Homework 4

Denis Ostroushko

2022-10-17

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

8.4

In this problem we will look at two additive multiple linear models. We will fit the models, and look for variables that are statistically associated with the calculated response variables.

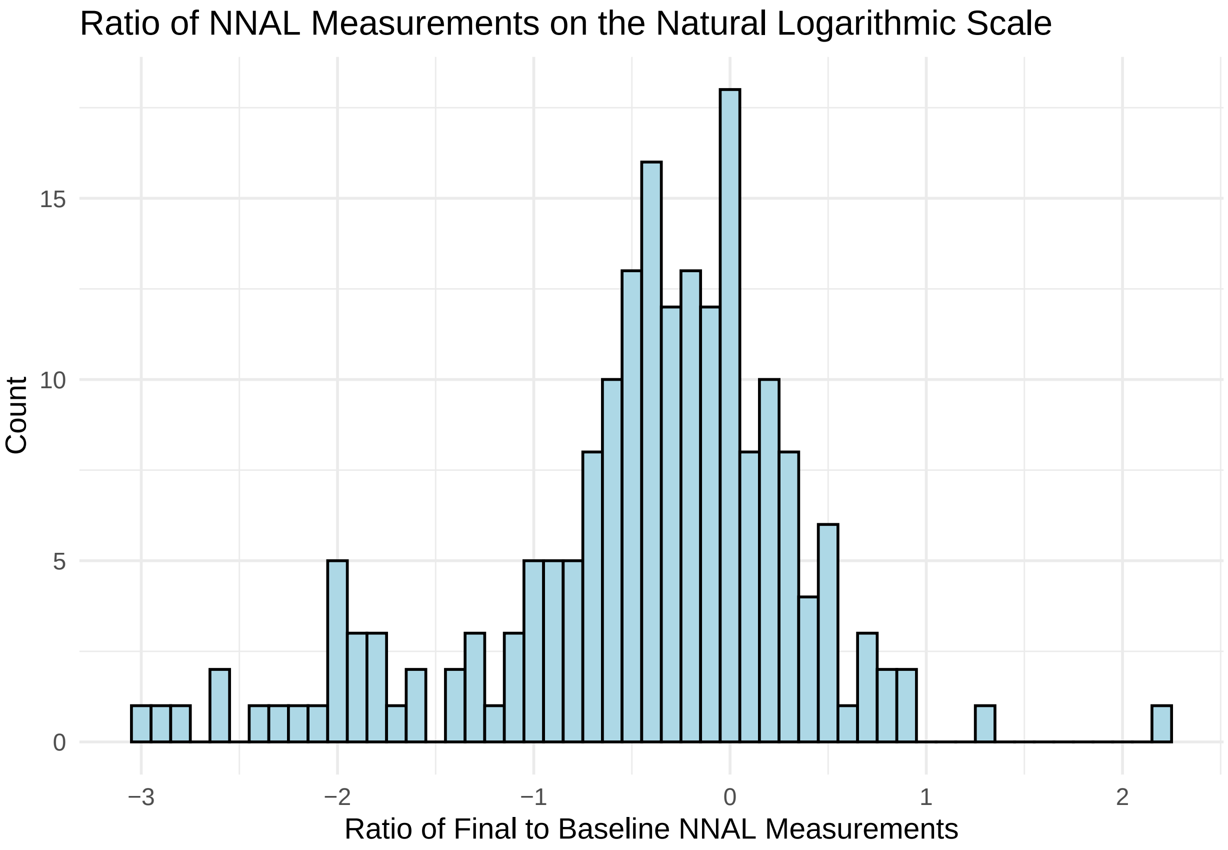
We can fit the model, get the summary and look at p-values that come out from a t-test for each $\hat{\beta}_i$.

However, to stay aligned with PUBH 7405 material, we will perform an ANOVA test for the model first, and then look at the individual t-test. Additionally, we will implement a Bonferroni, Holm, and Hockberg adjustments.

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 Measurements on the Natural Logarithmic Scale")
```



ANOVA Test for all predictors We will conduct a One-Way ANOVA test here to see how good the model is at explaining variation in the response variable.

For learning purposes, we will fit the models using built-in functions, and calculate the F statistic by hand.

In R, we need to fit a model with no predictors, i.e. the one that just predicts/fits the average value of the response for all observations in the data set. We then compare a model with more predictors to see if all coefficients are equal to zero, or not.

In the code chunk below we obtain the following estimates that we need to calculate F statistic:

- we obtain MSR and MSE from SSR and SSE respectively. Residuals and Fitted Values come from fitted model using an R function
- degrees on freedom in the numerator is the degrees of freedom of MSR, which is the number of predictors minus one
- degrees of freedom in the denominator is the degrees of freedom of MSE, which is the number of observations in the sample minus the number of predictors plus one

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

model_8.4 <- lm(Y1 ~ ., data = e_cig_3_model_data)

df_msr <- length(e_cig_3_model_data) - 1
df_mse <- nrow(e_cig_3_model_data) - length(e_cig_3_model_data)

MSR <- sum((mean(e_cig_3_model_data$Y1) - model_8.4$fitted.values)^2)/ # this is SSR: (fitted - mean)^2
            (df_msr) # this is DF = p - 1
```

```
MSE <- sum((e_cig_3_model_data$Y1 - model_8.4$fitted.values)^2)/ # this is SSE: (fitted - observed)^2
            (df_mse) # this is DF = n - p

F_stat <- MSR/MSE


F_star <- qf(1-.05/2, df1 = df_msr, df2 = df_mse)

P_F_star <- 1 - pf(F_stat, df1 = df_msr, df2 = df_mse)
```

Now we can conduct a test and see if all predictors are 0 or not. Test hypothesis and results are:

- Null Hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1} = 0$
- Alternative hypothesis: not all β_i are 0
- F- statistic: 1.62
- Cutoff F^* statistic: 2.48 with $df(MSR) = 6$ and $df(MSE) = 188$
- So, $F < F^*$, therefore, we do not reject the null hypothesis and we do not have enough evidence to conclude that coefficient estimates β_i are statistically different from 0.
- Additionally, $P(F^* > F) = 0.143$, which is kind of close to 0.05. So, when we do individual t-tests for each estimate β_i we will see that some of those coefficients are somewhat close to being significant, but we do not have enough data or a good enough model fit to detect any evidence that a given predictor is statistically related to the response variable

We can also check our work with the built in R function. We need to fit an "empty" model, that predicts/fits an average value of Y for each observation. Of course, we could also do it by hand.

```
empty <- lm(Y1 ~ 1, data = e_cig_3_model_data)

anova_res <- data.frame(anova(empty, model_8.4))

anova_res$model <- c("Empty Model", "Extended Model")

anova_res <- anova_res %>% dplyr::select(model, everything())

colnames(anova_res)[1] <- c("Model")

anova_res %>%
  kbl(booktabs = T, align = 'c', centering = T) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Model	Res.Df	RSS	Df	Sum.of.Sq	F	Pr..F.
Empty Model	194	125.3168	NA	NA	NA	NA
Extended Model	188	119.1532	6	6.163615	1.620827	0.1434007

As we can see, estimated calculated "by hand" align with the built in functions.

Summary of all coeffcients Coefficients and other statistics from the multiple regression model are given in the table below.

```
model_8.4_res <- summary(model_8.4)
model_8.4_res_df <- data.frame(model_8.4_res$coefficients)
```

```r
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)	-0.206	0.268	-0.766	0.445
age	-0.005	0.004	-1.082	0.281
gender2	-0.101	0.116	-0.870	0.385
white1	-0.113	0.123	-0.916	0.361
educ22	-0.064	0.118	-0.540	0.590
income302	-0.248	0.128	-1.936	0.054
FTND	0.061	0.045	1.347	0.180

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:

```r
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")

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.

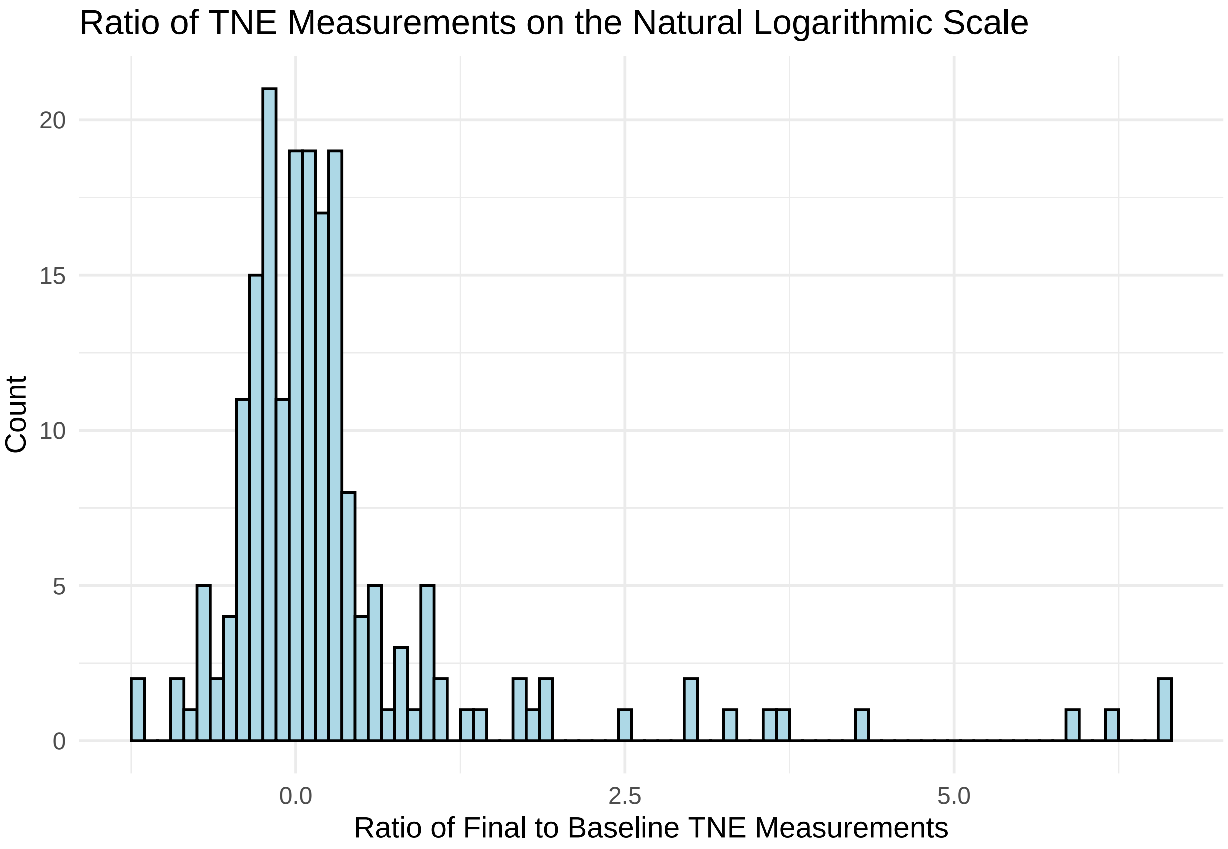
Income Levels	N	Average Response	Median Response
<= $30K/Yr.	135	-0.3531677	-0.2473906
> $30K/Yr.	60	-0.6406918	-0.4237410

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")
```



