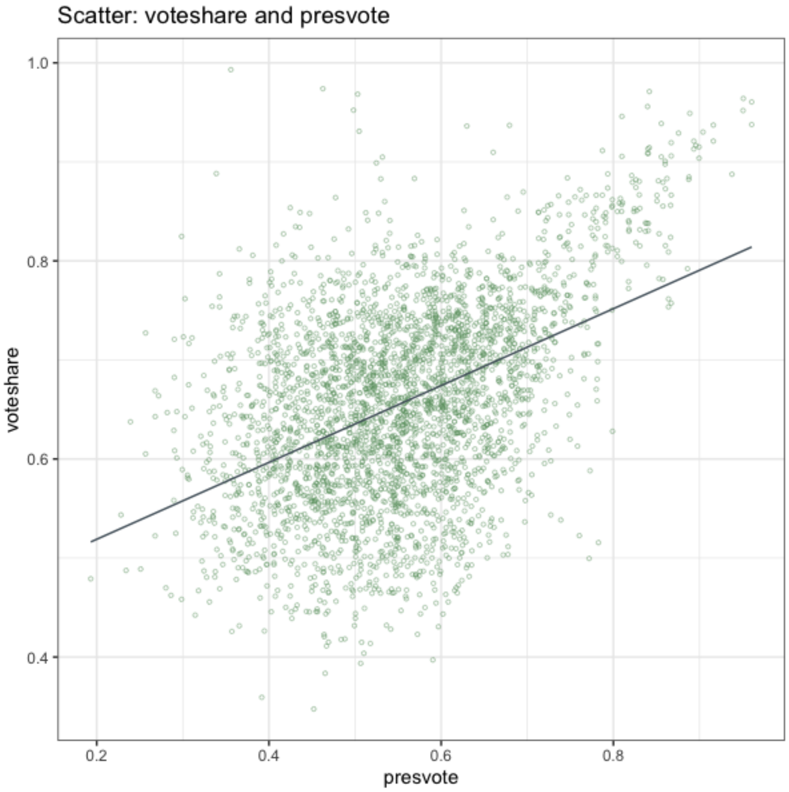
2. Make a scatterplot of the two variables and add the regression line.

   Below is the R code:

```r
##### (2) #####

q3_scatter <- ggplot(inc.sub, aes(x=presvote, y=voteshare)) +
  geom_point(shape=1, size=0.8, color="#609966", alpha=0.4) +
  geom_smooth(method="lm", se=FALSE, color="#52616B", size=0.5) +
  ggtitle("Scatter: voteshare and presvote") +
  theme_bw()
q3_scatter
```

Figure 3: Scatter - Voteshare and Presvote



- Generally speaking, voteshare and presvote are positively correlated.

3. Write the prediction equation.

   Because regression model in R is a list composed of coefficients, residuals, effects and many other elements.

   For coefficients, the first element is the intercept, and the second element is the slope.

   Below is the R code:

```r
##### (3) #####

q3_cof_vec <- coef(q3_reg_model)

q3_intercept <- q3_cof_vec[1]
q3_slope <- q3_cof_vec[2]

q3_pre_equation <- paste(
  "voteshare =", round(q3_intercept, 2), "+", round(q3_slope, 2), "*
    presvote")
q3_pre_equation
```

   Below is the output in R studio:

```
[1] "voteshare = 0.44 + 0.39 * 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.

   Below is the R code:

```r
####### Question 4 #######

##### (1) #####

q4_reg_model <- lm(q1_residuals ~ q2_residuals , data=inc.sub)
summary(q4_reg_model)
```

Table 4: Abstract of Regression Model in Question 4

	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	-1.942e-18	1.299e-03	0.00	1
q2residuals	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:  0.0733 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
```

We can conclude from the results:

- F test p-value for the entire regression model is 2.2e-16, which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis. So at least one slope in regression model is not 0.

- T test p-value for the intercept and slope is correspondingly 1 and 2e-16,the first one didn’t pass the statistics test, which means the intercept is 0. And the second one is very close to 0, which means we have a very high confidence in rejecting the null hypothesis. So the estimate of slope is not 0.

- The adjusted R-squared is 0.1298, which means that the fit of this model is very bad.

