Lab 8

Code

Julian Adames-Ng

2022-10-24

Hide

```
library(tidyverse)
library(openintro)
data('hfi', package='openintro')
```

Exercise 1

What are the dimensions of the dataset?

Hide

```
# Insert code for Exercise 1 here

dim(hfi)
```

```
## [1] 1458  123
```

Exercise 2

What type of plot would you use to display the relationship between the personal freedom score, pf_score, and one of the other numerical variables? Plot this relationship using the variable pf_expression_control as the predictor. Does the relationship look linear? If you knew a country's pf_expression_control, or its score out of 10, with 0 being the most, of political pressures and controls on media content, would you be comfortable using a linear model to predict the personal freedom score?

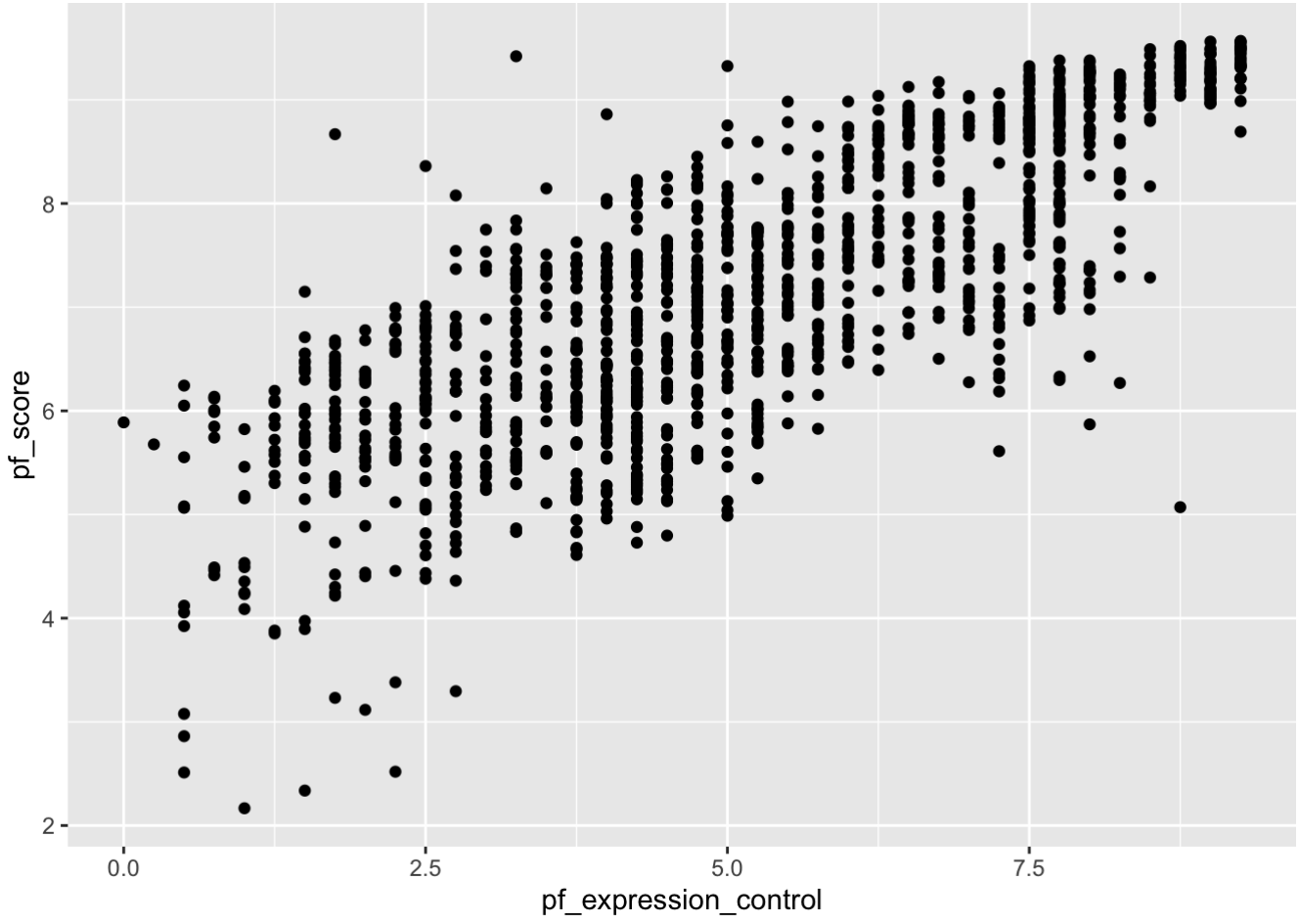
-I'd use a scatterplot to display the relationship between the personal freedom score and another numerical variable. Using expression control as the predictor variable, the relationship does appear to be linear, although more on the moderate side…for example; the linear correlation coefficient, r, may be +0.65. If I knew a country's pf_expression_control I'd have a reasonable level of comfort using a linear model.

