Homework #4

Graycen Mahon

Homework #4

```
# set your file path using the here function to organize the files needed and load in pack
here::set_here("/Users/graycenmahon/Downloads/ENVS 193DS/ENVS-193DS_homework-04_mahon-gray
```

```
File .here already exists in /Users/graycenmahon/Downloads/ENVS 193DS/ENVS-193DS_homework-04_
```

```
library(here)
```

```
here() starts at /Users/graycenmahon/Downloads/ENVS 193DS/ENVS-193DS_homework-04_mahon-grayce
```

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(naniar)
library(magrittr)
```

```
Attaching package: 'magrittr'

The following object is masked from 'package:purrr':

    set_names

The following object is masked from 'package:tidyr':

    extract
```

```
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```
library(dplyr)
library(ggplot2)
```

1) How does fish length predict fish weight for trout perch? (across all sample years)

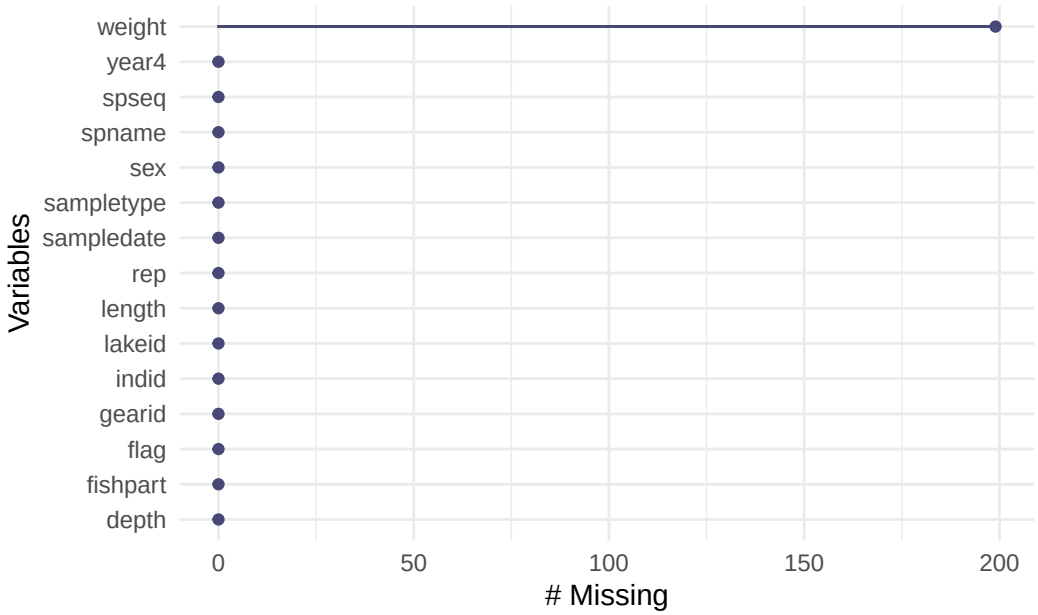
H0: There is no linear relationship between fish length and fish weight, B = 0 (B -> slope). Fish length does not predict fish weight in trout perch, there is no correlation between the two. HA: There is a linear relationship between fish length and fish weight, B =/ 0. Fish length and fish weight are positively correlated, as fish length increases, so does its length and vice versa.

```
# using the read.csv function read in the data set that will be used in the code
data <- read.csv("north_temperate_lakes.csv")
```

```
troutperch_data <- data %>%
# using the mutate and filter functions, adjust the column title and then filter the large
```

