DSC 385 - Project 2 Report

Name and EID

Name: Marina Witherell
EID: mw44788

Setup

Read in the data with `read_csv()` and store the data as an R object named `dataset`. Check the data to make sure all of the expected observations and variables are there.

```
dataset <- read.csv('maacs.csv')
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.6
✓ forcats   1.0.1     ✓ stringr   1.6.0
✓ ggplot2   4.0.1     ✓ tibble    3.3.1
✓ lubridate 1.9.4     ✓ tidyr     1.3.2
✓ purrr     1.2.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
library(broom)
library(patchwork)
```

Part 1

We will first consider the relationship between FEV1 and age. In general, it is expected that as children get older (and hence, larger in size), their FEV1 values should get higher. Verify this by fitting a linear regression model with FEV1 as the outcome and age as a predictor.

How much does FEV1 change for a 1-year increase in the child's age? (Write your answer below the code block.)

```
fit <- lm(fev1 ~ age, data = dataset, na.action = na.exclude)
fit
```

```

Call:
lm(formula = fev1 ~ age, data = dataset, na.action = na.exclude)
```

```
Coefficients:
(Intercept)          age
  -0.001199     0.171162
```

For a 1-year increase in the child's age, FEV1 changes by 0.171162.

Part 2

Make a scatterplot of FEV1 on the y-axis and age on the x-axis. Overlay the fitted regression line using the fitted model from Part 1.

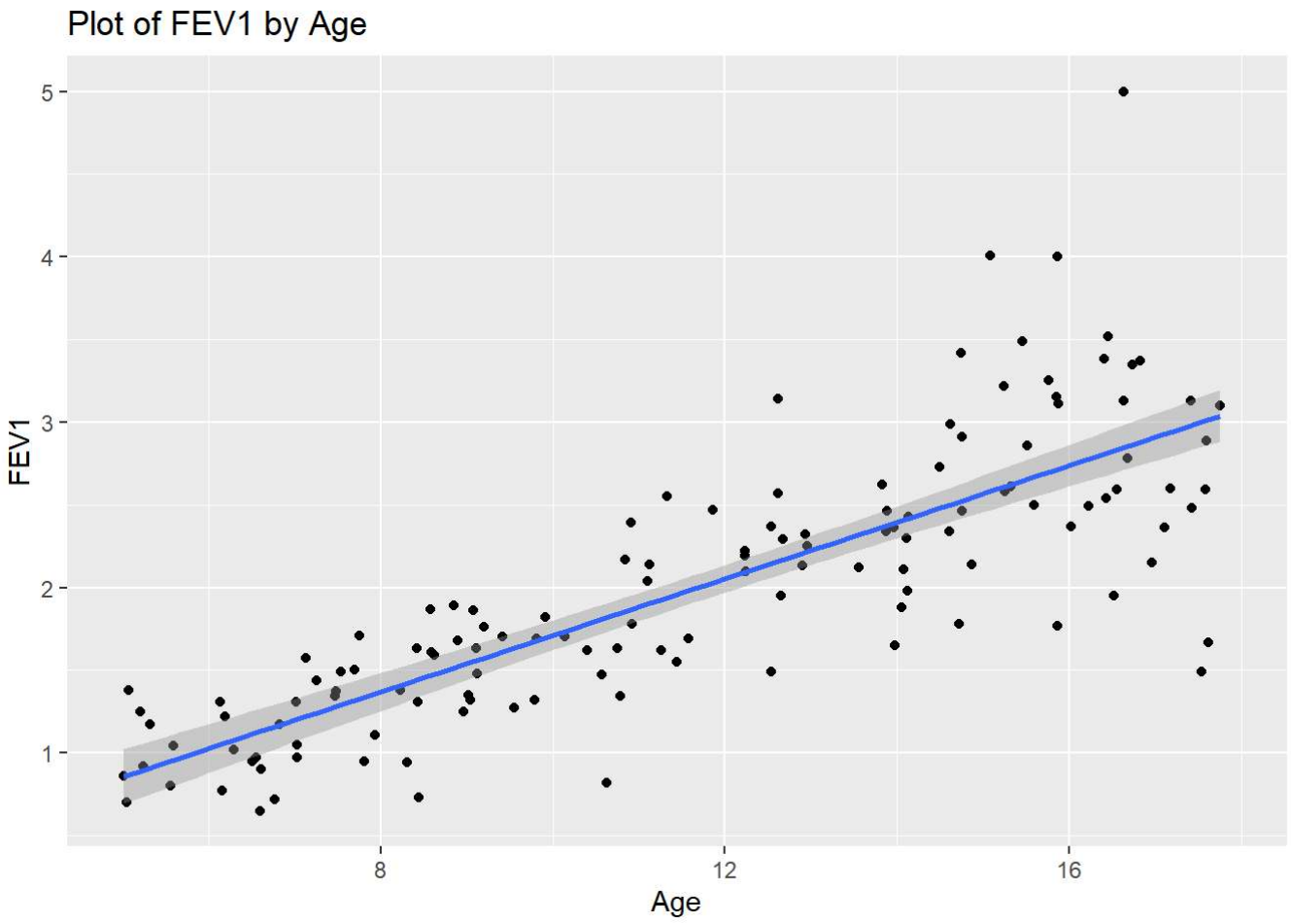
How would you characterize the fit of the regression line to the data? Is there anything unexpected in the plot?

```
fit |>
  augment(dataset) |>
  ggplot(aes(x = age, y = fev1)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(
    title = "Plot of FEV1 by Age",
    x = "Age",
    y = "FEV1"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 17 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 17 rows containing missing values or values outside the scale range
(`geom_point()`).
```



From the regression line, we can conclude that the data is positively correlated. This is not unexpected because, as stated earlier, it is expected that as children get older, their FEV1 values should get higher. However, it looks like as age increases, the residuals also increase, indicating that this model may not be the most accurate.

Part 3

Make a scatterplot of the fitted values (x-axis) from the regression model and the residuals (y-axis).