...

Hide

```
# Insert code for Exercise 1 here

#ggplot(hfi, aes(x=pf_score, y=pf_expression_control)) +
#  geom_point()

ggplot(hfi, aes(x=pf_expression_control, y=pf_score)) +
  geom_point()
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



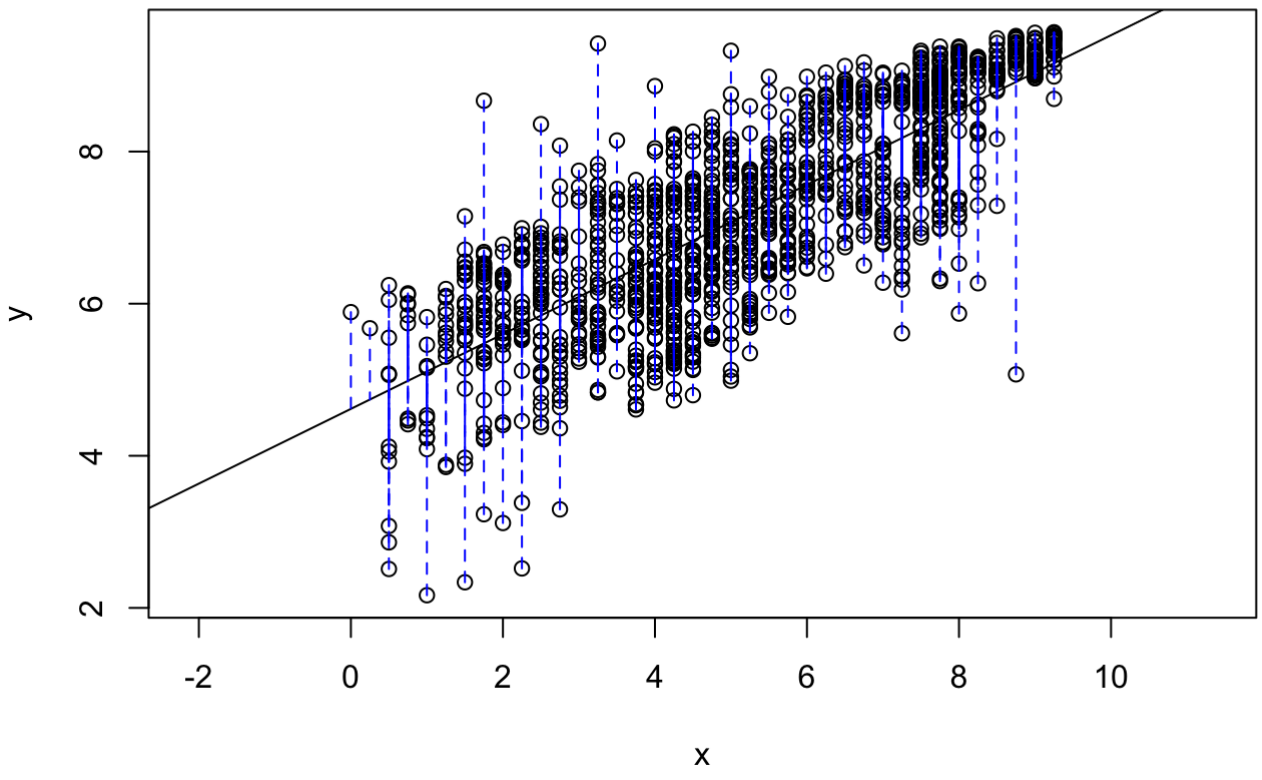
Exercise 3

Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

-As mentioned in my answer for Exercise 2, there is a linear correlation between the variables, but a more moderate and positive relationship. The regression line that is graphed along with the scattered data points has most data within the moderate range with about 20-25 data points that appear to be visual outliers. ...

Hide

