Homework 4

Stephanie Nguyen

2023-05-24

Table of contents

Github Link: https://github.com/stephanie-t-nguyen/ENVS-193DS\_homework-04\_Nguyen\_Stephanie.git

### Setup

library(tidyverse)  
library(here)  
library(naniar)  
#install.packages("ggeffects")  
library(ggeffects)  
#install.packages("flextable")  
library(flextable)  
library(broom)

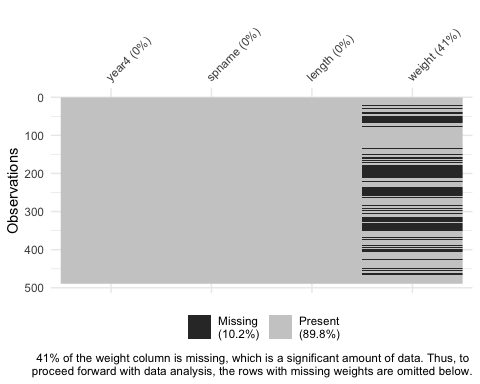
### How does fish length predict fish weight for trout perch (across all sample years)?

### Part 1 - Hypotheses

1. The null hypothesis is that there is no linear relationship between the length and weight of the trout perch (the p-value is greater than 0.05). The alternative hypothesis is that the fish length could be used to predict the fish weight in a linear model (the two variables have a linear relationship, the p-value is less than 0.05). In biological terms, this means that the null hypothesis states that fish weight is unrelated to fish length (there can be heavy, short fish and light, long fish). The alternative hypothesis would state that fish weight is correlated with fish length – longer fish should have a heavier weight because they are longer.

### Part 2 - Missing Data

#Load in and filter data set  
fish\_data <- read.csv("~/Documents/Rprojects\_ENVS193/github/ENVS-193DS\_homework-04\_Nguyen\_Stephanie/data/ntl6\_v12.csv") |> #Read in data  
 filter(grepl("TROUTPERCH", spname)) |> #Filter rows containing trout perch  
 select(c("year4", "spname", "length", "weight")) #Select relevant columns  
  
#Problem 1b - create visualization of missing data   
missing\_fish\_data <- vis\_miss(fish\_data) + labs(caption = "41% of the weight column is missing, which is a significant amount of data. Thus, to   
 proceed forward with data analysis, the rows with missing weights are omitted below.")  
missing\_fish\_data

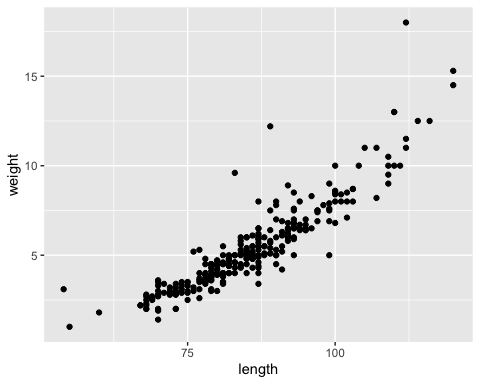


#41% of the weight data is missing. This is relevant because one of the variables being tested in the linear regression model is weight.

### Part 3 - Run tests

#Address missing data  
fish\_data <- fish\_data |> #Fix fish\_data   
 drop\_na(weight) #Drop rows with NAs in weight

#Exploratory data visualization, looks like there is relationship between length and weight, but we need to further investigate   
ggplot(data = fish\_data, aes(x = length, y = weight)) + geom\_point()

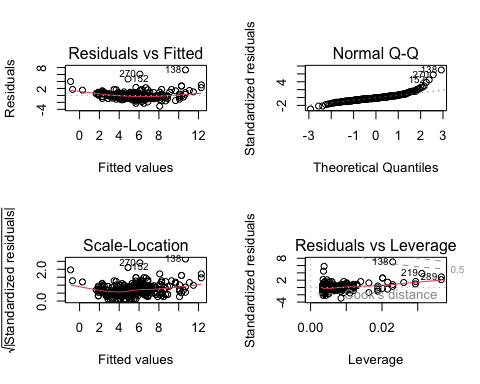


### Part 4 - Diagnostic Plots

#Create model to further explore relationship  
modelobject <- lm(weight ~ length, data = fish\_data)  
modelobject

Call:  
lm(formula = weight ~ length, data = fish\_data)  
  
Coefficients:  
(Intercept) length   
 -11.7025 0.1999

#Visualize models - diagnostic plots in a grid   
par(mfrow = c(2, 2))  
plot(modelobject)



### Part 5 - Meaning of diagnostic plots

The residuals vs fitted line nicely fits the data, with the exception of some very few outliers, suggesting that a linear relationship with a slight bow fits the data well. This means we could proceed with our analysis and that the p-values and confidence intervals are reliable for this test.

