Homework 4

Victoria Yang

5/25/23

Link to github repository:

https://github.com/v-yc/ENVS-193DS_homework-04_Yang-Victoria

1. Setup

Load in required packages.

```
library(tidyverse)
library(here)
library(performance)
library(broom)
library(flextable)
library(ggeffects)
library(car)
library(naniar)
```

Read in the data and subset data of interest (length and weight of trout perch species).

```
# read in data
fish <- read_csv(here("data", "knb-lter-ntl.6.34", "ntl6_v12.csv")) %>%

  # only include troutperch species
  filter(spname == "TROUTPERCH") %>%

  # select the columns of interest
  select(length, weight)
```

2. Initial data visualization

In mathematical terms:

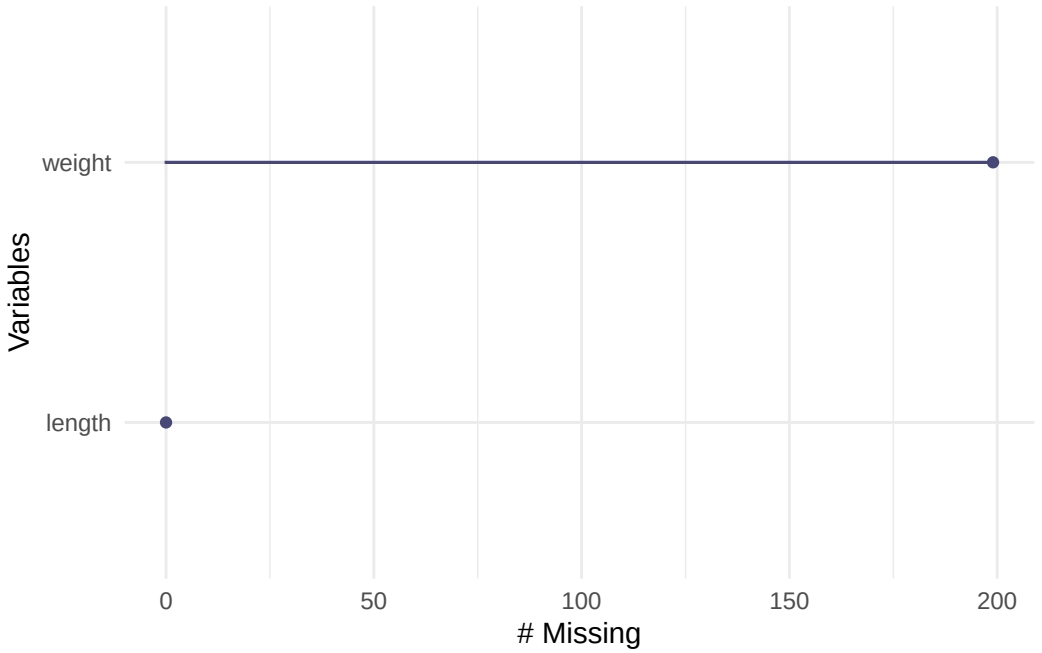
- The null hypothesis is that the slope of the linear model (where fish length is the predictor variable and fish weight is the response variable) is 0.
- The alternative hypothesis is that the slope of the linear model (where fish length is the predictor variable and fish weight is the response variable) is not 0.

In biological terms,

- The null hypothesis is that fish length is not a good predictor of fish weight for trout perch.
- The alternative hypothesis is that fish length is a good predictor of fish weight for trout perch.

Visualize the missing data:

```
gg_miss_var(fish)
```



There are about 200 missing data points for the weight variable, which is significant because there are only 489 observations in this dataset. This would mean that about 41% of the data is missing.

However, we do not know why these data points are missing because this is from an online data set. For our analysis, we will remove the NA values by subsetting the data.

```
fish_subset <- fish %>%

  # drops rows with NA values in the columns specified
  drop_na(length, weight)
```

3. Create a linear model and check assumptions

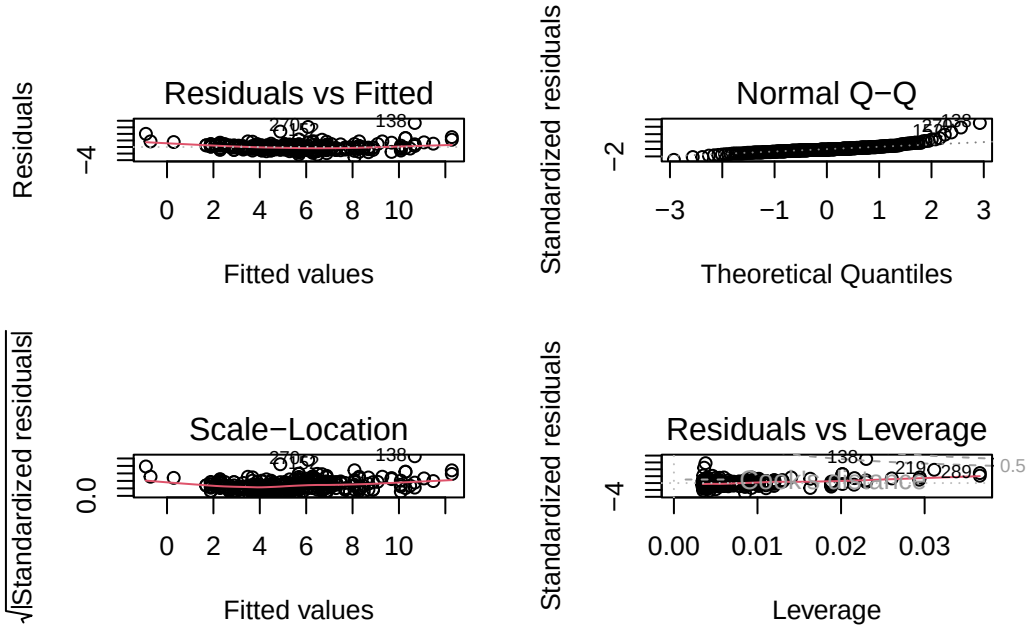
Creating a linear model where fish length is the predictor variable and fish weight is the response variable:

```
modelobject <- lm(weight ~ length, data = fish_subset)
```

Check assumptions by plotting residual diagnostic plots.

```
# display diagnostic plots in a 2x2 grid
par(mfrow = c(2, 2))

# plot diagnostic plots to check linear model assumptions
plot(modelobject)
```



Interpretation of diagnostic plots:

- Residuals vs fitted plot:
- Normal Q-Q plot:
- Scale-location plot:
- Residuals vs Leverage: