Homework 4

Katie Miller

2023-05-25

# Set Up

# loading in the necessary packages  
library(tidyverse)  
library(here)  
library(naniar)  
library(dplyr)  
library(performance)  
library(broom)  
library(flextable)  
library(ggeffects)  
library(car)

# reading in the data using here package  
fish <- read\_csv(here("data", "ntl6\_v12.csv"))

# filtering data frame to extract relevant info  
trout <- fish %>%   
 filter(spname %in% "TROUTPERCH") %>%   
 reframe(year = case\_when(spname == "TROUTPERCH" ~ year4),  
 species = case\_when(spname == "TROUTPERCH" ~ spname),  
 length = case\_when(spname == "TROUTPERCH" ~ length),  
 weight = case\_when(spname == "TROUTPERCH" ~ weight))

# Problem 1

### 1.

**Biological:**

Null: The length of a trout perch is not an accurate predictor of its weight.

Alternative: The length of a trout perch accurately predicts its weight.

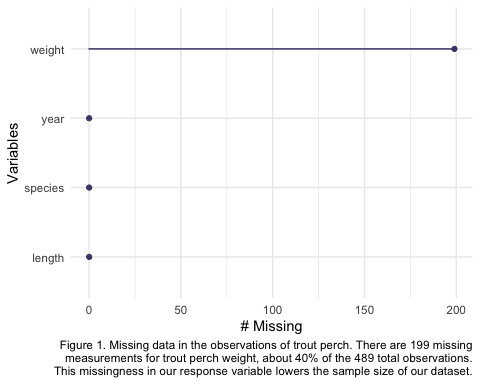
**Mathematical:**

Null: the predictor variable does not predict the response variable.

Alternative: the predictor variable predicts the response variable.

### 2.

# visualizing missing data in the filtered perch trout dataset  
gg\_miss\_var(trout) +   
 labs(caption = "Figure 1. Missing data in the observations of trout perch. There are 199 missing\nmeasurements for trout perch weight, about 40% of the 489 total observations.\nThis missingness in our response variable lowers the sample size of our dataset.")

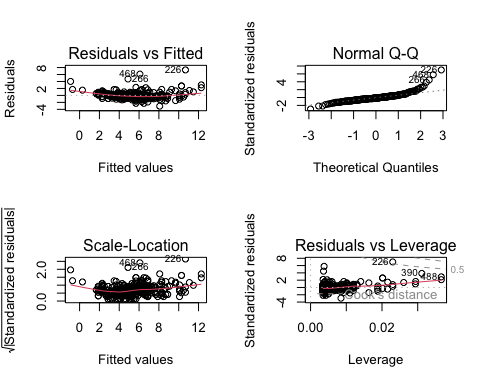


### 3.

# creating a linear model for the trout perch dataset   
trout\_model <- lm(weight ~ length, data = trout)

### 4.

# plotting the trout model in a 2 by 2 grid  
par(mfrow = c(2, 2))  
plot(trout\_model)

 Figure 2. Linear models for the trout perch data set help visualize residuals and errors.

dev.off()

## null device   
## 1

### 5.

**Residuals vs Fitted:** There is a relatively even distribution above and below the line, however, the points are clustered towards the center of the plot. Based on past examples I would say the assumption of homoscedasticity is violated.

**Scale-Location:** the residuals are distributed slightly more evenly in this plot, however there is still a bias towards the center.

**Normal Q-Q:** The points follow a linear path until a slight deviation around positive 2. Based on past examples I would say the residuals are normally distributed.

**Residuals vs Leverage:** there are a few outliers measured influential by Cook’s distance.

### 6.

# creating a summary of the trout perch model  
model\_summary <- summary(trout\_model)  
  
model\_summary

##   
## Call:  
## lm(formula = weight ~ length, data = trout)  
##   
## 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

## 7.

# setting up the initial ANOVA table  
model\_anova <- anova(trout\_model)

# creating a summary table  
model\_anova\_table <- tidy(model\_anova) %>%   
 # rounding the number of significant digits shown in the table  
 mutate(across(sumsq:meansq, ~ round(.x, digits = 3))) %>%   
 mutate(statistic = round(statistic, digits = 1)) %>%   
 # removing the scientific notation  
 mutate(p.value = case\_when(p.value < 0.001 ~ "< 0.001")) %>%   
 # meaningful term title  
 mutate(term = case\_when(term == "length" ~ "Trout Perch Length",   
 TRUE ~ term)) %>%  
 # turning the data frame into a flextable  
 flextable() %>%   
 # giving the headers clearer labels   
 set\_header\_labels(df = "Degrees of Freedom",   
 sumsq = "Sum of squares",  
 meansq = "Mean squares",  
 statistic = "F-statistic",  
 p.value = "p-value")  
   
model\_anova\_table

| term | Degrees of Freedom | Sum of squares | Mean squares | F-statistic | p-value |
| --- | --- | --- | --- | --- | --- |
| Trout Perch Length | 1 | 1,432.288 | 1,432.288 | 1,280.8 | < 0.001 |
| Residuals | 288 | 322.052 | 1.118 |  |  |

Table 1. Analysis of variance table summarizing the statistical restults.

## 8.

Though the ANOVA table provides it more succinctly, both the ANOVA and summary functions extract and represent important information about our trout perch model. The test statistic and the p-value are represented by both of these functions and can show how accurately our model predicts the correlation between the predictor variable (trout perch length) and the response variable (trout perch weight).

## 9.

We performed a linear regression in order to analyze whether the length of a trout perch predicts its weight, our null hypothesis being that length is not an accurate predictor of weight. Based on our observations,

For each decrease (or increase) in the x variable, we expect the y variable to increase/decrease by the amount of the slope An ANOVA (Table 1) showed the test statistic to be 1,280.8, and the p-value to be < 0.001, meaning that

The sample size of our study was initially 489 observations, however, missing data for trout perch weights reduced that number to 290. The ANOVA table (Table 1) shows the test statistic to be 1,280.8, and the p-value to be < 0.001 . was equal to 0.8164, which shows that our model fits the observed values fairly well.

context (i.e. what do these stats mean in the context of the original question)? null hypothesis test sample size test statistic degrees of freedom 1 p-value < 0.001 significance level R2 0.8164 interpretation of model equation

## 10.

# creating the initial predictions object  
predictions <- ggpredict(trout\_model, terms = "length")

# creating the plot of predictions  
# underlying data of trout perch dataset  
predict\_plot <- ggplot(data = trout,  
 aes(x = length, y = weight)) +  
 geom\_point() +  
 # plotting predictions  
 geom\_line(data = predictions,   
 aes(x = x, y = predicted),   
 color = "purple", linewidth = 1) +  
 # plotting the confidence interval   
 geom\_ribbon(data = predictions,   
 aes(x = x, y = predicted, ymin = conf.low,   
 ymax = conf.high), alpha = 0.3) +  
 #theme and labels  
 theme\_bw() +  
 labs(x = "Trout perch length",   
 y = "Trout perch mass",  
 caption = "Figure 3. Predicted relationship between trout perch length and trout perch weight.\nPoints represent collected data and the line is the predicted relationship with shaded confidence interval")  
  
  
predict\_plot