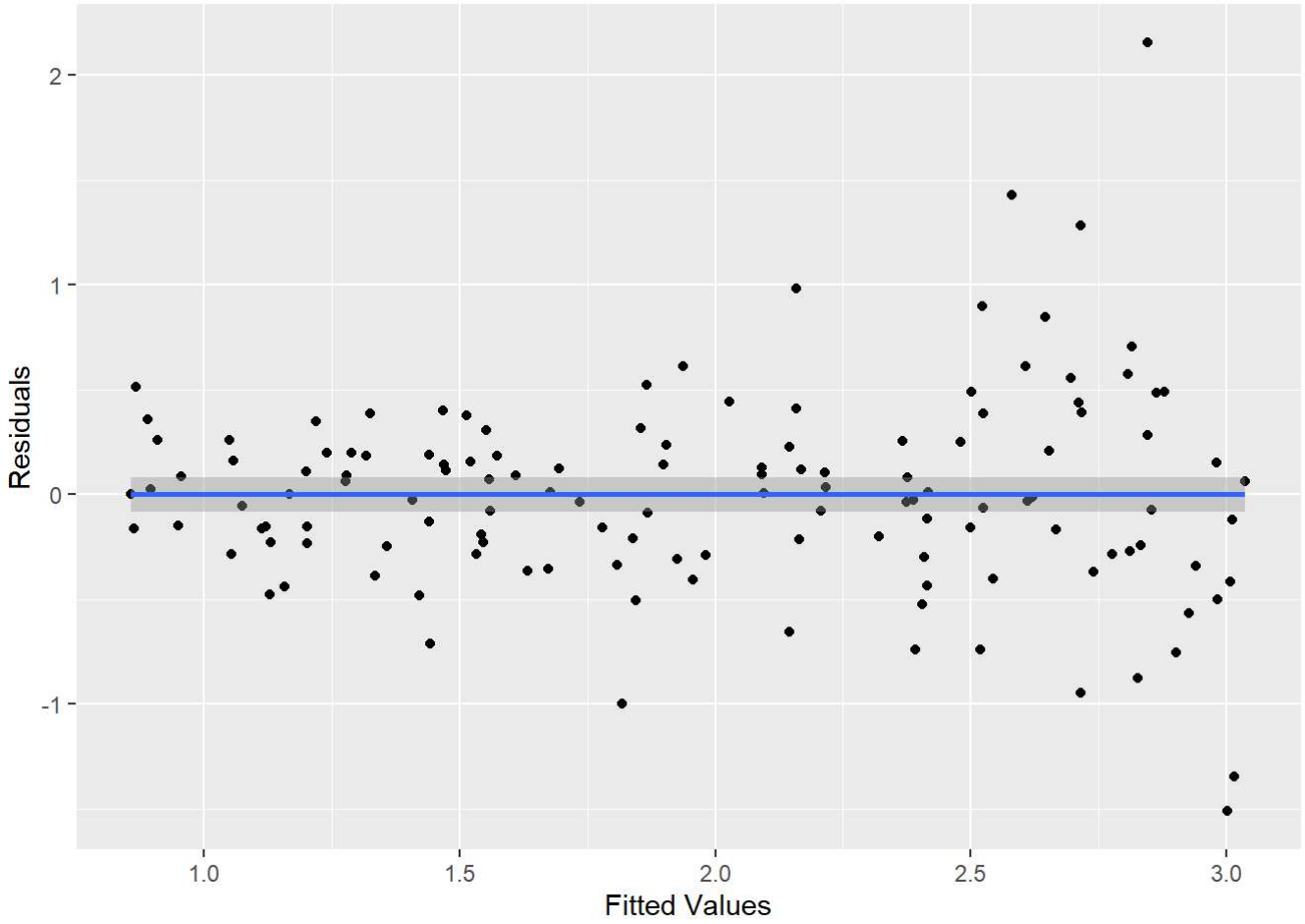
1. Describe the pattern of the residuals in this plot.
2. Recall the assumptions required for making inference from a linear model. Do any of those assumptions appear to be violated?

```
fit |>
  augment(dataset) |>
  ggplot(aes(.fitted, .resid)) +
  geom_point() +
  geom_smooth(method = "gam") +
  labs(x = "Fitted Values", y = "Residuals")
```

```
`geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```

```
Warning: Removed 17 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 17 rows containing missing values or values outside the scale range
(`geom_point()`).
```



1. The residuals do appear to be linear, as they are equally distributed around the zero line. However, they make a "megaphone" shape as the fitted values increase (their spread around the zero line increases).
2. From this graph, it appears that the assumption of homoscedasticity is violated. In other words, variance does not appear to be constant.

Part 4

In previous studies, males have tended to have higher FEV1 values than females. Fit a linear regression model with FEV1 as the outcome and gender as the predictor. **Using the results of the model fit**, answer the following questions:

1. On average in this dataset, which gender has higher FEV1 values?
2. On average in this dataset, what is the difference in FEV1 values between males and females?

NOTE: For both of these questions, you are **not** being asked to make a statistical inference.

```r
fit1 <- lm(fev1 ~ gender, data = dataset, na.action = na.exclude)
fit1
```

```
Call:
lm(formula = fev1 ~ gender, data = dataset, na.action = na.exclude)

Coefficients:
(Intercept)   gendermale
     2.1325      -0.2492
```

1. On average, males have lower FEV1 values than females (in this dataset).
2. On average in this dataset, the difference in FEV1 values between males and females is 0.2492.

Part 5

Now fit a linear regression model with FEV1 as the outcome and gender *and* age as predictors. Using the results of the model fit, answer the following questions:

1. On average in this dataset, which gender has higher FEV1 values when age is held constant?
2. On average in this dataset, what is the difference in FEV1 values between males and females, when age is held constant?

```r
fit2 <- lm(fev1 ~ gender + age, data = dataset, na.action = na.exclude)
fit2
```

```
Call:
lm(formula = fev1 ~ gender + age, data = dataset, na.action = na.exclude)

Coefficients:
(Intercept)   gendermale          age
   -0.06745      0.07178      0.17342
```

1. On average in this dataset, males have higher FEV1 values when age is held constant.
2. On average in this dataset, the difference in FEV1 values between males and females when age is held constant is 0.07178.

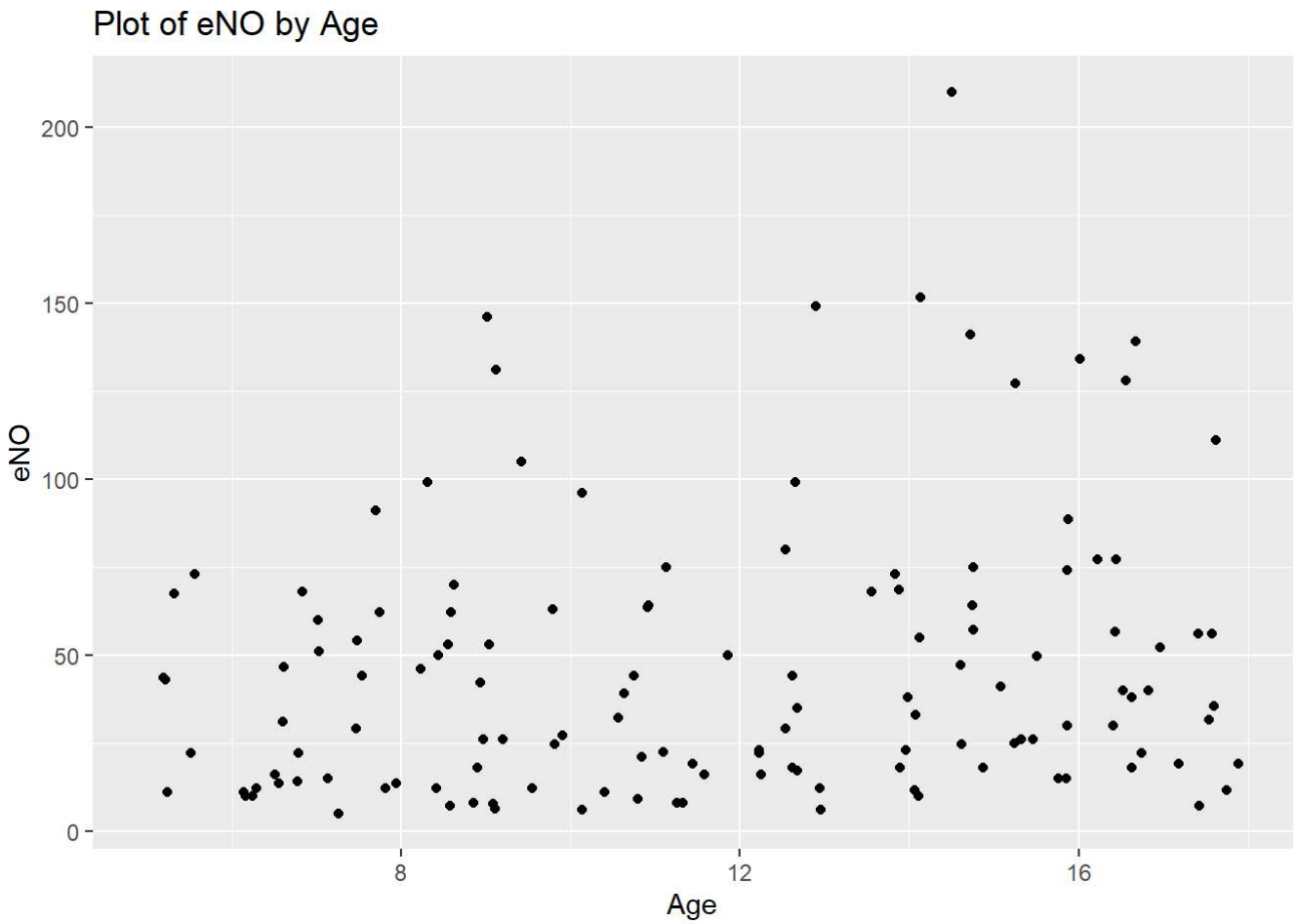
Part 6

We will now move to looking at the eNO variable.

1. Make a scatterplot of eNO (y-axis) and age (x-axis). Briefly describe the relationship between eNO and age.

