homework-04-code

## Quarto

Link to repo can be found [here](https://github.com/bianca-berron/ENVS-193DS_homework-04_berron-bianca).

# load in packages  
library(tidyverse)  
library(here)  
library(performance)  
library(broom)  
library(flextable)  
library(ggeffects)  
library(car)  
library(naniar)  
library(lterdatasampler)  
library(ggplot2)

# Problem 1. “How does fish length *predict* fish weight for trout perch (across all sample years)?”

1. Biological: Null Hypothesis: Fish length does not predict fish weight for trout perch.

Alternative Hypothesis: Fish length does predict fish weight for trout perch.

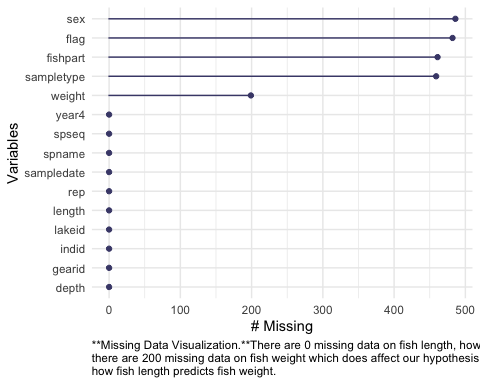
Mathematical: Null hypothesis: β\_1 = 0

Alternative hypothesis: β\_1 ≠ 0

# load in data and save as "fish\_data" object  
fish\_data <- read\_csv(here("data", "ntl6\_v12.csv"))

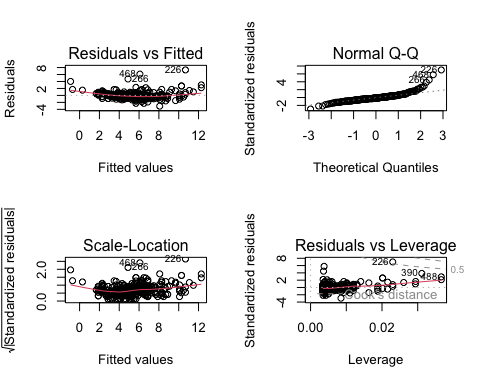
# filter the data to only include trout perch species  
fish\_data\_filtered <- fish\_data %>%   
 filter(spname == "TROUTPERCH")

# create a missing data visualization using gg\_miss\_var  
gg\_miss\_var(fish\_data\_filtered) +  
 # add a meaningful caption   
 labs (caption = str\_wrap("\*\*Missing Data Visualization.\*\*There are 0 missing data on fish length, however there are 200 missing data on fish weight which does affect our hypothesis in how fish length predicts fish weight.")) +  
 theme(  
 plot.caption = element\_text(hjust = 0)  
 )



3 and 4)

# create modelobject to determine if there is a relationship between fish length and weight  
modelobject <- lm(weight ~ length, data = fish\_data\_filtered)   
  
# makes the viewer pane show a 2x2 grid of plots  
par(mfrow = c(2, 2))  
plot(modelobject)



1. Check assumptions:

Residual vs fitted plot: plots the residuals (difference between observed and expected values) versus the fitted line. Based on the how the residuals seem to be somewhat evenly and randomly distributed around the horizontal dotted line, one can confirm homoscedasticity. In addition, the red line is relatively flat.

Normal Q-Q plot: looking at the qqplot, the residuals look normally distributed, as the residuals follow a linear path.

Scale-Location plot: plots fitted values versus the square root of the standardized residuals; plots seem to be somewhat evenly and randomly distributed around horizontal (and relatively flat) red line, implying homoscedasticity.

Residuals vs leverage plot: there doesn’t appear to be any significant residual outliers that are affecting the model predictions.

# store model summary as an object  
model\_summary <- summary(modelobject)  
  
model\_summary

Call:  
lm(formula = weight ~ length, data = fish\_data\_filtered)  
  
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 <2e-16 \*\*\*  
length 0.199852 0.005584 35.79 <2e-16 \*\*\*  
---  
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: < 2.2e-16

# store the ANOVA table as an object and use anova () function to retrieve variance table analysis for model  
model\_squares <- anova(modelobject)  
  
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 < 2.2e-16 \*\*\*  
Residuals 288 322.05 1.12   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

model\_squares\_table <- tidy(model\_squares) %>%   
 # replace the extremely small p values with < 0.001  
 mutate(p.value = case\_when(  
 p.value < 0.001 ~ "< 0.001"  
 )) %>%   
 # rename the "length" cell to be meaningful as "Trout Perch Fish Length (mm)"   
 mutate(term = case\_when(  
 term == "length" ~ "Trout Perch Fish Length (mm)", TRUE ~ term  
 )) %>%   
 # make 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 |
| --- | --- | --- | --- | --- | --- |
| Trout Perch Fish Length (mm) | 1 | 1,432.2877 | 1,432.287687 | 1,280.844 | < 0.001 |
| Residuals | 288 | 322.0525 | 1.118238 |  |  |

1. The ANOVA table directly relates to the information received from the summary() object. Both have a calculated F-statistic of about 1281 on 1 and 288 degrees of freedom, a p-value < 0.001, and the summary () object has an R^2 value of 0.8164.
2. We tested the hypothesis of whether or not fish length predicts fish weight for trout perch in which the null hypothesis was that fish length does not predict fish weight for trout perch. Based on the samples and data, we found that fish length does predict fish weight for trout perch based on our sample size of 489 and a significance level of 0.05 (F\_1,288 = 1280.8, p < 0.001, R^2 = 0.82). Based on our observations, for every 1.0 mm increase in fish length, we expect a 0.2 g increase in fish weight.

# conduct model predictions using ggpredict  
predictions <- ggpredict(modelobject, 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]

plot\_predictions <- ggplot(data = fish\_data\_filtered, aes(x = length, y = weight)) +  
 # plot the underlying data from fish\_data\_filtered  
 geom\_point() +  
 # plot the predictions  
 geom\_line(data = predictions, aes(x = x, y = predicted), color = "hotpink", linewidth = 1) +  
 # plot 95% confidence interval using ggpredict  
 geom\_ribbon(data = predictions, aes(x = x, y = predicted, ymin = conf.low, ymax = conf.high), alpha = 0.2) +  
 # add a theme, meaningful labels, and a caption  
 theme\_bw() +  
 labs (x = "Fish Length (mm)", y = "Fish Weight (g)",   
 caption = str\_wrap("\*\*Figure 1.\*\* Based on our observations, for every 1.0 mm increase in fish length, we expect a 0.2 g increase in fish weight. Pink line represents predicted fish weight based on fish length while black points represent the observed fish weights. The shaded gray area around red line represents the confidence level of the predicted values.")) +  
 theme(  
 # change axis titles font and boldness  
 axis.title = (element\_text(color = "black", family = "Times New Roman", face = "bold")),  
 axis.text = element\_text(family = "Times New Roman", color = "black"),  
 # adjust caption  
 plot.caption = element\_text(family = "Times New Roman", hjust = 0))  
  
plot\_predictions