```
# Insert code for Exercise 1 here

#hfi %>%
#  summarise(cor(pf_expression_control, pf_score, use = "complete.obs"))


hfi1 <- hfi[c("pf_score", "pf_expression_control")]

hfi1 <- drop_na(hfi1)
row.names <- NULL


#hfi_n <- drop_na(hfi[c('pf_score','pf_expression_control')])

DATA606::plot_ss(x = hfi1$pf_expression_control, y = hfi1$pf_score)
```



```
## Click two points to make a line.

## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##      4.6171       0.4914
##
## Sum of Squares:  952.153
```

Exercise 4

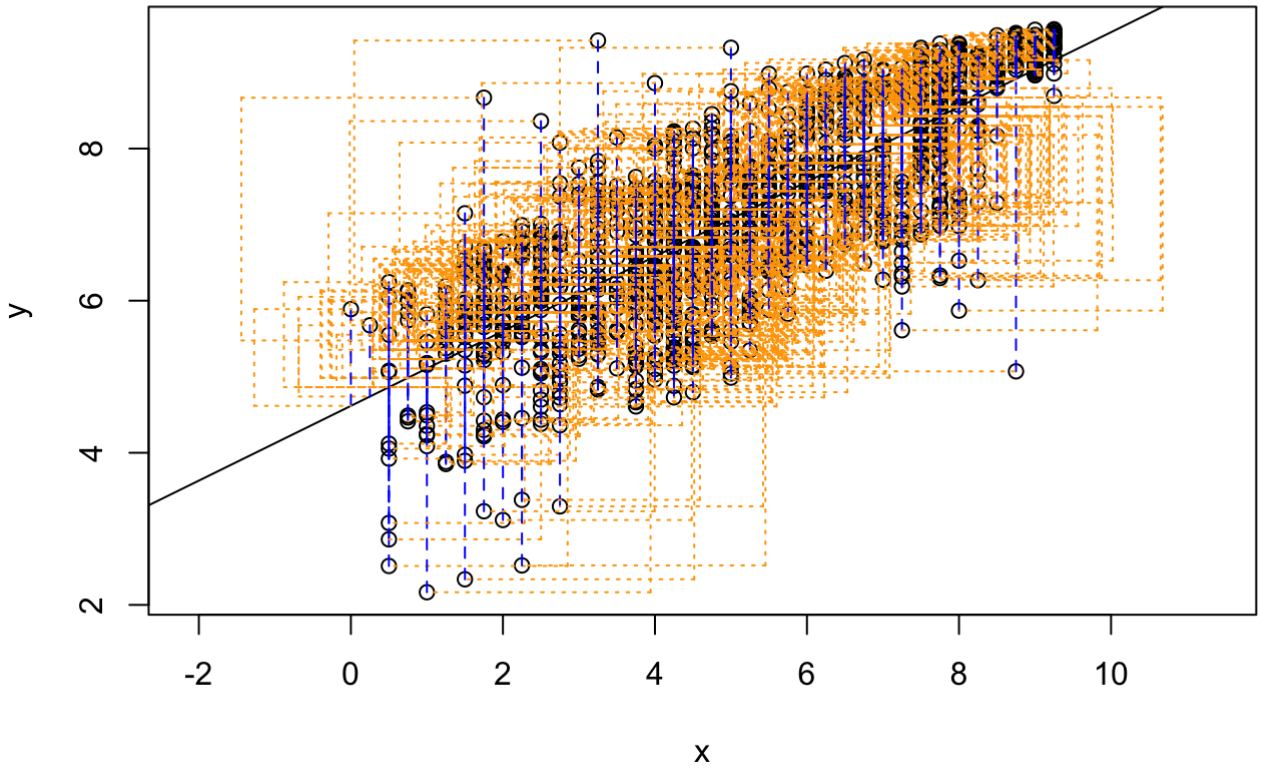
Using plot_ss, choose a line that does a good job of minimizing the sum of squares. Run the function several times. What was the smallest sum of squares that you got? How does it compare to your neighbors?

-The smallest sum of squares that I got was 952.153.

...

Hide