```r
dataset |>
  ggplot(aes(x = age, y = eNO)) +
  geom_point() +
  labs(
    title = "Plot of eNO by Age",
    x = "Age",
    y = "eNO"
  )
```

```
Warning: Removed 17 rows containing missing values or values outside the scale range
(`geom_point()`).
```



2. Fit a linear model with eNO as the outcome and age as the predictor variable. Overlay the fitted regression line on the scatterplot of eNO and age. How would you characterize the quality of the fit of this model? Do any of the linear regression model assumptions appear to be violated?

```r
fit3 <- lm(eNO ~ age, data = dataset, na.action = na.exclude)
fit3
```

```
Call:
lm(formula = eNO ~ age, data = dataset, na.action = na.exclude)
```

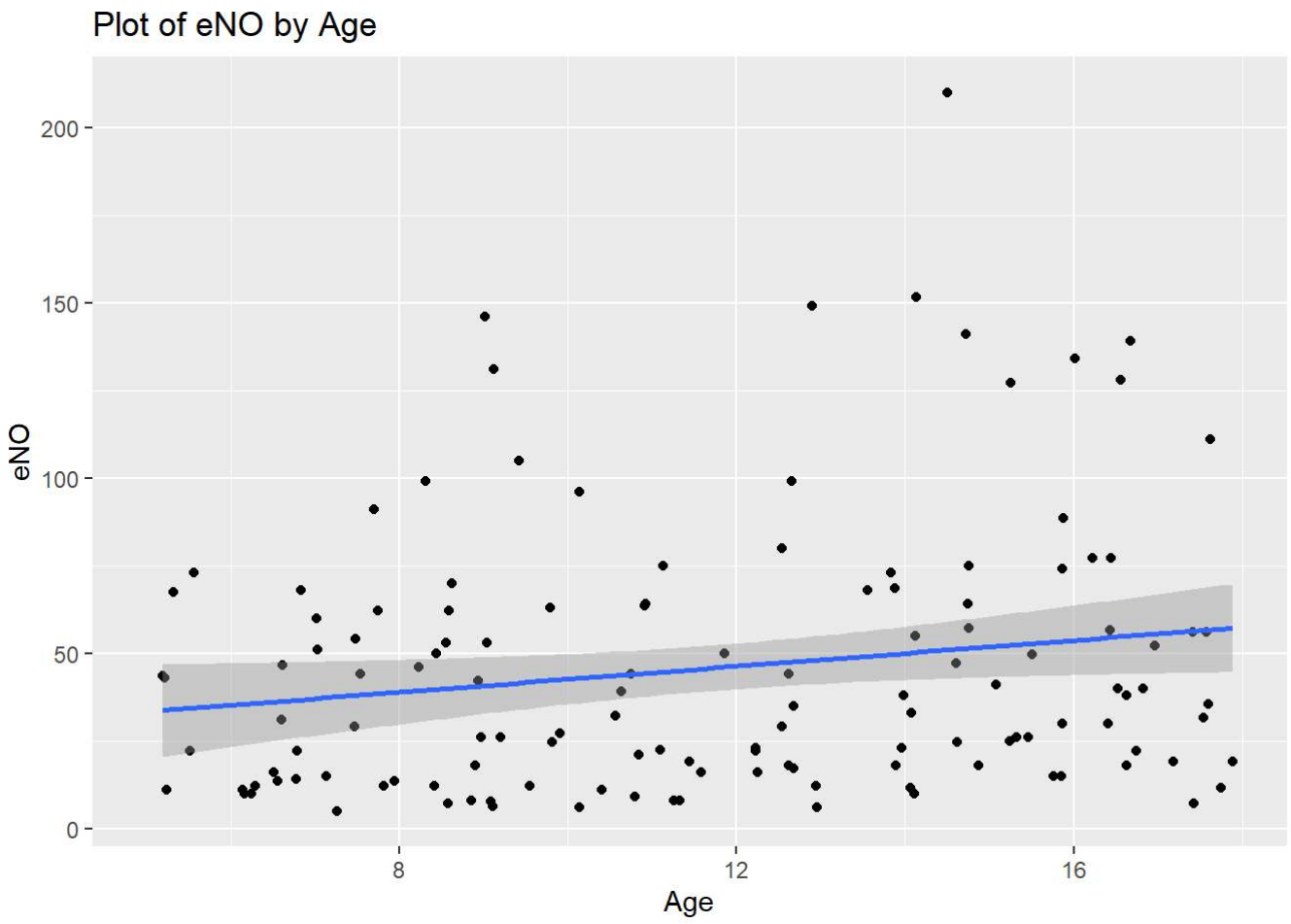
```
Coefficients:
(Intercept)          age
     24.085        1.847
```

