code for ENVS193DS HW4

Maddie Manzagol

5/25/23

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

#load in packages  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.1 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(here)

here() starts at /Users/maddiemanzagol/Desktop/ENVS 193DS/github/ENVS-193DS\_homework-04\_manzagol-maddie

library(lterdatasampler)  
library(naniar)  
library(performance)  
library(broom)  
library(flextable)

Attaching package: 'flextable'  
  
The following object is masked from 'package:purrr':  
  
 compose

library(ggeffects)  
library(car)

Loading required package: carData  
  
Attaching package: 'car'  
  
The following object is masked from 'package:dplyr':  
  
 recode  
  
The following object is masked from 'package:purrr':  
  
 some

#read in the data   
fishdata <- read\_csv(here::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)

Rows: 349229 Columns: 15  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (8): lakeid, gearid, spname, sampletype, indid, fishpart, spseq, flag  
dbl (5): year4, depth, rep, length, weight  
lgl (1): sex  
date (1): sampledate  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#Problem 1

1. Null and alternative hypotheses in mathematical and biological terms

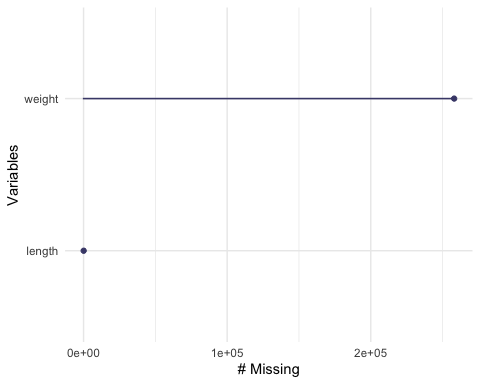
Mathematical: HO: β = 0; HA: β ≠ 0

Biological: The null hypothesis will be “fish length does not predict fish weight for trout perch”; meanwhile, the alternative hypothesis will be “fish length does predict fish weight for trout perch”

1. Create a visualization of the missing data for the filtered data set containing the observations you will use and also write an accompanying caption explaining how/if the missing data is relevant to your hypotheses.

#create a subset of the data to only include length and weight   
fish\_summary <- subset(fishdata, select = c(length, weight))

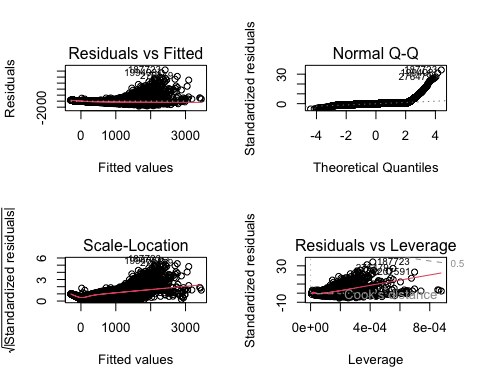
#visualize the missing data  
gg\_miss\_var(fish\_summary)



The missing data shown from the gg\_miss\_var plot highlights a substantial amount of missing data for fish weight. This will be important to note because our statistical question is asking if fish length predicts fish weight; however, if a lot of data is missing for fish weight then it will likely be difficult to accurately determine a relationship between fish length and fish weight.

1. Run your test (Linear regression test)

#perform a linear regression test   
fishlinearregression <- lm(weight~length, data= fish\_summary)  
  
# makes the viewer pane show a 2x2 grid of plots   
# format: par(mfrow = c(number of rows, number of columns))  
#display the diagnostic tests  
par(mfrow = c(2, 2))  
plot(fishlinearregression)



plot 1–Residual vs Fitted: The residual vs fitted plot generated from “fishlinearregression” shows residuals that are not evenly and randomly distributed around the horizontal dotted line. Therefore, the assumption of homoscedasticity is violated.

plot 2–Normal QQ: The normal QQ plot shows residuals that are not entirely linearly distributed, meaning the points do not follow a straight line. There is a portion distributed between quantiles -2 to 2 that follow a very linear line but after quantile 2 the slope changes.

plot 3–Scale-Location: The scale-location plot highlights that the square root of residuals does not highlight an even and random spread around the red line. This means that the variances change and homoscedasticity is violated.

plot 4–Residual vs Leverage: The residual vs leverage plot highlights some outliers (labeled) indicating that there are data points that are influencing the model. This means that if the model was calculated without these labeled outliers, then the model would predict a different estimate.

