hw4

library(here)

here() starts at /Users/larry/github/hw4

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.1 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1

── 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(lterdatasampler)  
library(naniar)  
# would be nice to have  
library(performance)  
library(broom)  
library(flextable)

Attaching package: 'flextable'  
  
The following object is masked from 'package:purrr':  
  
 compose

library(ggeffects)  
library(car)

Loading required package: carData  
  
Attaching package: 'car'  
  
The following object is masked from 'package:dplyr':  
  
 recode  
  
The following object is masked from 'package:purrr':  
  
 some

fish\_size <- read\_csv(here("ntl6\_v12.csv"))

Warning: One or more parsing issues, call `problems()` on your data frame for details,  
e.g.:  
 dat <- vroom(...)  
 problems(dat)

Rows: 349229 Columns: 15  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (8): lakeid, gearid, spname, sampletype, indid, fishpart, spseq, flag  
dbl (5): year4, depth, rep, length, weight  
lgl (1): sex  
date (1): sampledate  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

1.

**Mathematical**

H0: β = 0

Ha: β ≠ 0

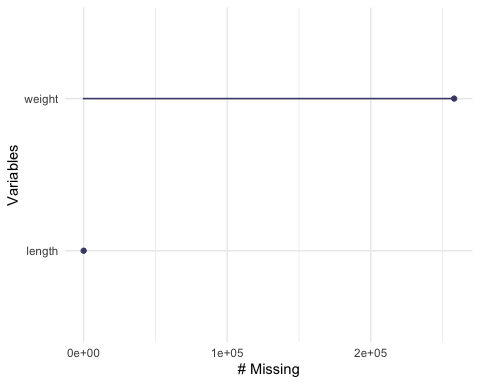
**Biological**

Null: Fish length does not predict fish weight for trout perches

Alternative: Fish length does predict fish weight for trout perches

2.

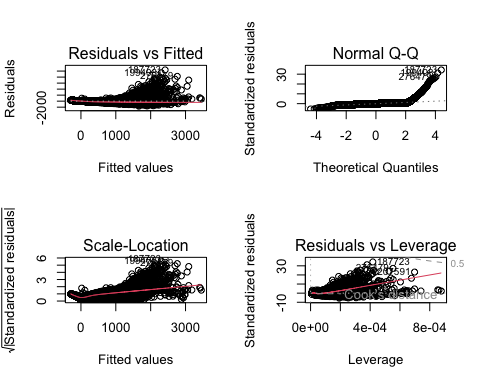
filtered\_fish\_size <- subset(fish\_size, select = c(length, weight))  
gg\_miss\_var(filtered\_fish\_size)



2a. The missing data

3 & 4.

linear\_fish <- lm(weight ~ length, data = filtered\_fish\_size)  
# makes the viewer pane show a 2x2 grid of plots  
par(mfrow = c(2, 2))  
plot(linear\_fish)



Residuals vs Fitted: Residuals are not evenly or randomly distributed around the horizontal dotted line, violating the assumption of homoscedasticity.

Normal Q-Q: The normal Q-Q shows residuals that are not normal because they do not follow a straight line (although they are rather straight until 2 on the theoretical quantile.

Scale-Location This plot essentially tells us the same thing as the residuals vs. fitted plot (except using the square root of the standardized residuals), which is that since the residuals are not evenly or randomly distributed around the horizontal dotted line, the assumption of homoscedasticity is violated.