```r
fit3 |>
  augment(dataset) |>
  ggplot(aes(x = age, y = eNO)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(
    title = "Plot of eNO by Age",
    x = "Age",
    y = "eNO"
  )
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 17 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 17 rows containing missing values or values outside the scale range
(`geom_point()`).
```



The data points appear to not be randomly scattered around the fitted line. There are more points under the line, meaning that some of the points high above the line are causing leverage on the model. This suggests that the assumption of normality may be violated.

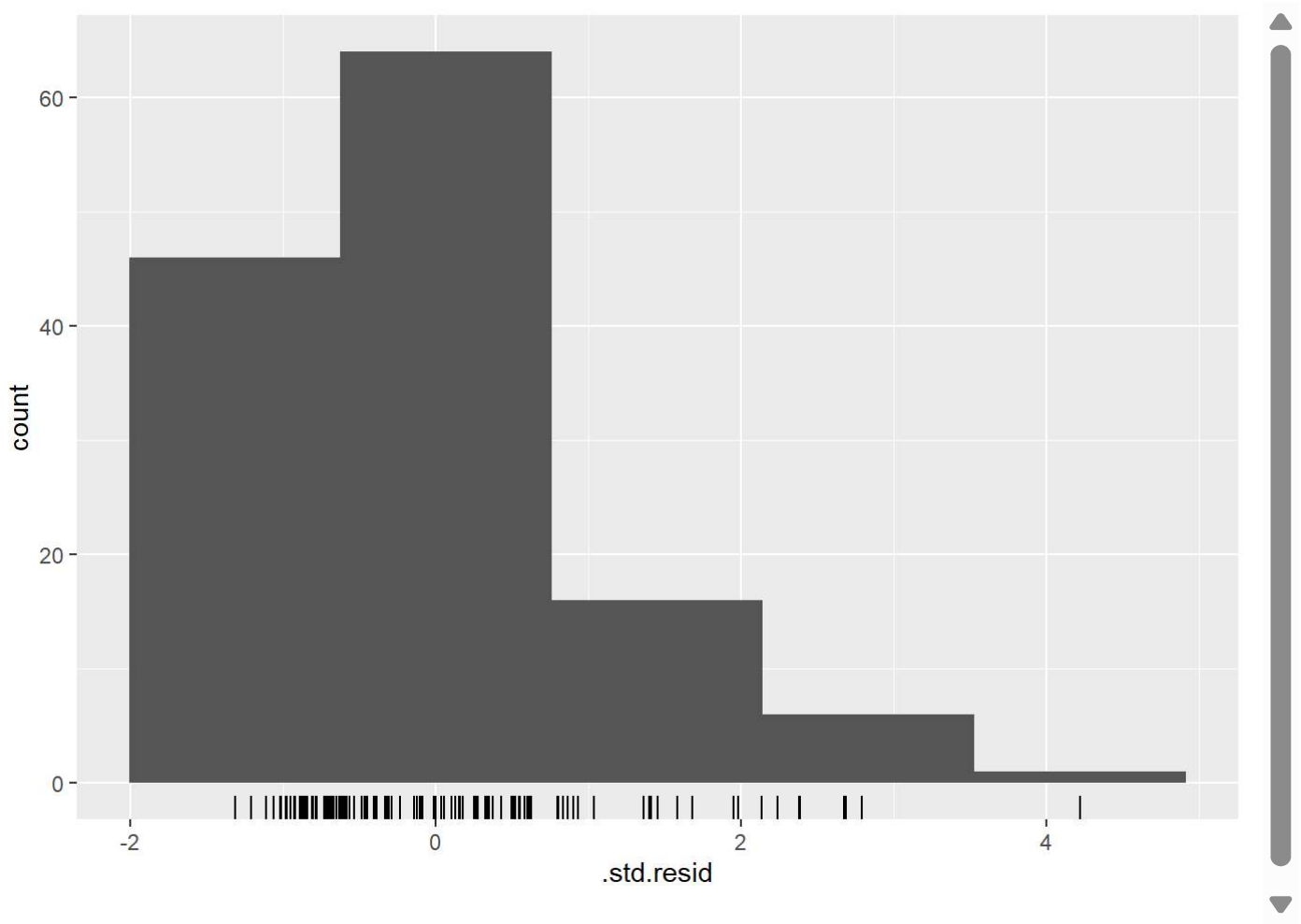
Part 7

Make a histogram of the standardized residuals for the fitted model of eNO and age. Use a few different `bin` choices in the `geom_histogram()` function to see how the histogram changes. How would you characterize the shape of the histogram here? Do any of the linear regression model assumptions appear to be violated?

```
fit3 |>
  augment(dataset) |>
  ggplot(aes(x= .std.resid)) +
  geom_histogram(bins = 5) +
  geom_rug()
```

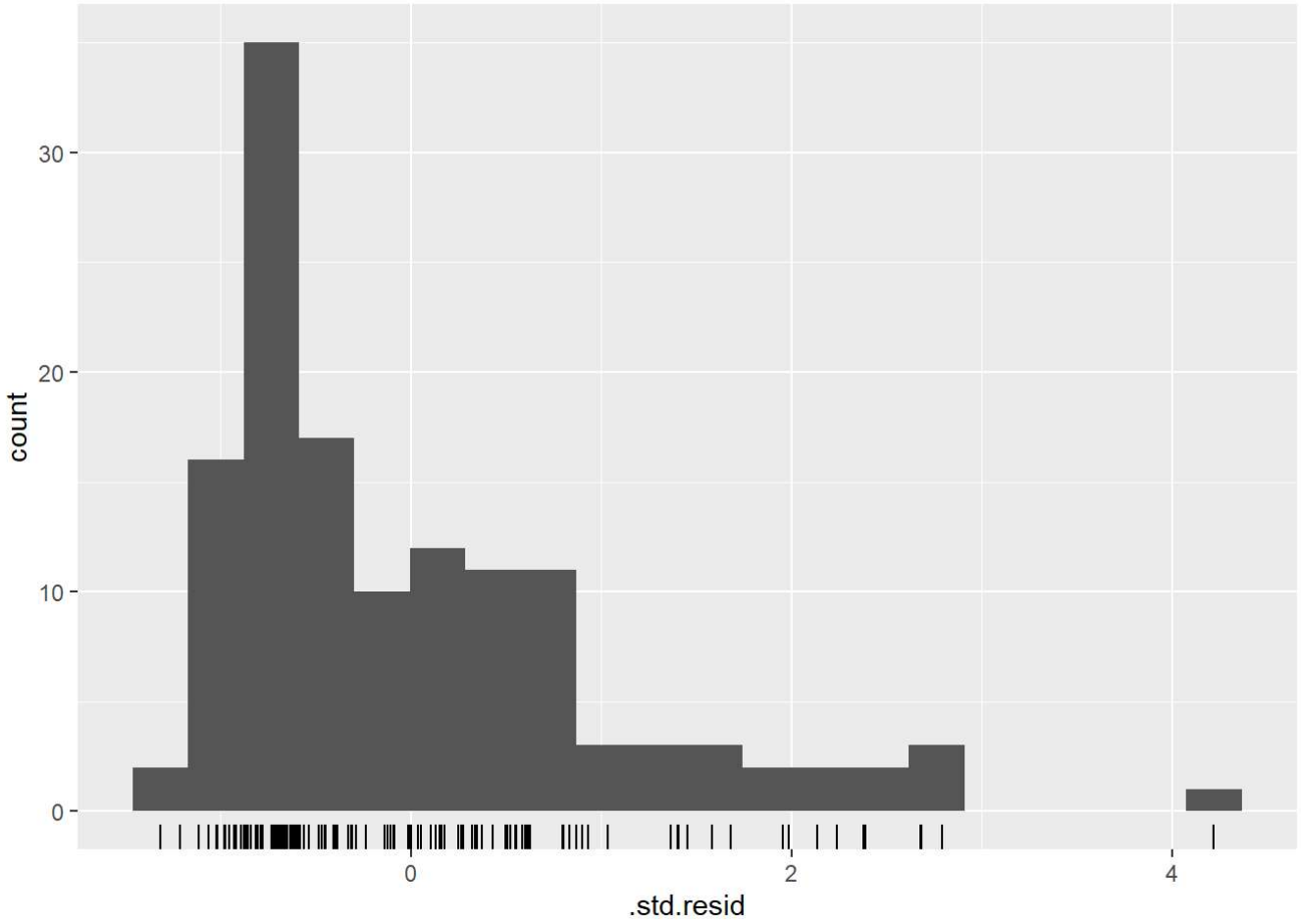
```
Warning: Removed 17 rows containing non-finite outside the scale range
(`stat_bin()`).