```
  mutate(spname = case_when(spname == "TROUTPERCH" ~ "trout_perch")) %>%
  filter(spname == "trout_perch")
```

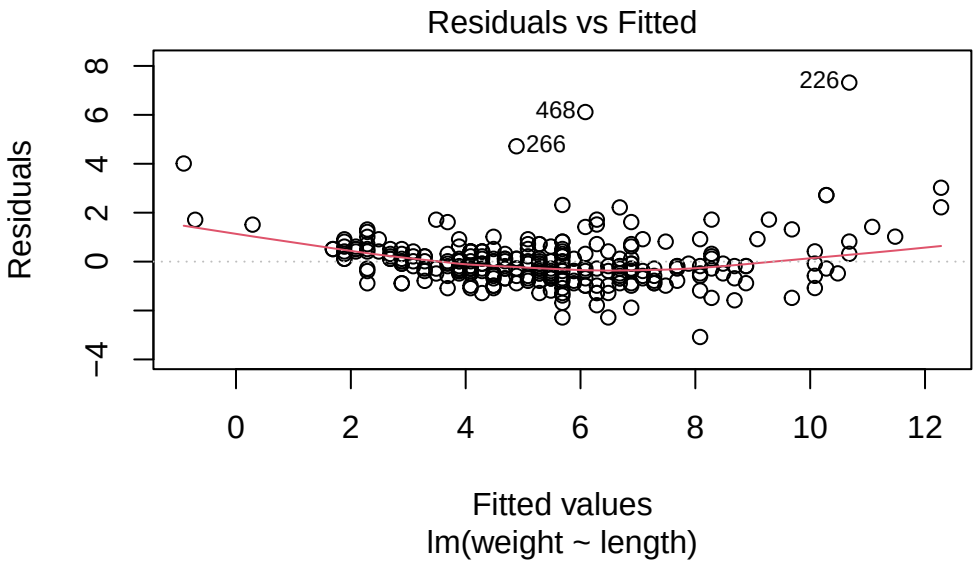
```
# using the gg missing data function, call on the new data to make a visualization showing
gg_miss_var(troutperch_data) + (element_text = labs(caption = "Visualizing the Missing Dat
```

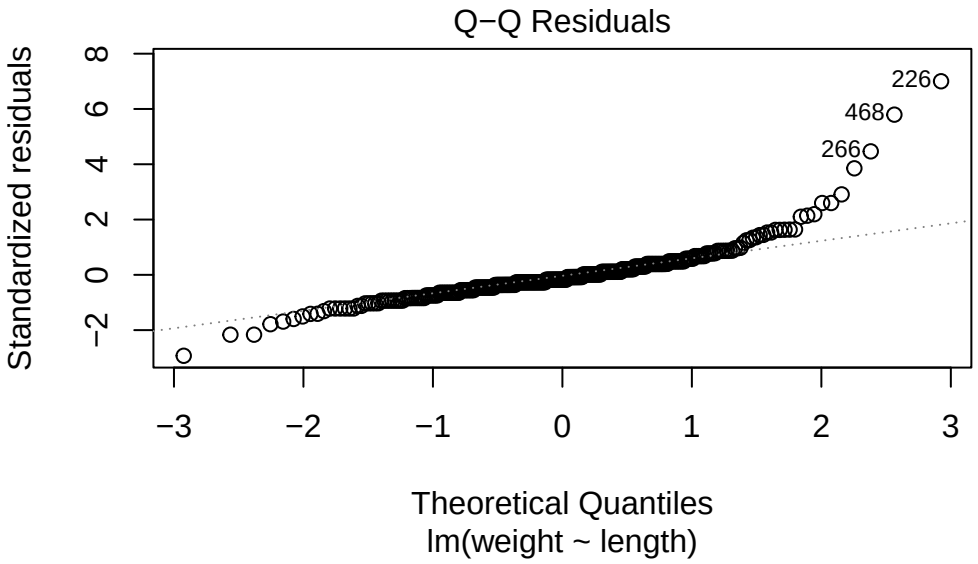


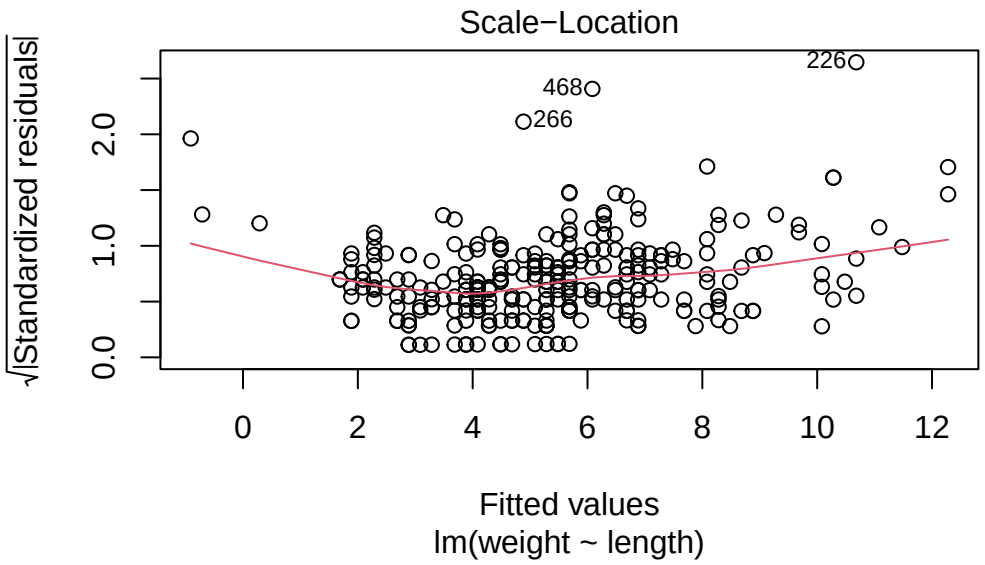
Visualizing the Missing Data from the North Temperate Lakes dataset

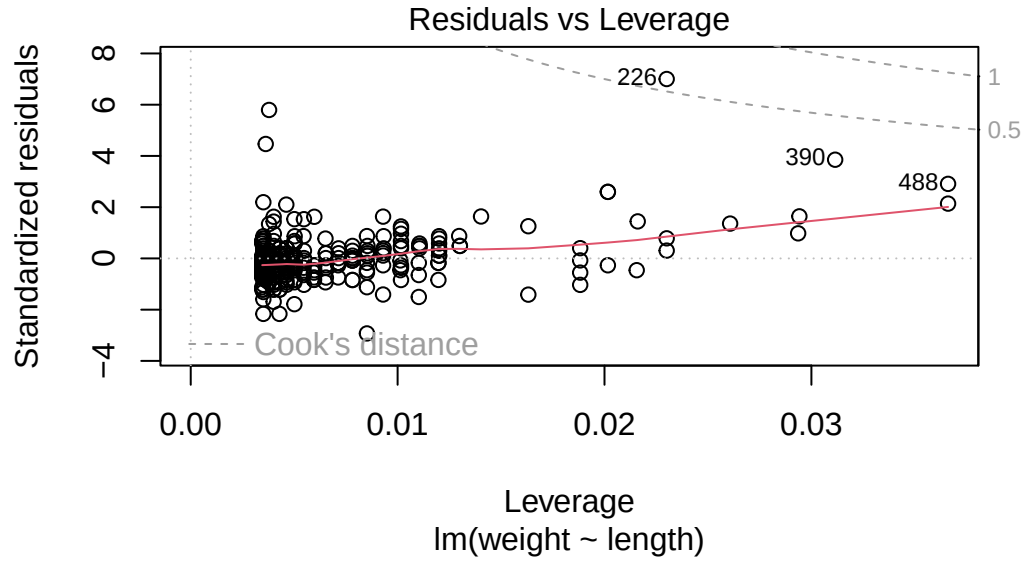
```
# here we can see that there are just under 200 missing values for the trout perch weight
```