Residuals vs Leverage This plot shows the influence outliers have on the model, and although the observations are within cook’s distance, some are labeled illustrating that if they were taken out they would vjange the estimates

options(scipen = 999)  
model\_summary <- summary(linear\_fish)  
model\_summary

Call:  
lm(formula = weight ~ length, data = filtered\_fish\_size)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1661.7 -142.1 -6.5 118.8 10385.8   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -309.55373 1.60785 -192.5 <0.0000000000000002 \*\*\*  
length 3.00965 0.00814 369.7 <0.0000000000000002 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 303 on 91088 degrees of freedom  
 (258139 observations deleted due to missingness)  
Multiple R-squared: 0.6001, Adjusted R-squared: 0.6001   
F-statistic: 1.367e+05 on 1 and 91088 DF, p-value: < 0.00000000000000022

model\_squares <- anova(linear\_fish)  
model\_squares

Analysis of Variance Table  
  
Response: weight  
 Df Sum Sq Mean Sq F value Pr(>F)   
length 1 12549293413 12549293413 136700 < 0.00000000000000022 \*\*\*  
Residuals 91088 8362023669 91802   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

model\_squares\_table <- tidy(model\_squares) %>%   
 # round the sum of squares and mean squares columns to have 5 digits (could be less)  
 mutate(across(sumsq:meansq, ~ round(.x, digits = 5))) %>%   
 # round the F-statistic to have 1 digit  
 mutate(statistic = round(statistic, digits = 1)) %>%   
 # replace the very very very small p value with < 0.001  
 mutate(p.value = case\_when(  
 p.value < 0.001 ~ "< 0.001"  
 )) %>%   
 # rename the stem\_length cell to be meaningful  
 mutate(term = case\_when(  
 term == "length" ~ "Fish Length",  
 TRUE ~ term  
 )) %>%   
 # make the data frame a flextable object  
 flextable() %>%   
 # change the header labels to be meaningful  
 set\_header\_labels(df = "Degrees of Freedom",   
 sumsq = "Sum of squares",  
 meansq = "Mean squares",  
 statistic = "F-statistic",  
 p.value = "p-value")  
  
model\_squares\_table

| term | Degrees of Freedom | Sum of squares | Mean squares | F-statistic | p-value |
| --- | --- | --- | --- | --- | --- |
| Fish Length | 1 | 12,549,293,413 | 12,549,293,413.06 | 136,700.2 | < 0.001 |
| Residuals | 91,088 | 8,362,023,669 | 91,801.59 |  |  |

1. We hypothesized that fish length predicted fish weight. Our null hypothesis was that fish length did not predict fish weight. Fish weight was significantly predicted by differing fish lengths (analysis of variance, F(1, 91088) = 136700.2, ⍺ = 0.05, p < 0.001).

predictions <- ggpredict(linear\_fish, terms = "length")  
  
predictions

# Predicted values of weight  
  
length | Predicted | 95% CI  
---------------------------------------  
 0 | -309.55 | [-312.71, -306.40]  
 200 | 292.38 | [ 290.28, 294.47]  
 300 | 593.34 | [ 590.29, 596.39]  
 500 | 1195.27 | [1189.41, 1201.13]  
 700 | 1797.20 | [1788.27, 1806.13]  
 800 | 2098.16 | [2087.68, 2108.65]  
 1000 | 2700.09 | [2686.46, 2713.73]  
 1300 | 3602.99 | [3584.60, 3621.37]

#caption was to large, so had to put up here instead of in 'labs'  
caption\_text <- "This plot is illustrating the predicted values of fish weight (assuming a certain fish length) within a 95% confidence level. The blue line shows the trend line for the correlation between the length and weight."  
wrapped\_caption <- str\_wrap(caption\_text, width = 100)  
  
  
plot\_predictions <- ggplot(data = filtered\_fish\_size,   
 aes(x = length, y = weight)) +  
 # first plot the underlying data from maples\_data  
 geom\_point() +  
 # then plot the predictions  
 geom\_line(data = predictions,   
 aes(x = x, y = predicted),   
 color = "blue", linewidth = 1) +  
 # then plot the 95% confidence interval from ggpredict  
 geom\_ribbon(data = predictions,   
 aes(x = x, y = predicted, ymin = conf.low, ymax = conf.high),   
 alpha = .2) +  
 # theme and meaningful labels  
 theme\_bw() +  
 labs(x = "Fish length",  
 y = "Fish weight")  
  
plot\_predictions +labs(caption = wrapped\_caption)

Warning: Removed 258139 rows containing missing values (`geom\_point()`).