Warning: Removed 17 rows containing missing values or values outside the scale range
(`geom_rug()`).
```



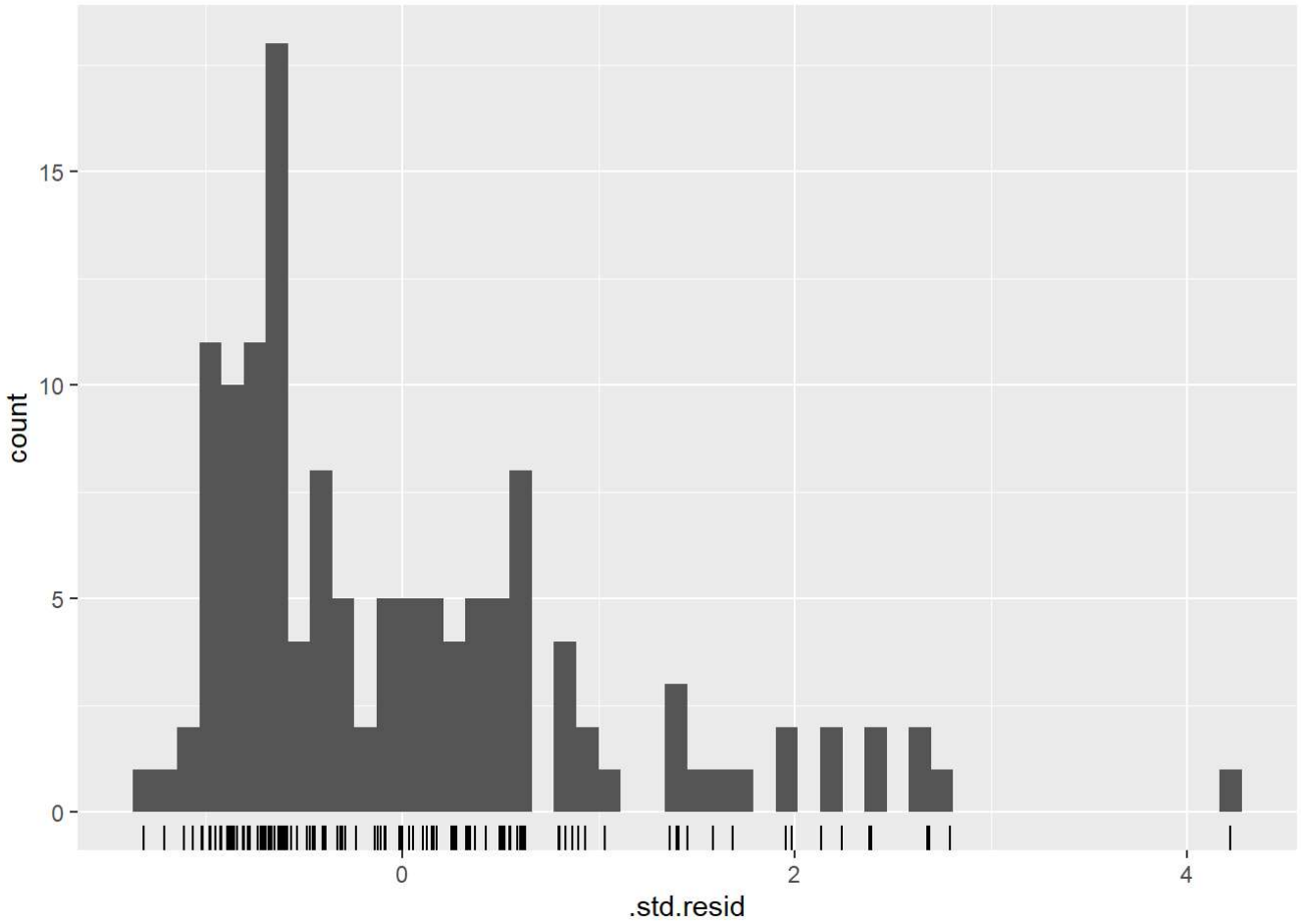
```
fit3 |>
  augment(dataset) |>
  ggplot(aes(x= .std.resid)) +
  geom_histogram(bins = 20) +
  geom_rug()
```

```
Warning: Removed 17 rows containing non-finite outside the scale range (`stat_bin()`).
Removed 17 rows containing missing values or values outside the scale range
(`geom_rug()`).
```




```
fit3 |>
  augment(dataset) |>
  ggplot(aes(x= .std.resid)) +
  geom_histogram(bins = 50) +
  geom_rug()
```

```
Warning: Removed 17 rows containing non-finite outside the scale range (`stat_bin()`).
Removed 17 rows containing missing values or values outside the scale range
(`geom_rug()`).
```



The shape of the histogram shows evidence of right-skewness. This means that the normality assumption is violated.

Part 8

Create a new dataset that replaces eNO with the log of eNO (using the `log()` function). Store this new dataset in an object called `dataset2`.

Using `dataset2`, fit a new linear regression model with the (now log-transformed) eNO variable as the response and age as the predictor.

1. Make a plot of the residuals vs. fitted values for this new model (residuals on the y-axis, fitted values on the x-axis).

2. Also make a histogram of the standardized residuals from this new model.

How would you characterize the fit of this model versus the fit of the previous model that did not use the log-transformed eNO variable? Which model is preferable, based on the diagnostic plots you have seen?