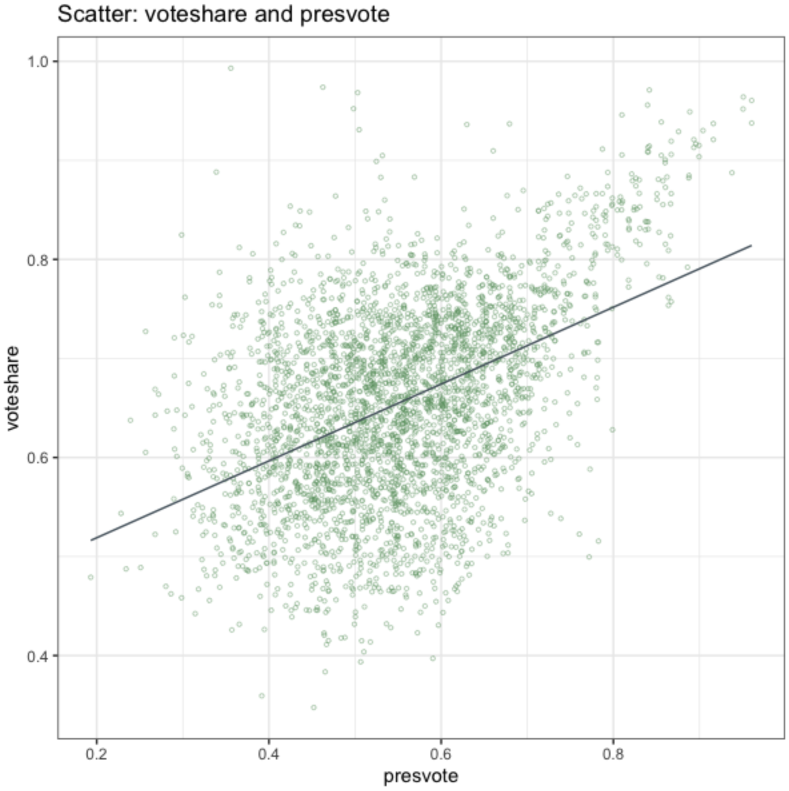
2. Make a scatterplot of the two residuals and add the regression line.

   Below is the R code:

```r
##### (2) #####

q3_scatter <- ggplot(inc.sub, aes(x=presvote, y=voteshare)) +
  geom_point(shape=1, size=0.8, color="#609966", alpha=0.4) +
  geom_smooth(method="lm", se=FALSE, color="#52616B", size=0.5) +
  ggtitle("Scatter: voteshare and presvote") +
  theme_bw()
q3_scatter
```

Figure 4: Scatter - q1residuals and q2residuals



- Generally speaking, q1residuals and q2residuals are positively correlated.

3. Write the prediction equation.

   Because regression model in R is a list composed of coefficients, residuals, effects and many other elements.

   For coefficients, the first element is the intercept, and the second element is the slope.

   Below is the R code:

```r
##### (3) #####

q4_cof_vec <- coef(q4_reg_model)

q4_intercept <- q4_cof_vec[1]
q4_slope <- q4_cof_vec[2]

q4_pre_equation <- paste(
  "q1_residuals =", round(q4_intercept, 2), "+", round(q4_slope, 2), "* q2_residuals")
q4_pre_equation
```

   Below is the output in R studio:

```
[1] "q1_residuals = 0 + 0.26 * q2_residuals"
```

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`.

   Below is the R code:

```r
####### Question 5 #######

##### (1) #####

q5_reg_model <- lm(voteshare ~ difflog + presvote, data=inc.sub)
summary(q5_reg_model)
```

Table 5: Abstract of Regression Model in Question 5

	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:  0.0739 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
```

We can conclude from the results:

- F test p-value for the entire regression model is 2.2e-16, which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis. So at least one slope in regression model is not 0.

- T test p-value for the intercept and slope is both 2e-16, which is very close to 0, which means we have a very high confidence in rejecting the null hypothesis. So the estimate of slope and intercept is not 0.

- The adjusted R-squared is 0.4493, which means that the fit of this model is average.

2. Write the prediction equation.

   Because regression model in R is a list composed of coefficients, residuals, effects and many other elements.

   For coefficients, the first element is the intercept, and the second element is the slope.

   Below is the R code:

```r
##### (2) #####

q5_cof_vec <- coef(q5_reg_model)

q5_intercept <- q5_cof_vec[1]
q5_slope1 <- q5_cof_vec[2]
q5_slope2 <- q5_cof_vec[3]

q5_pre_equation <- paste(
  "voteshare =", round(q5_intercept, 2),
  "+", round(q5_slope1, 2), "* difflog",
  "+", round(q5_slope2, 2), "* presvote")
q5_pre_equation
```

   Below is the output in R studio:

```
[1] "voteshare = 0.45 + 0.04 * difflog + 0.26 * 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?

   I found that the residuals in question 4's regression model is the same as the one in question 5. We can check that in R code and box plot. Below is the R code:

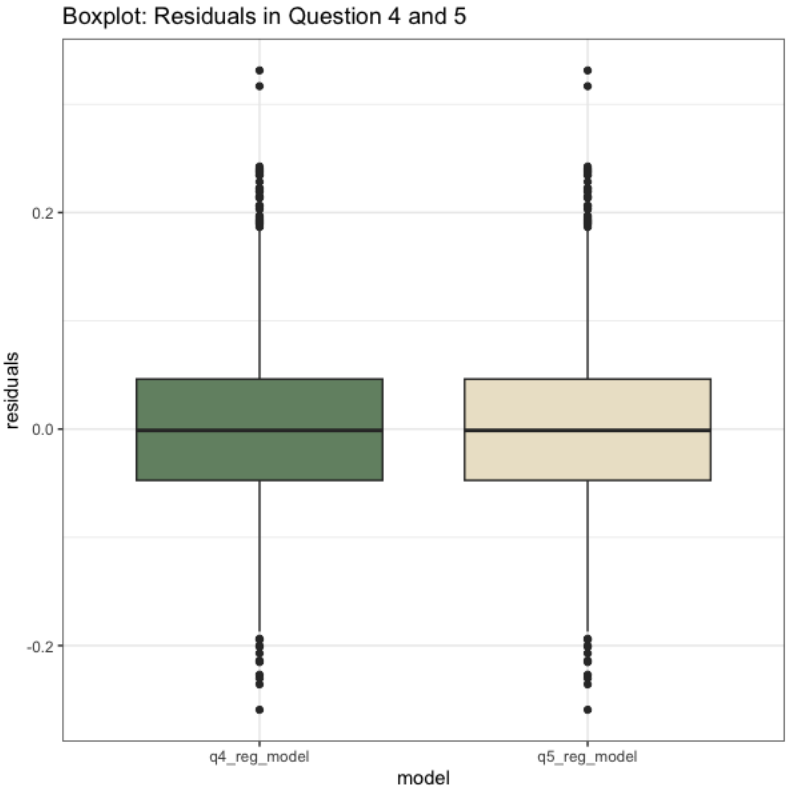
```r
##### (3) #####

q4_residuals <- as.vector(resid(q4_reg_model))
q5_residuals <- as.vector(resid(q5_reg_model))
str(q4_residuals); str(q5_residuals)

q5_data <- data.frame(
  model = rep(c("q4_reg_model", "q5_reg_model"),
              each=length(q4_residuals)),
  residuals = c(q4_residuals, q5_residuals)
)

q5_boxplot <- ggplot(q5_data, aes(x=model, y=residuals, fill=model)) +
  geom_boxplot(fill=c("#739072", "#ECE3CE")) +
  ggtitle("Boxplot: Residuals in Question 4 and 5") +
  theme_bw()
q5_boxplot
```

Figure 5: Boxplot - Residuals in Question 4 and 5



If two regression model's residuals is the same, we can conclude that:

- The homoskedasticity assumption is met: the residual variances of the two models are similar at different levels of the explanatory variables.
- The model fits the data well: The presence of homoscedasticity may indicate that the model fits the data well because it meets the basic assumptions of linear regression.
- Reliability of model comparisons and interpretations: Model comparisons and interpretation of model results can be made more reliably if the residuals of two models are the same.