hw4

library(here)  
library(tidyverse)  
library(lterdatasampler)  
library(naniar)  
# would be nice to have  
library(performance)  
library(broom)  
library(flextable)  
library(ggeffects)  
library(car)  
library(ggplot2)  
#read in fish data csv  
fish\_data <- read\_csv(here("data", "ntl6\_v12.csv"))

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

1.

**Mathematical**

H0: β = 0

Ha: β ≠ 0

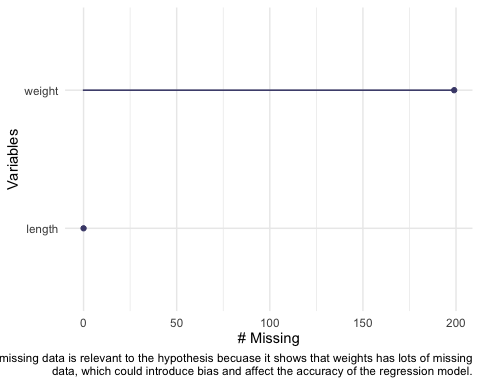
**Biological**

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

Alternative: Fish length does predict fish weight for trout perches

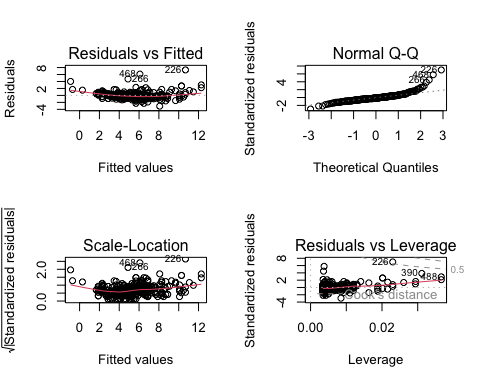
2 & 2a.

#filter fish data to only include the columns length and weight  
filtered\_fish <- fish\_data %>%  
 filter(spname == "TROUTPERCH") %>%  
 select(length, weight)  
#caption was to large, so had to put up here instead of in 'labs'  
caption\_text <- "The missing data is relevant to the hypothesis becuase it shows that weights has lots of missing data, which could introduce bias and affect the accuracy of the regression model."  
wrapped\_caption <- str\_wrap(caption\_text, width = 100)  
#shows missing data (specifically in weights, where there is over 250,000 missing observations)  
gg\_miss\_var(filtered\_fish)+labs(caption = wrapped\_caption)



3 & 4.

#linear regression of the filtered fish data, were length is the predictor and weight is the response  
linear\_fish <- lm(weight ~ length, data = filtered\_fish)  
# makes the viewer pane show a 2x2 grid of plots  
par(mfrow = c(2, 2))  
plot(linear\_fish)



Residuals vs Fitted: Residuals are seemingly evenly and randomly distributed around the horizontal dotted line, confirming 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 evenly and randomly distributed around the horizontal dotted line, the assumption of homoscedasticity is confirmed.

Residuals vs Leverage This plot shows the influence outliers have on the model, and although nearly all the observations are within cook’s distance (except 226), some are labeled illustrating that if they were taken out they would change the estimates.

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

Call:  
lm(formula = weight ~ length, data = filtered\_fish)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.0828 -0.4862 -0.1830 0.4128 7.3191   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -11.702476 0.481564 -24.30 <0.0000000000000002 \*\*\*  
length 0.199852 0.005584 35.79 <0.0000000000000002 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.057 on 288 degrees of freedom  
 (199 observations deleted due to missingness)  
Multiple R-squared: 0.8164, Adjusted R-squared: 0.8158   
F-statistic: 1281 on 1 and 288 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 1432.29 1432.29 1280.8 < 0.00000000000000022 \*\*\*  
Residuals 288 322.05 1.12   
---  
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 small p value with < 0.001  
 mutate(p.value = case\_when(  
 p.value < 0.001 ~ "< 0.001"  
 )) %>%   
 # rename the row names 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 | 1,432.2877 | 1,432.28769 | 1,280.8 | < 0.001 |
| Residuals | 288 | 322.0525 | 1.11824 |  |  |

1. The Anova table relates to the summary object because both say the f statistic, degrees of freedom, and p-value. However, the summary object provides more information (such as standard errors) and other statistics for interpreting the data.
2. We hypothesized that fish length predicted fish weight. Our null hypothesis was that fish length did not predict fish weight. Our results provided evidence to reject the null hypothesis, as fish weight was significantly predicted by differing fish lengths (analysis of variance, F(1, 91088) = 136700.2, ⍺ = 0.05, p < 0.001).

10.

# extract model predictions using ggpredict  
predictions <- ggpredict(linear\_fish, terms = "length")  
  
predictions

# Predicted values of weight  
  
length | Predicted | 95% CI  
-----------------------------------  
 50 | -1.71 | [-2.12, -1.30]  
 60 | 0.29 | [-0.02, 0.59]  
 65 | 1.29 | [ 1.03, 1.54]  
 75 | 3.29 | [ 3.12, 3.45]  
 85 | 5.28 | [ 5.16, 5.41]  
 95 | 7.28 | [ 7.12, 7.44]  
 105 | 9.28 | [ 9.04, 9.53]  
 120 | 12.28 | [11.88, 12.68]

#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,   
 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),   
 # makes the confidence interval less opaque and more transparent  
 alpha = .2) +  
 # theme and meaningful labels  
 theme\_bw() +  
 labs(x = "Fish length",  
 y = "Fish weight")  
  
plot\_predictions +labs(caption = wrapped\_caption)

Warning: Removed 199 rows containing missing values (`geom\_point()`).