```
# Insert code for Exercise 1 here

#DATA606::plot_ss(x = hfi_n$pf_expression_control, y = hfi_n$pf_score, showSquares
         = TRUE)
DATA606::plot_ss(x = hfi1$pf_expression_control, y = hfi1$pf_score, showSquares = T
        RUE)
```




```
## Click two points to make a line.

## Call:
## lm(formula = y ~ x, data = pts)
##
## Coefficients:
## (Intercept)            x
##      4.6171       0.4914
##
## Sum of Squares:  952.153
```

Exercise 5

Fit a new model that uses pf_expression_control to predict hf_score, or the total human freedom score. Using the estimates from the R output, write the equation of the regression line. What does the slope tell us in the context of the relationship between human freedom and the amount of political pressure on media content?

-The Intercept is 5.153687 and the slope is 0.49143, so the equation of the regression line is hf_score = 5.153687 + 0.349862*(pf_expression_control).

-The slope tells us that as the amount of political pressure increases by one increment, human freedom score increases by 0.349862 of an increment.

...

Hide

```
# Insert code for Exercise 1 here

m1 <- lm(pf_score ~ pf_expression_control, data = hfi)
summary(m1)
```

```
##
## Call:
## lm(formula = pf_score ~ pf_expression_control, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8467 -0.5704  0.1452  0.6066  3.2060
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.61707    0.05745   80.36   <2e-16 ***
## pf_expression_control  0.49143    0.01006   48.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8318 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.6342, Adjusted R-squared:  0.634
## F-statistic:  2386 on 1 and 1376 DF,  p-value: < 2.2e-16
```

Hide

```
m2 <- lm(hf_score ~ pf_expression_control, data = hfi)
summary(m2)
```

```
##
## Call:
## lm(formula = hf_score ~ pf_expression_control, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6198 -0.4908  0.1031  0.4703  2.2933
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.153687   0.046070  111.87   <2e-16 ***
## pf_expression_control 0.349862   0.008067   43.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.667 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.5775, Adjusted R-squared:  0.5772
## F-statistic:  1881 on 1 and 1376 DF,  p-value: < 2.2e-16
```

Exercise 6

If someone saw the least squares regression line and not the actual data, how would they predict a country's personal freedom score for one with a 6.7 rating for pf_expression_control? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

-We use the equation pf_score = 4.61707 + 0.49143 * (pf_expression_control) with a pf_expression_control value of 6.7 to predict a country's personal freedom score. It seems to be an overestimate of 0.48 because the residual for this prediction is -0.48. …

Hide

```
# Insert code for Exercise 1 here

#Using the formula from above and taking pf_expression_control = 6.7 we have
pf_score67 <- 4.61707 + 0.49143 * 6.7
pf_score67
```

```
## [1] 7.909651
```

Hide

```
hfi %>%
  group_by(pf_score) %>%
  filter(pf_expression_control == 6.7)
```

```
## # A tibble: 0 × 123
## # Groups:   pf_score [0]
## # … with 123 variables: year <dbl>, ISO_code <chr>, countries <chr>,
## #   region <chr>, pf_rol_procedural <dbl>, pf_rol_civil <dbl>,
## #   pf_rol_criminal <dbl>, pf_rol <dbl>, pf_ss_homicide <dbl>,
## #   pf_ss_disappearances_disap <dbl>, pf_ss_disappearances_violent <dbl>,
## #   pf_ss_disappearances_organized <dbl>,
## #   pf_ss_disappearances_fatalities <dbl>, pf_ss_disappearances_injuries <dbl>,
## #   pf_ss_disappearances <dbl>, pf_ss_women_fgm <dbl>, …
```

Hide