```
dataset2 <- dataset |>
  mutate(eNO = log(eNO))
```

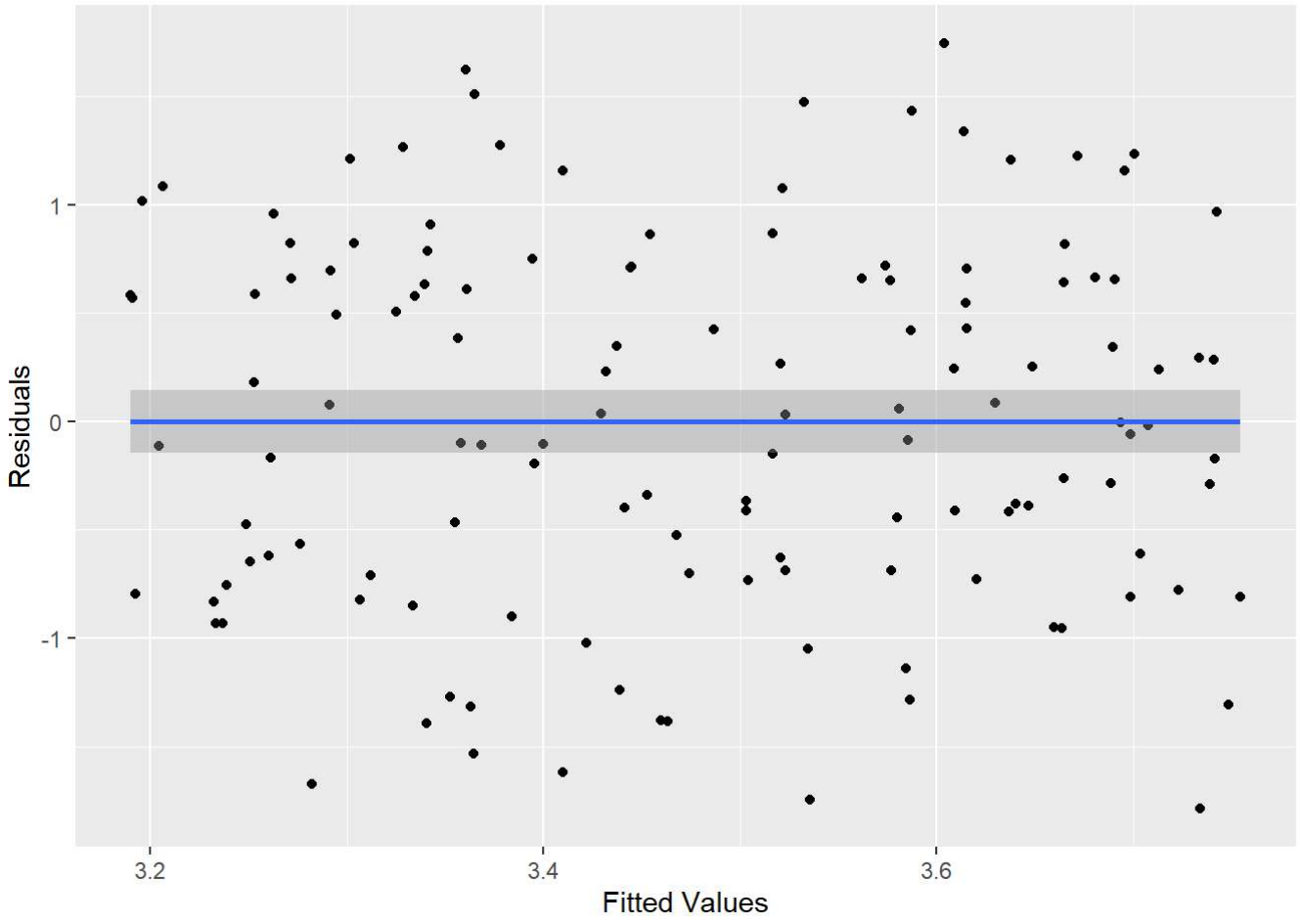
```r
fit4 <- lm(eNO ~ age, data = dataset2, na.action = na.exclude)

fit4 |>
  augment(dataset2) |>
  ggplot(aes(.fitted, .resid)) +
  geom_point() +
  geom_smooth(method = "gam") +
  labs(x = "Fitted Values", y = "Residuals")
```

```
`geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'

Warning: Removed 17 rows containing non-finite outside the scale range
(`stat_smooth()`).

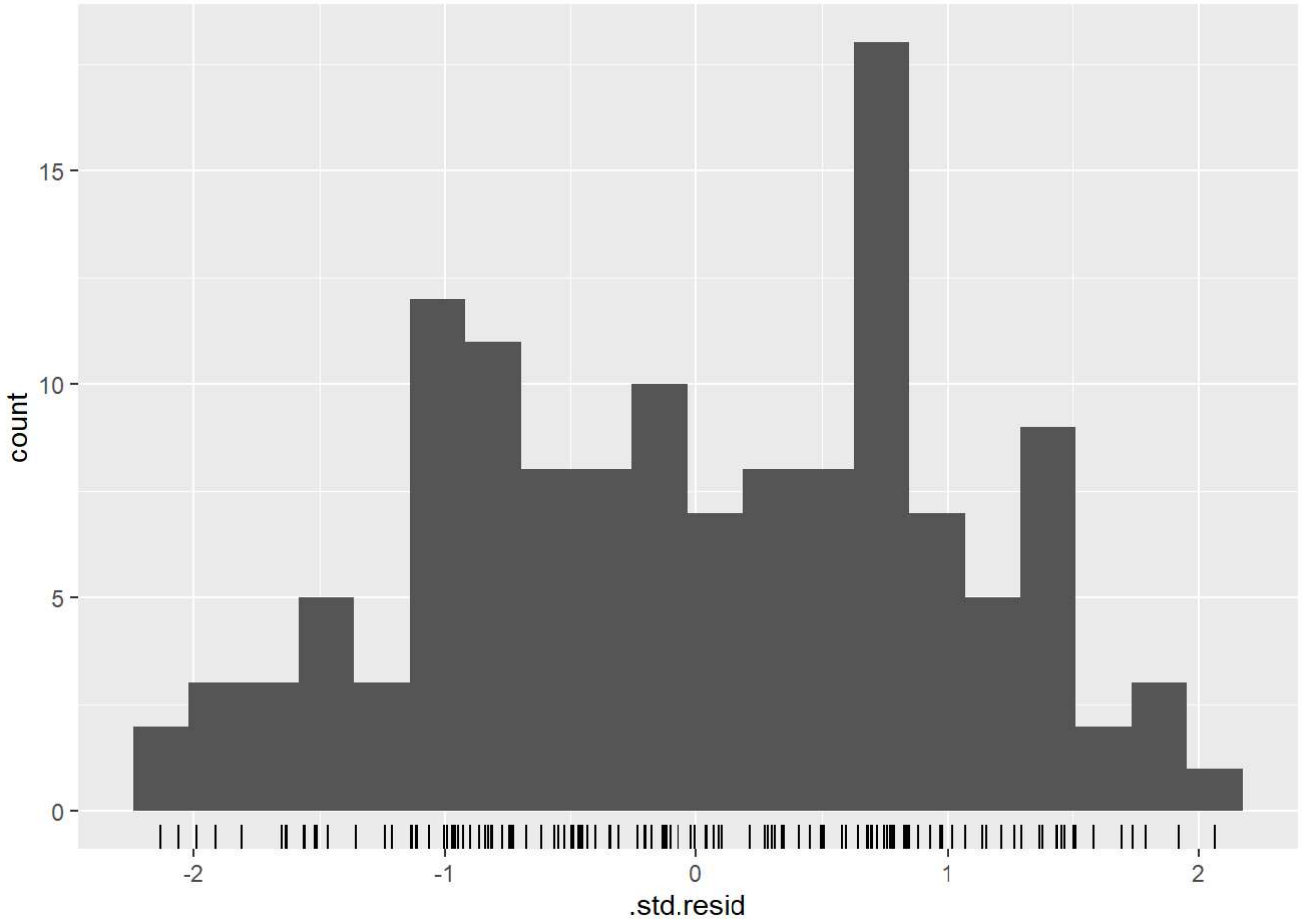
Warning: Removed 17 rows containing missing values or values outside the scale range
(`geom_point()`).
```




```r
fit4 |>
  augment(dataset2) |>
  ggplot(aes(x= .std.resid)) +
  geom_histogram(bins = 20) +
  geom_rug()
```

```
Warning: Removed 17 rows containing non-finite outside the scale range
(`stat_bin()`).

