Homework 4

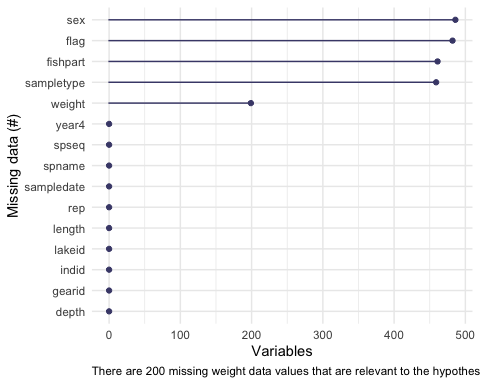
Rylee Pupa

library(here)  
library(janitor)  
library(tidyverse)  
library(naniar)  
library(visdat)  
library(tidyverse)  
library(here)  
library(lterdatasampler)  
library(performance)  
library(broom)  
library(flextable)  
library(ggeffects)  
library(car)

1. Write your null and alternative hypotheses in mathematical and biological terms.
   * null: fish length does not predict fish weight for trout perch across all sample years
     + math: In statistical terms, H0: B1 = 0
   * alternative: fish length does predict fish weight for trout perch across all sample years
     + math: In statistical terms, HA: B1 ≠ 0

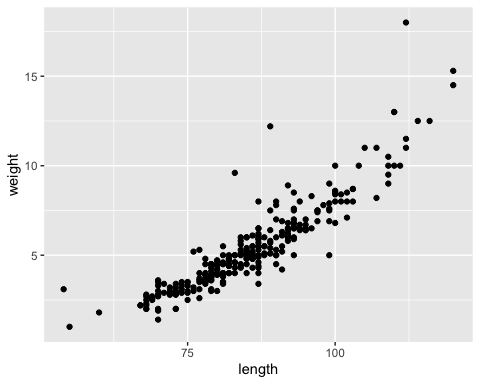
#reading in files using here function, clean names to make lowercase and more readable  
fish <- read\_csv(here("data", "ntl6\_v12.csv")) %>%   
 clean\_names()  
  
#filter fish to only include trout perch  
trout\_perch <- fish %>%   
 filter(spname == "TROUTPERCH")

#2.visualize missing data  
gg\_miss\_var(trout\_perch) + labs(x = "Missing data (#)",  
 y = "Variables", caption = "There are 200 missing weight data values that are relevant to the hypothesis") + theme(plot.caption = element\_text(hjust = 0))

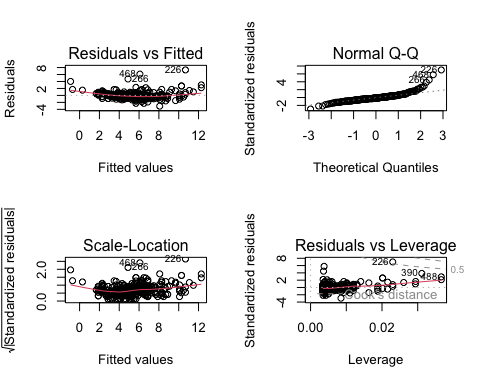


#Extended Caption: There is missing data for sex, flag, fish part, sample, type, and weight. There are 200 missing data values for weight, which is the only missing data relevant to our hypothesis.

#plotting points to explore if check if there may be a linear relationship in the data  
ggplot(data = trout\_perch, aes(x = length, y = weight)) +  
 geom\_point()



#3 running our linear tests (can check assumptions from this)  
#creating a linear model of the length and weight  
troutperch\_model <- lm(weight ~ length, data = trout\_perch)  
  
#make pane show grid of diagnostic plots from linear model   
par(mfrow = c(2,2))  
plot(troutperch\_model)



#to turn off two by two grid   
#dev.off()  
  
#remove outliers

#5) For each diagnostic plot, describe in 1-2 sentences what it is showing you, and what youdecide after looking at the plot. Check assumptions:

1. The Residuals vs. fitted plot shows us if the errors or residuals in the relationship are homoscedastic, in other words whether or not there is a constant variance. The data does not look homoscedastic because the errors plotted are not mostly constant throughout the range of x-axis values, both the data and the outliers are not evenly and randomly distributed across the horizontal line
2. The Normal QQ plot shows us if the errors are normally distributed, with the straight line through the plot I would assume yes it is normally distibuted.
3. The scale location plot also checks for homoscedasticy of variance but instead uses square root of standardized residuals. The variance of residuals is not evenly and not randomly distributed so I would also say that the residuals are not homockedastic.
4. The residuals versus leverage plot (also called Cook’s model) helps us check if outliers are influencing our model estimates by measuring the influence of a single observation on the model. Since there are several points labeled in the graph, I would say yes outliers are influencing.

Other assumptions not shown in grid:

Assumption 1: There is a linear relationship between variables: Yes! From our exploratory data visualization we can see a mostly linear relationship despite outliers.

Assumption 2: The errors are independent (no correlation): Yes we would assume so because the data collected was independent, so we can assume that errors in one measurement wouldn’t influence errors in other observations.

#storing our summary model as an object   
model\_summary <- summary(troutperch\_model)  
  
#storing ANOVA table as an object  
model\_anova <- anova(troutperch\_model)  
  
#6) displaying results from summary using the model object  
model\_summary

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

model\_anova

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

summarizing model in a table

#7 creating a table that summarizes the anova table   
  
model\_anova\_table <- tidy(model\_anova) %>%   
 #rounding the statistics, sum of mean squares and sum of squares, to two digits  
 mutate(across(sumsq:meansq, ~ round(.x, digits = 2))) %>%   
 #round the F-statistic value to have 2 digits  
 mutate(statistic = round(statistic, digits = 2)) %>%   
 ##replace the p value with < 0.001 rather than a super small number  
 mutate(p.value = case\_when(p.value < 0.001 ~ "< 0.001")) %>%   
 #rename the length to be more proper  
 mutate(term = case\_when(term == "length" ~ "Length", TRUE ~ term)) %>%   
 #create a flex table object from the data frame  
 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\_anova\_table

| term | Degrees of Freedom | Sum of Squares | Mean Squares | F-statistic | P-value |
| --- | --- | --- | --- | --- | --- |
| Length | 1 | 1,432.29 | 1,432.29 | 1,280.84 | < 0.001 |
| Residuals | 288 | 322.05 | 1.12 |  |  |

1. In 1-2 sentences, describe how the ANOVA table relates to the information you get from the **summary()** object
   * The Anova table is formed from the anova function which extracts the summary information from the model object, including the degrees of freedom, sum of squares, mean squares, F-statistic, and P-value. This data is meant to compare the means and test for significant differences between them.
2. In 2-3 sentences, summarize your results in prose with in-text references to test results. Include all relevant information.
   * I hypothesized that fish length would not predict fish weight for trout perch across our sample years. The results of the analysis show that length does predict fish weight in trout perch across all years (analysis of variance, F(1, 288) = 1280.84 p < 0.001, ⍺ =0.05).

#10) creating visualization with model predictions and confidence intervals   
  
#find model predictions for the linear model using ggpredict   
troutperch\_predict <- ggpredict(troutperch\_model, terms = "length")  
troutperch\_predict

# 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]

#plotting predictions from the ggpredict, with length on x and weight on y  
predict\_plot <- ggplot(data = troutperch\_model,   
 aes(x = length, y = weight)) +  
 #plotting the troutperch data lone  
 geom\_point() +  
 #plotting predictions of linear model, changing aesthetics of plot line  
 geom\_line(data = troutperch\_predict,   
 aes(x = x, y = predicted),   
 color = "red", linewidth = 0.8) +  
 #plotting the confidence interval from the gg predict function (95%)  
 geom\_ribbon(data = troutperch\_predict,   
 aes(x = x, y = predicted, ymin = conf.low, ymax = conf.high),   
 alpha = 0.3) +  
  
 #theme and meaningful labels, expand on caption  
 theme\_classic() +  
 labs(x = "Length (millimeter)",  
 y = "Weight (gram)", caption = "Model predictions for if fish length predicts fish weight")  
  
predict\_plot