```
residual <- 7.43 - 7.91
residual
```

```
## [1] -0.48
```

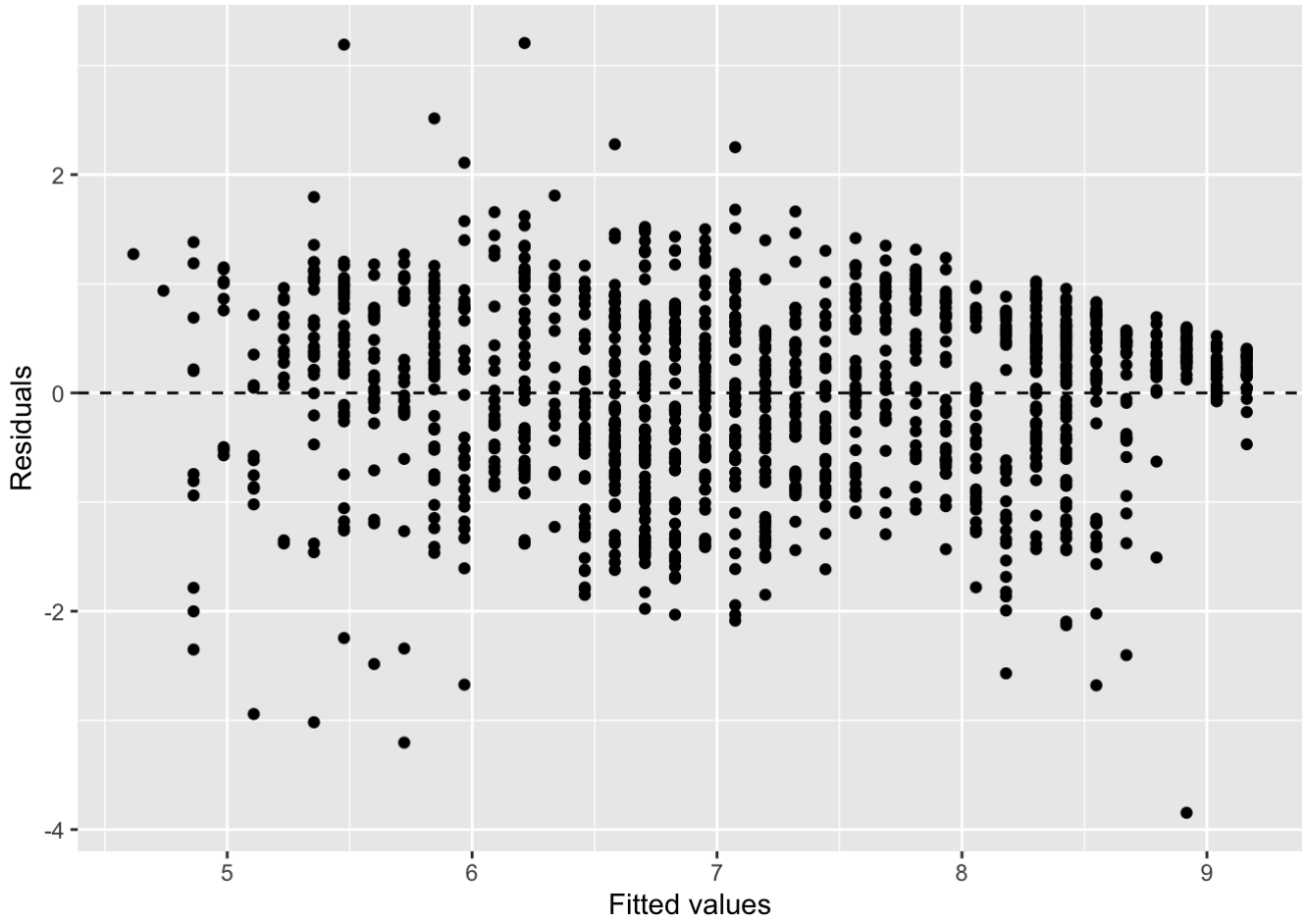
Exercise 7

Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between the two variables?

-There does not seem to be a discernable pattern between the predictor and response variables so it is safe to say that they may have a linear relationship. …

Hide