Warning: Removed 17 rows containing missing values or values outside the scale range
(`geom_rug()`).
```



The fit of this model looks much more evenly distributed than the previous model that did not use the log-transformed eNO variable. Based on these plots, this model is preferable (using that log-transformed eNO variable).

Part 9

Create a new dataset that takes the original `dataset` object and transforms the `mouse` variable by taking the log transformation of `mouse`. Store this new dataset in an R object named `dataset3`.

For this part we will focus on the outcome `cough_days`. Based on what we know of allergic disease, children should only have a response to mouse allergen if they are in fact allergic to mouse allergen. Let's examine this hypothesis as follows:

1. Create a new dataset called `dataset3yes` that contains only those individuals in `dataset3` who have `mouse_allergic == "yes"`. Using `dataset3yes`, fit a linear model with `cough_days` as the outcome and `mouse` as the predictor.

2. Create a new dataset called `dataset3no` that contains only those individuals in `dataset3` who have `mouse_allergic == "no"`. Using `dataset3no`, fit a linear model with `cough_days` as the outcome and `mouse` as the predictor.

For each analysis, describe what is the relationship between mouse allergen and days with coughing or wheezing (`cough_days`). Interpret the coefficient for `mouse` in each of the models.

```
dataset3 <- dataset |>
  mutate(mouse = log(mouse))

dataset3yes <- dataset3 |>
  filter(mouse_allergic == "yes")
dataset3no <- dataset3 |>
  filter(mouse_allergic == "no")
fit5 <- lm(cough_days ~ mouse, data = dataset3yes, na.action = na.exclude)
fit6 <- lm(cough_days ~ mouse, data = dataset3no, na.action = na.exclude)
fit5
```

```
Call:
lm(formula = cough_days ~ mouse, data = dataset3yes, na.action = na.exclude)

Coefficients:
(Intercept)        mouse
     7.2358      -0.5084
```

```
fit6
```

```
Call:
lm(formula = cough_days ~ mouse, data = dataset3no, na.action = na.exclude)

Coefficients:
(Intercept)        mouse
     0.1743       0.5328
```

For those who are allergic to the mouse allergen, it seems that as mouse allergen increases, cough days decrease by 0.5084. For those who are not allergic to the mouse allergen, as mouse allergen increases, cough days increase by 0.5328.

Part 10

Make a scatterplot of `cough_days` (y-axis) and `mouse` (x-axis), with a separate panel for those who are not mouse allergic and those who are mouse allergic (you can use `facet_wrap()` for this) using `dataset3`. In each plot, overlay the linear regression line fit using `geom_smooth(method = "lm")`.

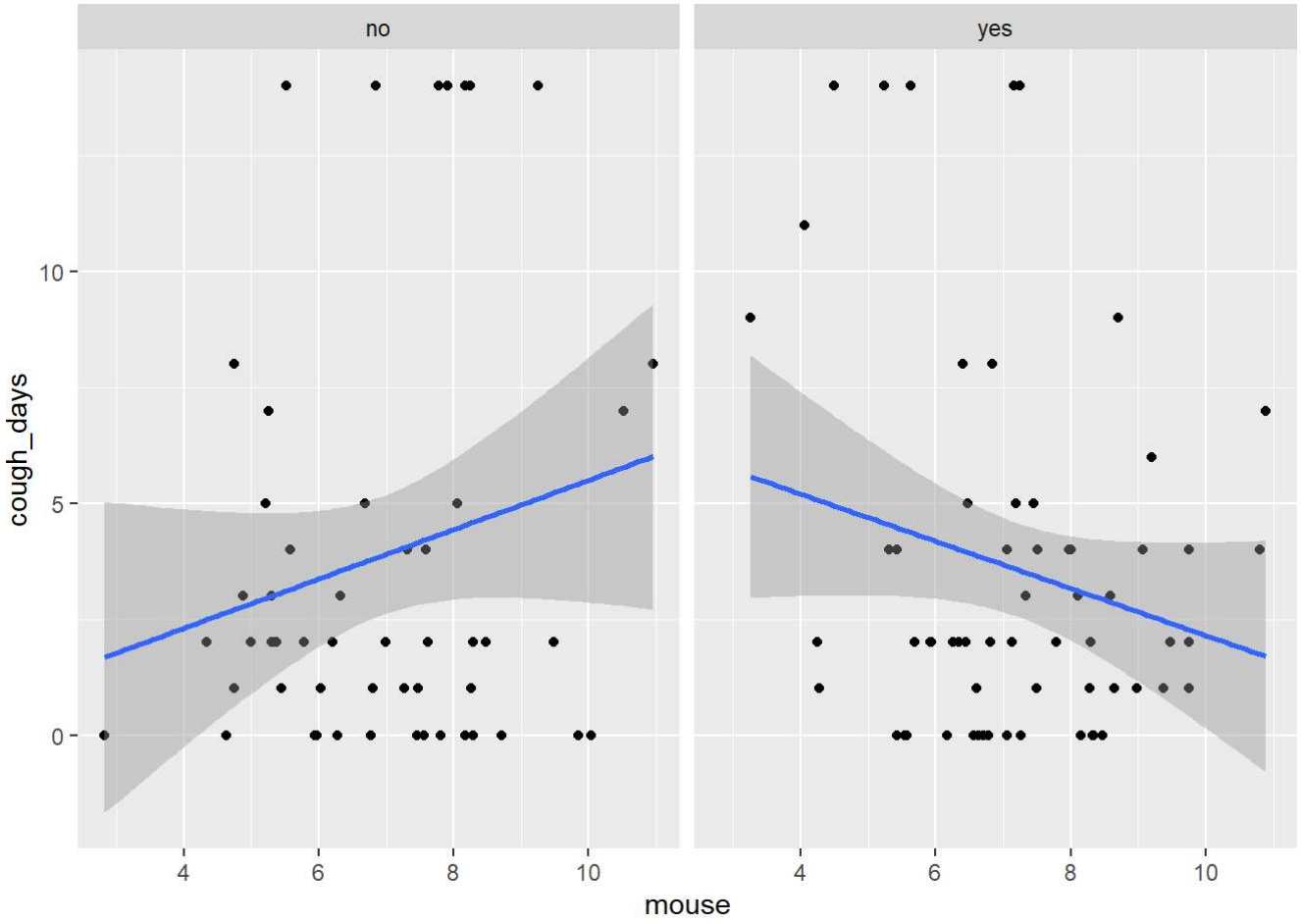
Are the relationships shown in the plots what you would have expected to see given what we know about allergic disease?

```
dataset3|>
  ggplot(aes(x = mouse, y = cough_days)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~mouse_allergic)
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning: Removed 34 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 34 rows containing missing values or values outside the scale range
(`geom_point()`).
```



The relationships shown in the plots are not what we would have expected to see given what we know about allergic disease, since we know that children should only have a response to the allergen if they are allergic to it. However, the plots show that some assumptions of linearity may be violated.

Part 11

Using the `dataset3` object, fit a linear regression model with FEV1 as the outcome and age, gender, PM2.5, NO2, mouse allergen, and mouse allergic status as the predictors in the model. Compute 95% confidence intervals for each of the coefficients in the model.

If we were going to collect data on a much larger sample of children, which of the three indoor pollutants—PM2.5, NO2, or mouse allergen—would most likely have a relationship similar to the one that we observe in this dataset?