#create a summary of the linear regression stats  
model\_summary <- summary(fishlinearregression)  
  
model\_summary

Call:  
lm(formula = weight ~ length, data = fish\_summary)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1661.7 -142.1 -6.5 118.8 10385.8   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -309.55373 1.60785 -192.5 <2e-16 \*\*\*  
length 3.00965 0.00814 369.7 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 303 on 91088 degrees of freedom  
 (258139 observations deleted due to missingness)  
Multiple R-squared: 0.6001, Adjusted R-squared: 0.6001   
F-statistic: 1.367e+05 on 1 and 91088 DF, p-value: < 2.2e-16

#perform an ANOVA test  
model\_squares <- anova(fishlinearregression)  
  
model\_squares

Analysis of Variance Table  
  
Response: weight  
 Df Sum Sq Mean Sq F value Pr(>F)   
length 1 1.2549e+10 1.2549e+10 136700 < 2.2e-16 \*\*\*  
Residuals 91088 8.3620e+09 9.1802e+04   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#clean up the ANOVA summary table   
model\_squares\_table <- tidy(model\_squares) %>%   
 mutate(across(sumsq:meansq,~round(.x, digits = 5))) %>%   
 mutate(statistic = round(statistic, digits = 1)) %>%   
 mutate(p.value = case\_when(p.value < 0.001~"<0.001")) %>%   
 mutate(term = case\_when(term == "stem\_length" ~ "Stem length (mm)",  
 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 |
| --- | --- | --- | --- | --- | --- |
| length | 1 | 12,549,293,413 | 12,549,293,413.06 | 136,700.2 | <0.001 |
| Residuals | 91,088 | 8,362,023,669 | 91,801.59 |  |  |

1. In 1-2 sentences, describe how the ANOVA table relates to the information you get from the summary() object.

The ANOVA table relates to the information you get from the summary object because both report the f-statistic, the degrees of freedom, and the p-value.

1. In 2-3 sentences, summarize your results in prose with in-text references to test results. Include all relevant information.

Given that we hypothesized that fish length would predict fish weights, our null hypothesis was that fish length did not predict fish weight. Fish weight was significantly predicted by differing fish lengths (analysis of variance, F(1, 91,088) = 136,700.2, p < 0.001,𝞪= 0.05).

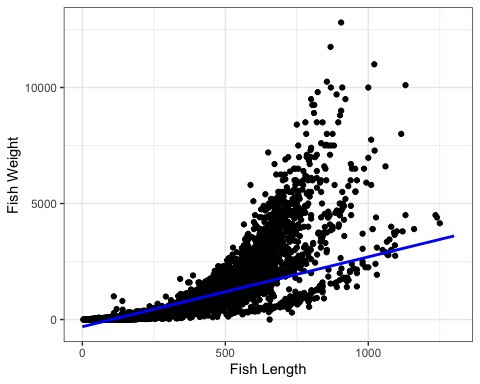
1. Create a visualization with model predictions and confidence intervals on top of the underlying data. Finalize your plot.
2. Write an accompanying caption

# extract model predictions using ggpredict  
predictions <- ggpredict(fishlinearregression, terms = "length")  
  
predictions

# Predicted values of weight  
  
length | Predicted | 95% CI  
---------------------------------------  
 0 | -309.55 | [-312.71, -306.40]  
 200 | 292.38 | [ 290.28, 294.47]  
 300 | 593.34 | [ 590.29, 596.39]  
 500 | 1195.27 | [1189.41, 1201.13]  
 700 | 1797.20 | [1788.27, 1806.13]  
 800 | 2098.16 | [2087.68, 2108.65]  
 1000 | 2700.09 | [2686.46, 2713.73]  
 1300 | 3602.99 | [3584.60, 3621.37]

plot\_predictions <- ggplot(data = fish\_summary,   
 aes(x = length, y = weight)) +  
 # first plot the underlying data from fish\_summary  
 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),   
 alpha = 0.2) +  
 # theme and meaningful labels  
 theme\_bw() +  
 labs(x = "Fish Length",  
 y = "Fish Weight")  
  
plot\_predictions

Warning: Removed 258139 rows containing missing values (`geom\_point()`).



1. The figure above plot fish weight prediction as a function of fish length using data collected about perch trout located in the North Temperate Lakes from 1981 to present day. The data points represent the predicted fish weight given a fish length; meanwhile, the blue line indicates the overall trendline for the predicted correlation between fish length and fish weight.