```
# create a new object called model object and using the lm fucntion (linear model) gather
modelobject <- lm(weight ~ length, data = troutperch_data)
plot(modelobject)
```


Residuals vs Fitted
Residuals
Fitted values
lm(weight ~ length)
226
468
266


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(weight ~ length)
226
468
266


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
0
2
4
6
8
10
12
226
468
266
Fitted values
lm(weight ~ length)


Residuals vs Leverage
Standardized residuals
−4
0
2
4
6
8
0.00
0.01
0.02
0.03
226
390
488
1
0.5
Cook's distance
Leverage
lm(weight ~ length)

Checking my assumptions:

There are four assumptions of linear models:

1) There is a linear relationship between dependent and independent variables According to the first plot (residuals vs fitted) the pattern and distributions of points show a fairly linear relationship between the two variables. There is no pattern or clear trend amongst the points.

2) Independent errors (no correlation between errors) The second plot (Q-Q Residuals) shows that the data set follows a mostly normal distribution. Towards the end of the graph the data points veer off the dotted line, indicating a data set that is not entirely normally distributed.

3) Homoscedasticity (constant variance) of errors In the third plot (Scale-location), testing for the ratio of variance, we can see that due to the data points being pretty evenly and randomly distributed along the fairly straight red line, that the data set is indeed homeoscedastic.

4) Normally distributed errors Lastly, the final plot (Residuals vs Leverage) is a plot that shows which, and how many, points in a data set are considered influential. If any point in this plot falls outside of Cook's distance (the black dashed lines in the upper right hand corner) then it is considered to be an influential observation and if removed from the dataset, the coefficients would change drastically. This specifc plot shows that there are very few influential points in the dataset, with only 1 points falling outside the dashed lines.