The Normal Q-Q plot demonstrates the normality of the data. The vast majority of the data fits on a linear plot suggesting normality, however, the tails at the end suggest some outliers.

The scale location plot shows us if the data is homoscedastic or heteroscedastic, based on the spread of points around a roughly red horizontal line. Since has no difference in spread throughout the red line, we can assume heteroscedasticity, one of the requirements for a linear regression model.

The residuals vs leverage plot shows us that there is one point (#138) that is outside Cook’s distance. This indicates an outlier that will significantly change the regression model if removed, so we must either verify the data point is not a mistake, or remove the data point altogether to better fit the linear regression model.

### Part 6 - Results of Model Object

#Display results of modelobject  
model\_summary <- summary(modelobject)  
model\_summary

Call:  
lm(formula = weight ~ length, data = fish\_data)  
  
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  
Multiple R-squared: 0.8164, Adjusted R-squared: 0.8158   
F-statistic: 1281 on 1 and 288 DF, p-value: < 2.2e-16

### Part 7 - ANOVA Table

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

#Create ANOVA table   
model\_squares\_table <- tidy(model\_squares) |> #Make dataframe with model\_squares  
 mutate(across(sumsq:meansq, ~ round(.x, digits = 5))) |> #round sum of squares and mean squares columns to have 5 digits  
 mutate(statistic = round(statistic, digits = 1)) |> #round F-statistic to have 1 digit  
 mutate(p.value = case\_when(  
 p.value < 0.001 ~ "0.001"  
 )) |> #Replace p-value with <0.001 if it is very small   
 mutate(term == case\_when( #Rename length column  
 term == "length" ~ "length (mm)",  
 TRUE ~ term  
 )) |>   
 flextable() |> #Make the data frame a flex table object  
 set\_header\_labels(df = "Degrees of Freedom", #Change header labels to be meaningful  
 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 | term == ... |
| --- | --- | --- | --- | --- | --- | --- |
| length | 1 | 1,432.2877 | 1,432.28769 | 1,280.8 | 0.001 | FALSE |
| Residuals | 288 | 322.0525 | 1.11824 |  |  | TRUE |

### Part 8

The ANOVA table lays out the overall model and significance of the linear regression model (F-statistic, p-value, sum and mean squares), while the summary() lays out individual predictors (adjusted R square values, standard error, etc). The ANOVA table still pulls much of its values from the summary() but calculates new values such as sum of squares and mean squares.

### Part 9

We ran an ANOVA test and a linear regression model in order to see how well length predicts weight. When running summary() and the ANOVA table, the degrees of freedom was 1 and 288, the p-value is 2e-16, the t-value for length was 35.79, the F-statistic is 1208.8, the significance level was 0.05, the R squared value was 0.81, and the sample size was 290. The p-value was much, much less than 0.001 meaning that our calculated F-statistic is very unlikely to have obtained through chance alone and we can reject our null hypothesis, meaning length is likely a good predictor of fish weight.

2-3 sentences, results

### Part 10

#Get predicted model values 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  
plot\_predictions <- ggplot(data = fish\_data, aes(x = length, y = weight)) +   
 geom\_point() + #Make scatterplot of weight and length  
 geom\_line(data = predictions, #Graph model (predicted values)  
 aes(x = x, y = predicted),   
 color = "red", linewidth = 1) +   
 geom\_ribbon(data = predictions,#Illustrate confidence interval  
 aes(x = x, y = predicted, ymin = conf.low, ymax = conf.high), alpha = 0.2) +   
 theme\_bw() + #Set theme   
 labs(x = "Length (mm)", #Set labels   
 y = "Weight (g)",   
 caption = "This is a linear regression model in which the red lines are the predicted values and the gray ribbon   
around it represents the confidence interval. According to our stats above and the plot, length predicts   
weight well. For every 20 mm longer the fish, its weight is expected to increase by about 5 grams.",   
 title = "Linear Regression Model of Trout-Perch Length vs Weight") +   
 theme\_bw() +   
 theme(plot.caption = element\_text(hjust = 0))  
  
plot\_predictions