```
# Insert code for Exercise 1 here

ggplot(data = m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



Exercise 8

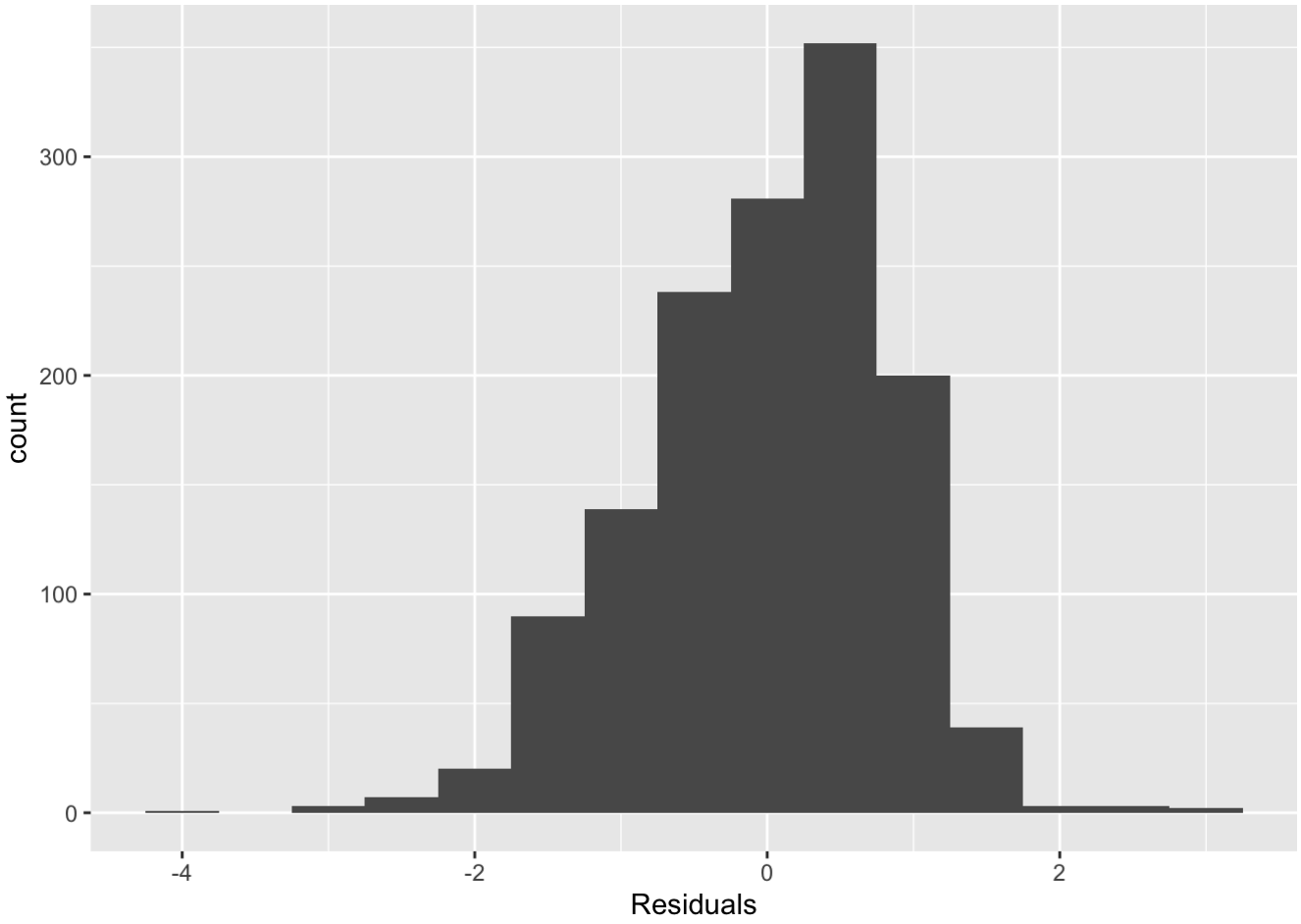
Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

-The nearly normal residuals condition appears to be met since the plots for the histogram and normal probability plot indicate normal data.

...

