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.

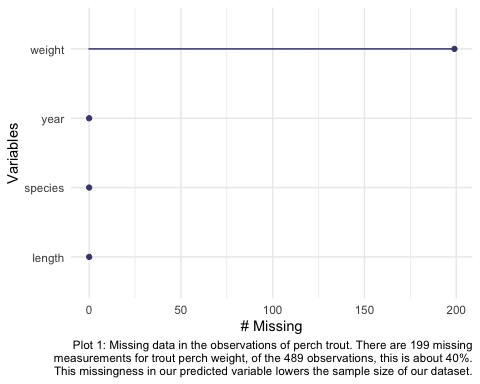
Mathematical:

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 = "Plot 1: Missing data in the observations of perch trout. There are 199 missing\nmeasurements for trout perch weight, of the 489 observations, this is about 40%.\nThis missingness in our predicted variable lowers the sample size of our dataset.")

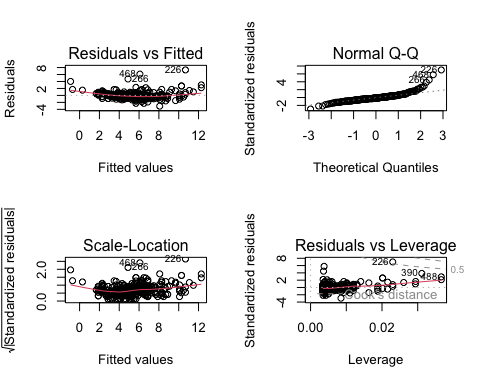


## 3.

# 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)



dev.off()

## null device   
## 1

check\_normality(trout\_model)

## Warning: Non-normality of residuals detected (p < .001).

check\_heteroscedasticity(trout\_model)

## Warning: Heteroscedasticity (non-constant error variance) detected (p < .001).

## 5.

Residuals vs Fitted: the residuals are somewhat evenly distributed. There is a relatively even distribution above and below the line, however, the points are concentrated towards the middle of the plot.

Scale-Location: there is slightly more homoscedasticity of variance in this plot, however there is still a bias towards the middle.

Normal Q-Q: the residuals are fairly normally distributed. The points follow a linear path until a slight deviation around positive 2.

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) %>%   
 mutate(across(sumsq:meansq, ~ round(.x, digits = 2))) %>%   
 mutate(statistic = round(statistic, digits = 1)) %>%   
 mutate(p.value = case\_when(p.value < 0.001 ~ "< 0.001")) %>%   
 mutate(term = case\_when(term == "length" ~ "Trout Perch Length",   
 TRUE ~ term)) %>%  
 flextable() %>%   
 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 | 1,432.29 | 1,432.29 | 1,280.8 | < 0.001 |
| Residuals | 288 | 322.05 | 1.12 |  |  |

## 8.

connection between ANOVA table and summary object

## 9.

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 p-value significance level R2 interpretation of model equation

## 10.

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

predict\_plot <- ggplot(data = trout,  
 aes(x = length, y = weight)) +  
 geom\_point() +  
 geom\_line(data = predictions,   
 aes(x = x, y = predicted),   
 color = "purple", linewidth = 1) +  
 geom\_ribbon(data = predictions,   
 aes(x = x, y = predicted, ymin = conf.low,   
 ymax = conf.high), alpha = 0.3) +  
 theme\_bw() +  
 labs(x = "Trout perch length", y = "Trout perch mass")  
  
  
predict\_plot