We begin this section again with the overall ANOVA test for the entire model. We will fit the full model and the empty model and conduct a built in ANOVA test.

```
e_cig_3_model_data <-
  e_cig_3 %>% select(age, gender, white, educ2, income30, FTND, Y2)

model_8.4 <- lm(Y2 ~ ., data = e_cig_3_model_data)
```

```r
empty <- lm(Y2 ~ 1, data = e_cig_3_model_data)

anova_res <- data.frame(anova(empty, model_8.4))

anova_res$model <- c("Empty Model", "Extended Model")

anova_res <- anova_res %>% dplyr::select(model, everything())

colnames(anova_res)[1] <- c("Model")

anova_res %>%
  kbl(booktabs = T, align = 'c', centering = T) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

Model	Res.Df	RSS	Df	Sum.of.Sq	F	Pr..F.
Empty Model	194	268.8892	NA	NA	NA	NA
Extended Model	188	261.0022	6	7.886923	0.9468256	0.4628066

Results of the one-way ANOVA are given in the table above. We will use these results to set up the test and interpret the results.

- Null Hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1}$
- Alternative Hypothesis: H_a : Not all coefficients β_i are zero
- $F-$statistic: 0.9468256
- Cutoff F^*-statistic: 2.1470705
- So, $F < F^*$, therefore we do not have enough evidence to reject the null hypothesis to conclude that some or all coefficients β_i are consistently different from zero.
- Moreover, $P(F^* > F) = 0.4628066$, which is quite different from zero. Therefore, when we evaluate a set of t-tests for each individual coefficient β_i we should not expect to see any predictors that are even close to being statistically significantly related to the response variable.

```
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)	0.152	0.397	0.384	0.702
age	0.003	0.006	0.473	0.637
gender2	0.083	0.171	0.484	0.629
white1	0.100	0.183	0.550	0.583
educ22	0.216	0.175	1.231	0.220
income302	0.222	0.189	1.169	0.244
FTND	-0.072	0.067	-1.089	0.278

Summary of all coefficients

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

data_9.3 %>% kbl(booktabs = T) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

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

```
data_9.3$int <- 1

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

res2 <- t(as.matrix(data_9.3 %>% dplyr::select(int, x))) %*% data_9.3$y

res3 <-  t(as.matrix(data_9.3 %>% dplyr::select(int, x))) %*%  as.matrix(data_9.3 %>% dplyr::select(int
```

- $Y`Y = res1 = 16663.85$
- In regression analysis X is a matrix of all predictors AND an additional column full of ones, which we add to the matrix in order to get the estimate for the intercept term. Thus, $X`$ is a $p \times n$ matrix (2 by 10 in our case) and $Y$ is a 10 by 1 vector. So, we obtain a 2 by 1 matrix $X`Y = res2 =$

```
##        [,1]
## int   408.1
## x   17245.6
```

- Matrix Explained:
  - The first entry in the matrix is the sum of all values of column(variable) Y in the data set, as presented above, and expressed as ΣY_i
  - The second variable is the linear combination of variables X and Y, we can represent it as $\Sigma X_i Y_i$
- Similar reasoning applies here. $X`X = res3 =$

```
##     int     x
## int  10   420
## x   420 18960
```

- Matrix Explained:

- 10 represents the N of the data set.
- Entries off the main diagonal represent the sum of the X variable, ΣX_i
- 18960 is the sum of squared X terms, ΣX_i^2