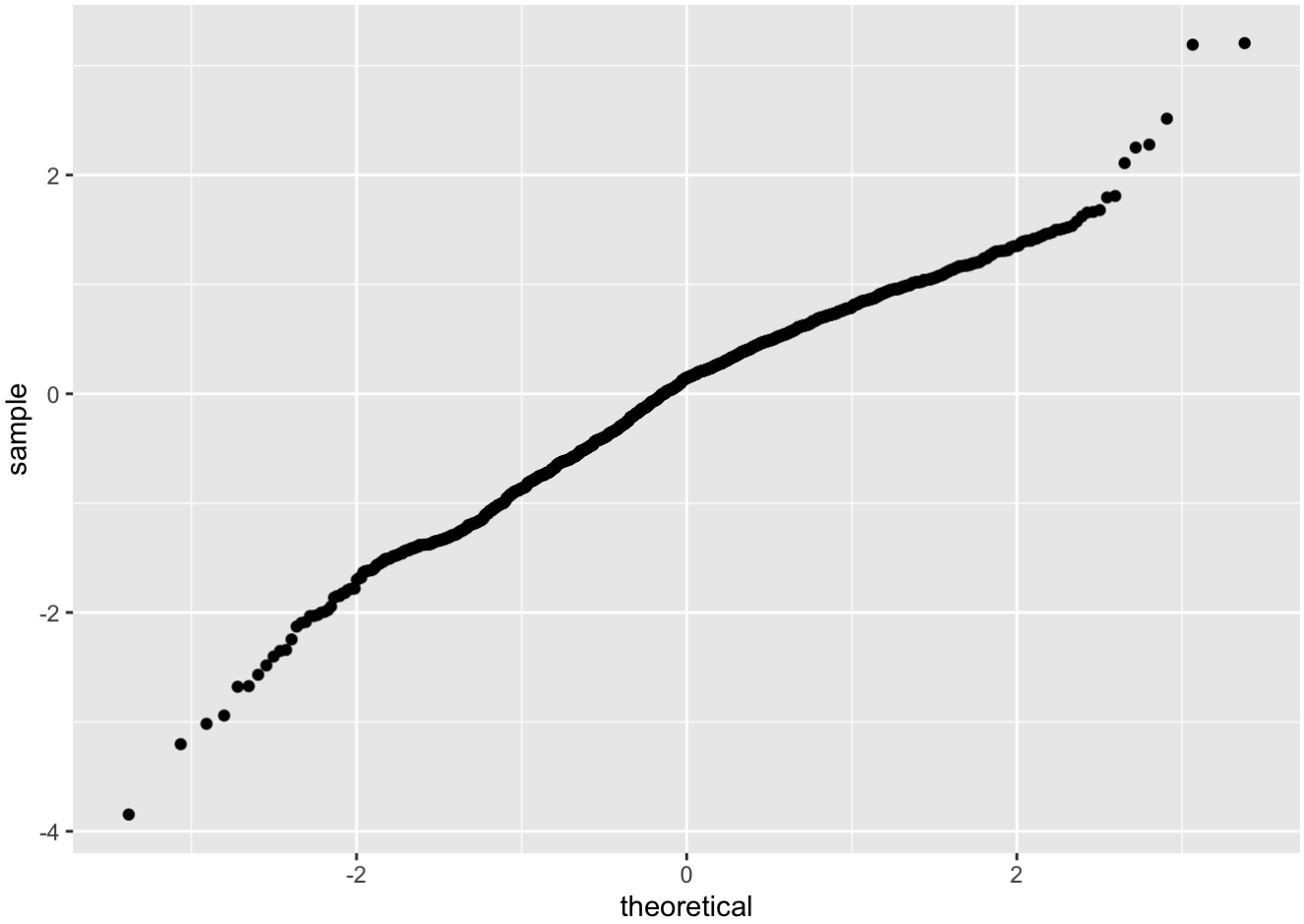
Hide

```
# Insert code for Exercise 1 here

ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(binwidth = 0.5) +
  xlab("Residuals")
```



Hide

```
ggplot(data = m1, aes(sample = .resid)) +
  stat_qq()
```



Exercise 9

Based on the residuals vs. fitted plot, does the constant variability condition appear to be met?

-Yes, the constant variability condition appears to be met. The plot shows a a grouping of data around 0, which is constant.

...