```
fit7 <- lm(fev1 ~ age + gender + pm25 + no2 + mouse + mouse_allergic, data = dataset3, na.action =
confint(fit7)
```

```
                        2.5 %       97.5 %
(Intercept)       -1.3144046484 -0.057499300
age                0.1683834095  0.232172723
gendermale        -0.0539616775  0.405841405
pm25              -0.0030981190  0.005754459
no2                0.0008807409  0.007746905
mouse             -0.0268822732  0.100273453
mouse_allergicyes -0.5032778879 -0.054187667
```

If we were going to collect data on a much larger sample of children, PM2.5 would most likely have a relationship similar to the one that we observe in this dataset since the slope is around 0.

Part 12

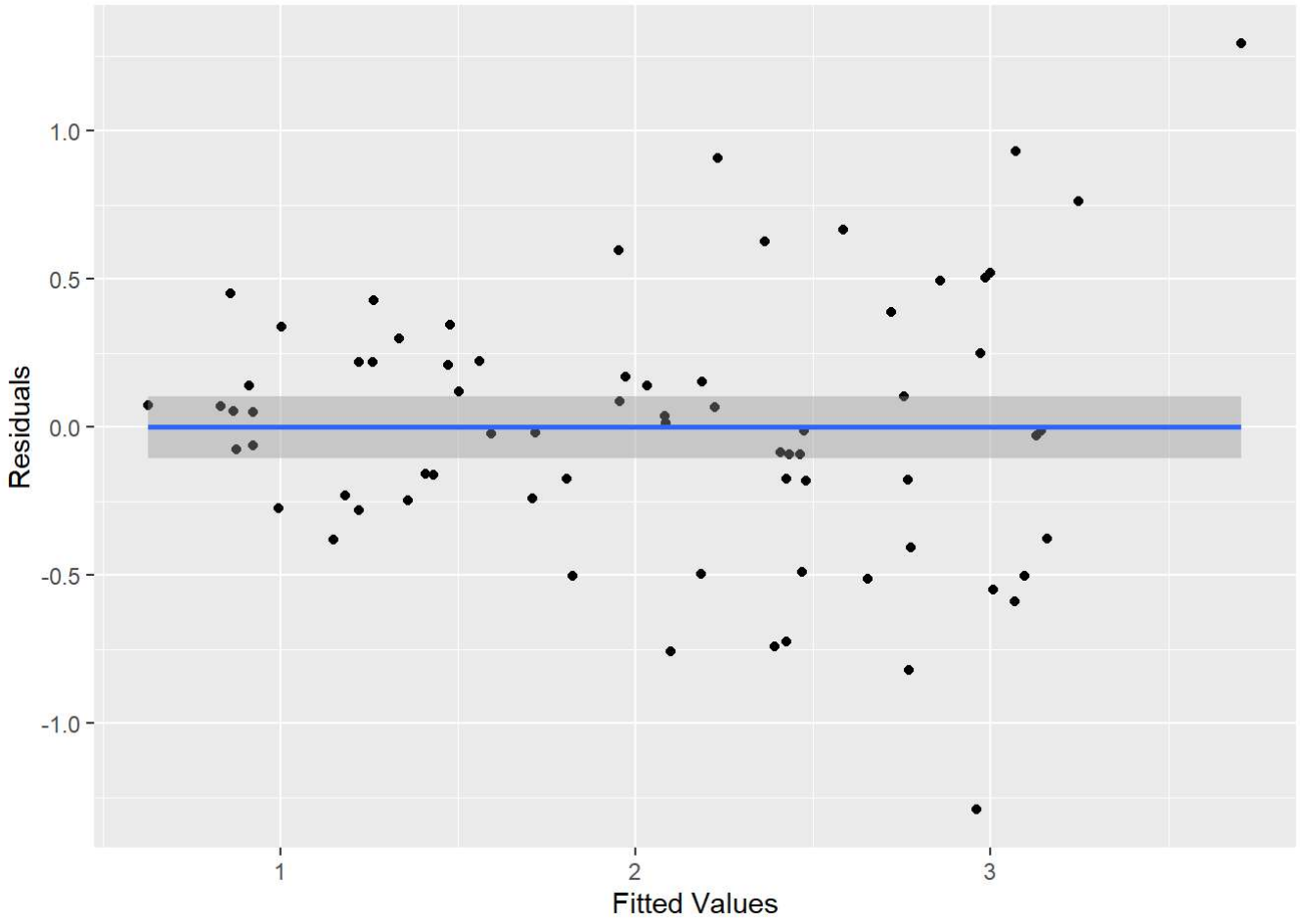
Make a scatterplot of the residuals versus the fitted values from this model. Describe what you see in the residuals and if there are any violations of the linear regression model assumptions evident in the plot.

```
fit7 |>
  augment(dataset3) |>
  ggplot(aes(.fitted, .resid)) +
  geom_point() +
  geom_smooth(method = "gam") +
  labs(x = "Fitted Values", y = "Residuals")
```

```
`geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'

Warning: Removed 79 rows containing non-finite outside the scale range
(`stat_smooth()`).

Warning: Removed 79 rows containing missing values or values outside the scale range
(`geom_point()`).
```



The residuals do appear to be linear, as they are equally distributed around the zero line. However, they make a "megaphone" shape as the fitted values increase (their spread around the zero line increases). So, the assumption of homoscedasticity appears to be violated.

Part 13

Identify the observations with the two largest residuals in absolute value. How would you describe these two children and why do you think they have such large residuals from the model?

```
fit7 |>
  augment(dataset3) |>
  filter(abs(.resid) > 1)
```

```
# A tibble: 2 × 16
  id          age gender  fev1   eNO cough_days  pm25   no2 mouse mouse_allergic
  <chr>     <dbl> <chr>  <dbl> <dbl>      <int> <dbl> <dbl> <dbl> <chr>
1 5186e2b6  16.6 male     5       38         14  39.3 124.   8.17 no
2 342d6b03  17.6 female   1.67   111         14  28.5  21.7  7.24 yes
# ℹ 6 more variables: .fitted <dbl>, .resid <dbl>, .hat <dbl>, .sigma <dbl>,
#   .cooksd <dbl>, .std.resid <dbl>
```

Both children displayed 14 cough days (which is every day in the observation period). This could mean that the children were affected from a source other than the indoor pollutants being measured.

Part 14

Suppose you had the ability to improve the fit of the model by collecting data on a new variable. What data would you most like to collect to improve this model? Why?

I would most like to collect data on time spent in their indoor home environment because if we are attempting to draw conclusions based on the indoor pollutants at home, it is important that we first draw the conclusion that increased time spent in the home has a positive correlation with cough days. For example, if the child is rarely at home, a high amount of mouse allergen would not imply causation of a higher number of coughing days compared to a child who sees an increase of coughing days with an increase of mouse allergen when they are home often.