```
# using the summary function, summarize the model object depicted above
summary_modelobject <- summary(modelobject)
```

```
# then use the anova function to make an anova table and summarize the data for the anova
anova_data <- anova(modelobject)
summary_anova_data <- summary.aov(modelobject)

# making the anova table with the as data frame function. I names both rows and columns an
anova_table <- as.data.frame(anova_data, row.names = c("Length", "Residuals"), stringsAsFa
colnames(anova_table) <- c("Degrees of Freedom", "Sum of Squares (SS)", "Mean of Squares (
```

8) Describing ANOVA Results in accordance with the Summary() Results

ANOVA is a form of linear regression, theoretically showing a linear relationship between all predictor (length) and response (weight) variables. The modelobject shows that the data is normally distributed, and with the Quartiles in the summary we can ssee that it is true; although slightly skewed left. Q1 and Q3 should be the same magnitude, but the Q1 is slightly more negative indicating a left skew.The F-ratio in the ANOVA table is the ratio is very high, and with a high t-value and low P-value in the modelobject summary, they reinforce the idea that there is a significant relationship between the variables.

9) According to the test run above, the model object, and the ANOVA table I can conclude that there is a significant relationsuip between the two variables: length and weight. A extremely low P-value (0.00000...2) and a high F-ratio support the idea that there is a relationship between fish length and their subsequent weight. Although the data is left skewed, it is still considered to have a normal distribution. The null hypothesis is rejected, as B is not equal to 0 (B = slope). This is proved by the linear regression model, where the first plot (residuals vs fitted) showed that the data had a postive linear relationship (denoting a positive slope!).

10) Create a Data Visualization with Model Predictions Confidence Intervals

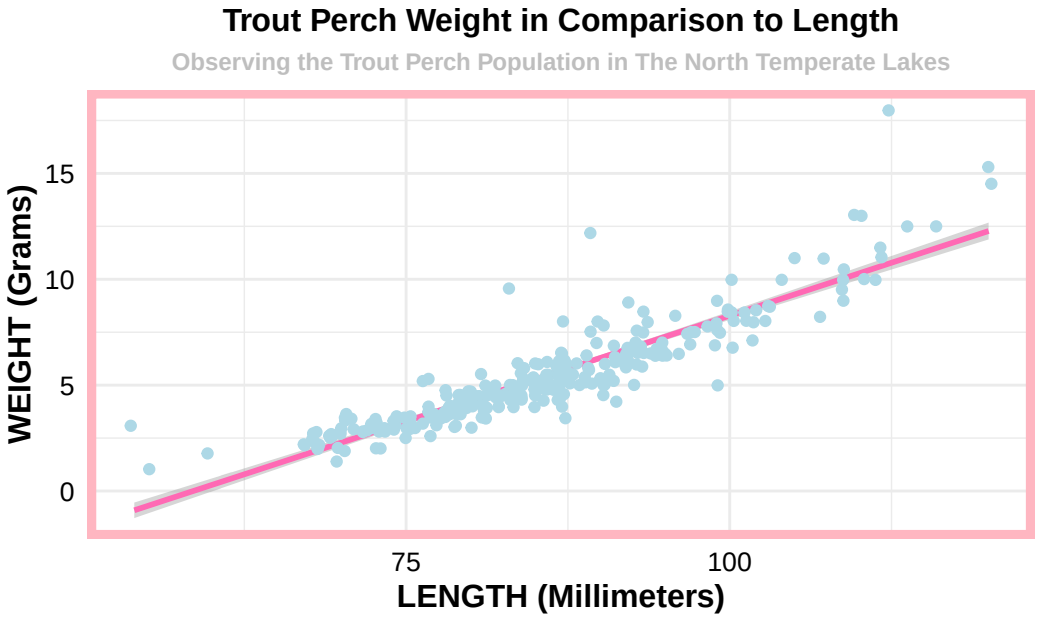
```
# calling in ggplot to make a graph!! Using the theme function I named and altered the app
ggplot(data = troutperch_data, aes(x = length, y = weight)) + geom_smooth(color = "hotpink
theme(axis.text = element_text(colour = "black", size=10),
        axis.title = element_text(size=12, face = "bold"),
        plot.title = element_text(color = "black", size = 12, face = "bold", hjust = 0.5),
  plot.subtitle = element_text(color = "gray", size = 9, face = "bold", hjust = 0.5),
  panel.border = element_rect(color = "lightpink", fill = NA, linewidth = 3),
  legend.text = element_text(color = "black", size = 7, face = "bold"),
  legend.position = c(0.85, 0.86)) +
```

```
    labs(x = "LENGTH (Millimeters)", y = "WEIGHT (Grams)", title = "Trout Perch Weight in Co
```

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

```
Warning: Removed 199 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 199 rows containing missing values (`geom_point()`).
```



of fish weights on fish length with the linear regression model predictions and confidence intervals