Homework 4

Denis Ostroushko

2022-10-11

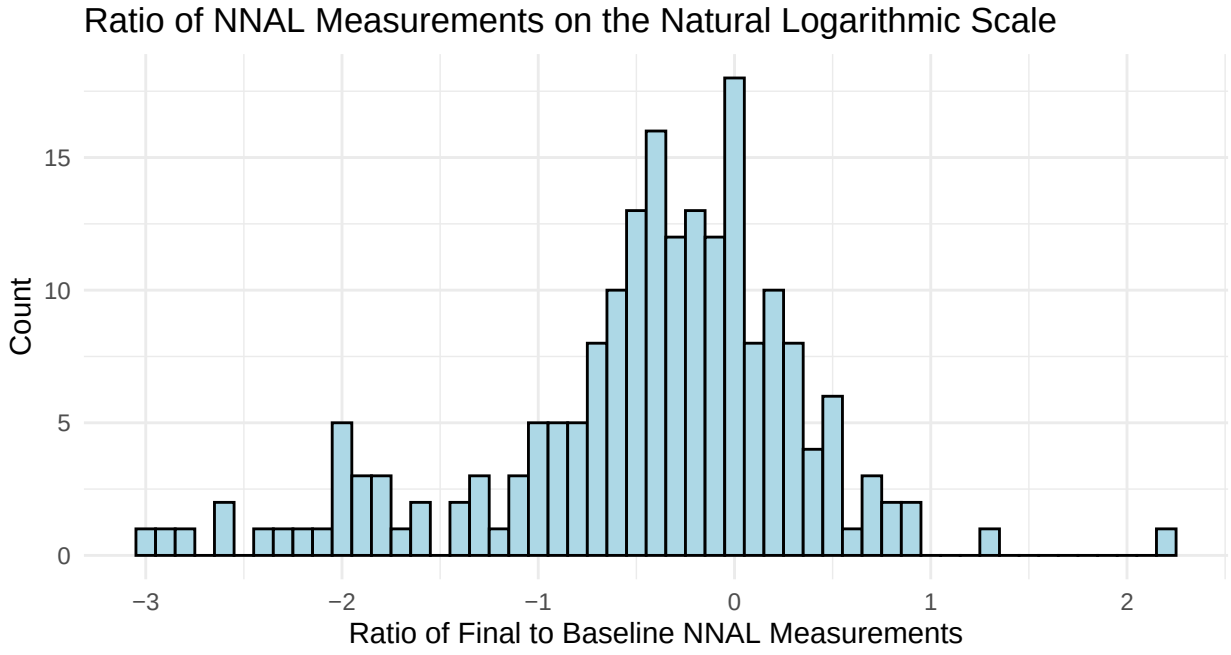
```
library(tidyverse)
library(kableExtra)
library(readxl)
```

8.4

8.4 - A

First, let's look at the distribution of the calculated response variable, it is a good practice to do so going forward for model development and diagnostics purpose.

```
ggplot(data = e_cig_3,
       aes(x = Y1)) +
  geom_histogram(binwidth = .1, color = "black", fill = "light blue") +
  theme_minimal() +
  ylab("Count") +  xlab("Ratio of Final to Baseline NNAL Measurements") + ggtitle("Ratio of NNAL Measur
```



Coefficients and other statistics from the multiple regression model are given in the table below.

```
e_cig_3_model_data <-
  e_cig_3 %>% select(arm, age, gender, white, educ2, income30, FTND, Y1)

model_8.4 <- lm(Y1 ~ ., data = e_cig_3_model_data)
model_8.4_res <- summary(model_8.4)
model_8.4_res_df <- data.frame(model_8.4_res$coefficients)
model_8.4_res_df$var <- rownames(model_8.4_res_df)
rownames(model_8.4_res_df) <- NULL
model_8.4_res_df <- model_8.4_res_df %>% select(var, everything())
model_8.4_res_df <-
  model_8.4_res_df %>% mutate_at(vars(Estimate, `Std..Error`, t.value, `Pr...t..`),
                                 funs(round(., 3)
                                      )
                                 )
colnames(model_8.4_res_df) <- c("Predictor", "Estiamte", "Standard Error", "T Value", "P value")

model_8.4_res_df %>%
  kbl(booktabs = T, align = c('l','c', 'c', 'c', 'c')) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Predictor	Estiamte	Standard Error	T Value	P value
(Intercept)	-0.093	0.477	-0.196	0.845
arm	-0.016	0.057	-0.285	0.776
age	-0.005	0.004	-1.068	0.287
gender2	-0.100	0.116	-0.863	0.389
white1	-0.113	0.124	-0.913	0.363
educ22	-0.068	0.119	-0.567	0.571
income302	-0.250	0.129	-1.944	0.053
FTND	0.060	0.045	1.323	0.187

Comments:

- None of the variables appear to be statistically significantly related to the response, after adjusting for other variables, at the 5% level.
- However, p-value for the income variable is suggestive that there might be some relationship going on, which we potentially can uncover either with a better model or with more data. Income summary is given below:

```
sum_income <-
  e_cig_3 %>%
    group_by(income30) %>%
    dplyr::summarise(
      n = n(),
      mean = mean(Y1),
      median = median(Y1)
    )


sum_income$income30 <- c("<= $30K/Yr.", "> $30K/Yr.")

colnames(sum_income) <- c("Income Levels", "N", "Average Response", "Median Response")
```

Income Levels	N	Average Response	Median Response
<= $30K/Yr.	135	-0.3531677	-0.2473906
> $30K/Yr.	60	-0.6406918	-0.4237410

```
sum_income %>%
  kbl(align = 'c', booktabs = T) %>%
  kable_styling(latex_options = 'striped')
```

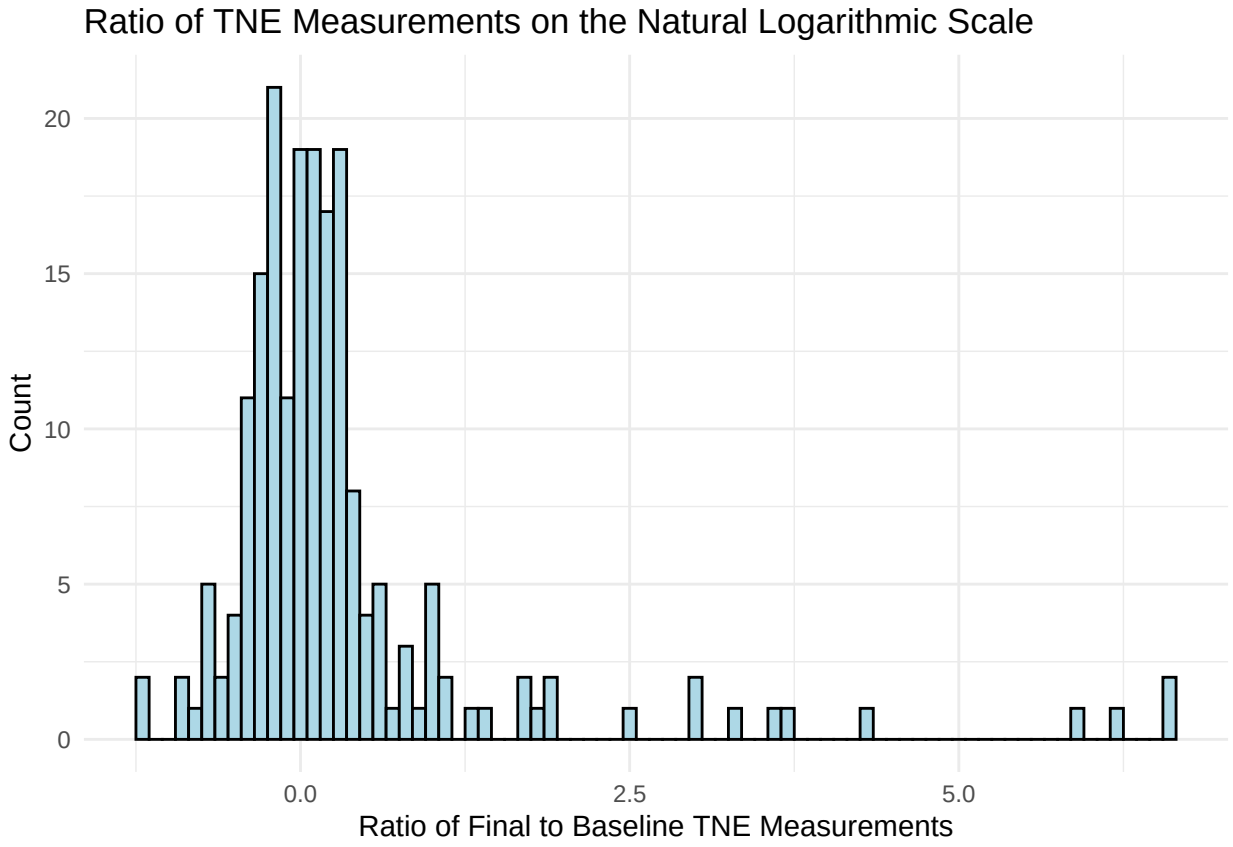
- While the average response appears to be quite different between the two groups, other variables in the multiple linear model might have an effect on this relationship.

8.4 - B

The distribution of the response variable below is highly skewed, so, perhaps, we should expect an even more poor fit of the model, and less statistically significant number of predictors.

```
#response
e_cig_3$Y2 <-log( e_cig_3$TNE_vt0_creat /e_cig_3$TNE_vt4_creat )

ggplot(data = e_cig_3,
       aes(x = Y2)) +
  geom_histogram(binwidth = .1, color = "black", fill = "light blue") +
  theme_minimal() +
  ylab("Count") +
  xlab("Ratio of Final to Baseline TNE Measurements") +
  ggtitle("Ratio of TNE Measurements on the Natural Logarithmic Scale")
```



```
e_cig_3_model_data <-
  e_cig_3 %>% select(arm, age, gender, white, educ2, income30, FTND, Y2)
model_8.4 <- lm(Y2 ~ ., data = e_cig_3_model_data)
model_8.4_res <- summary(model_8.4)
model_8.4_res_df <- data.frame(model_8.4_res$coefficients)
model_8.4_res_df$var <- rownames(model_8.4_res_df)
rownames(model_8.4_res_df) <- NULL
model_8.4_res_df <- model_8.4_res_df %>% select(var, everything())
model_8.4_res_df <-
  model_8.4_res_df %>% mutate_at(vars(Estimate, `Std..Error`, t.value, `Pr...t..`),
                                 funs(round(., 3)
                                      )
                                 )
colnames(model_8.4_res_df) <- c("Predictor", "Estiamte", "Standard Error", "T Value", "P value")
model_8.4_res_df %>%
  kbl(booktabs = T, align = c('l','c', 'c', 'c', 'c')) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Predictor	Estiamte	Standard Error	T Value	P value
(Intercept)	0.433	0.706	0.613	0.541
arm	-0.041	0.085	-0.481	0.631
age	0.003	0.007	0.490	0.625
gender2	0.084	0.172	0.492	0.624
white1	0.101	0.183	0.551	0.582
educ22	0.206	0.177	1.168	0.244
income302	0.216	0.190	1.138	0.257
FTND	-0.074	0.067	-1.114	0.267

- None of the variables here are close to being statistically significant
- Therefore, none of the predictors help us explain the variance of the biomarker change over time.

9.3

```
data_9.3 <-
data.frame(
  x = c(
    24,
    28,
    32,
    36,
    40,
    44,
    48,
    52,
    56,
    60
    ),
  y = c(
    38.8,
    39.5,
```

x	y
24	38.8
28	39.5
32	40.3
36	40.7
40	41.0
44	41.1
48	41.4
52	41.6
56	41.8
60	41.9

```
    40.3,
    40.7,
    41.0,
    41.1,
    41.4,
    41.6,
    41.8,
    41.9
    )
  )

data_9.3 %>% kbl() %>%
  kable_styling(latex_options = c("striped"))
```

```
res1 <- t(data_9.3$y) %*% data_9.3$y

res2 <- t(data_9.3$x) %*% data_9.3$y

res3 <- t(data_9.3$x) %*% data_9.3$x
```

- $Y`Y = res1 = 16663.85$
- $X`Y = res2 = 17245.6$
- $X`X = res3 = 18960$