Homework 04

Adam Guerra

5/17/23

Problem 1

1. Null and Alternative Hypothesis

Mathematical

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

Biological

- H_0 : Fish length is not a significant predictor of fish weight for trout perch.
- H_A : Fish length is a significant predictor of fish weight for trout perch.

2. Visualize the missing data

```
#clean and select data
fish_data <- sqldf("SELECT year4, spname, length, weight
                    FROM fish_raw_data
                    WHERE spname = 'TROUTPERCH'")

#visualize missing data
vis_miss(fish_data)
```

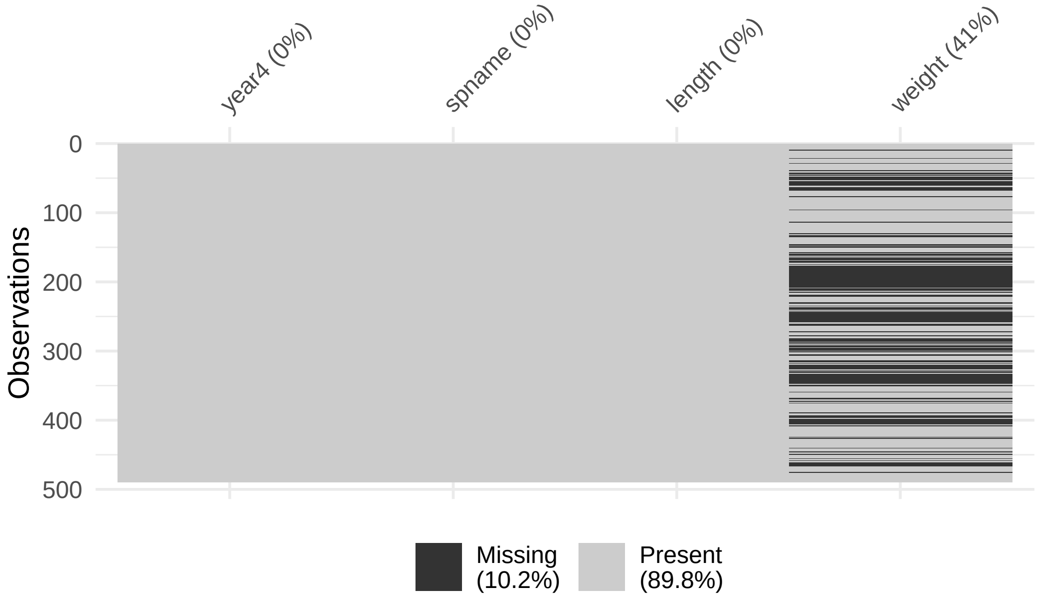

Figure 1: As shown in the figure above, 41% of the weight category has NAs. This is going to limit how many observations I have to use in the linear model, reducing its statistical power.

3. Run test

```
#create linear model
fish_lm <- lm(weight ~ length, data = fish_data)

#get residuals
fish_res <- fish_lm$residuals
```

4. Visually check assumptions

```
#diagnostic plots
par(mfrow = c(2,2)) #set 2x2 format
plot(fish_lm)
```

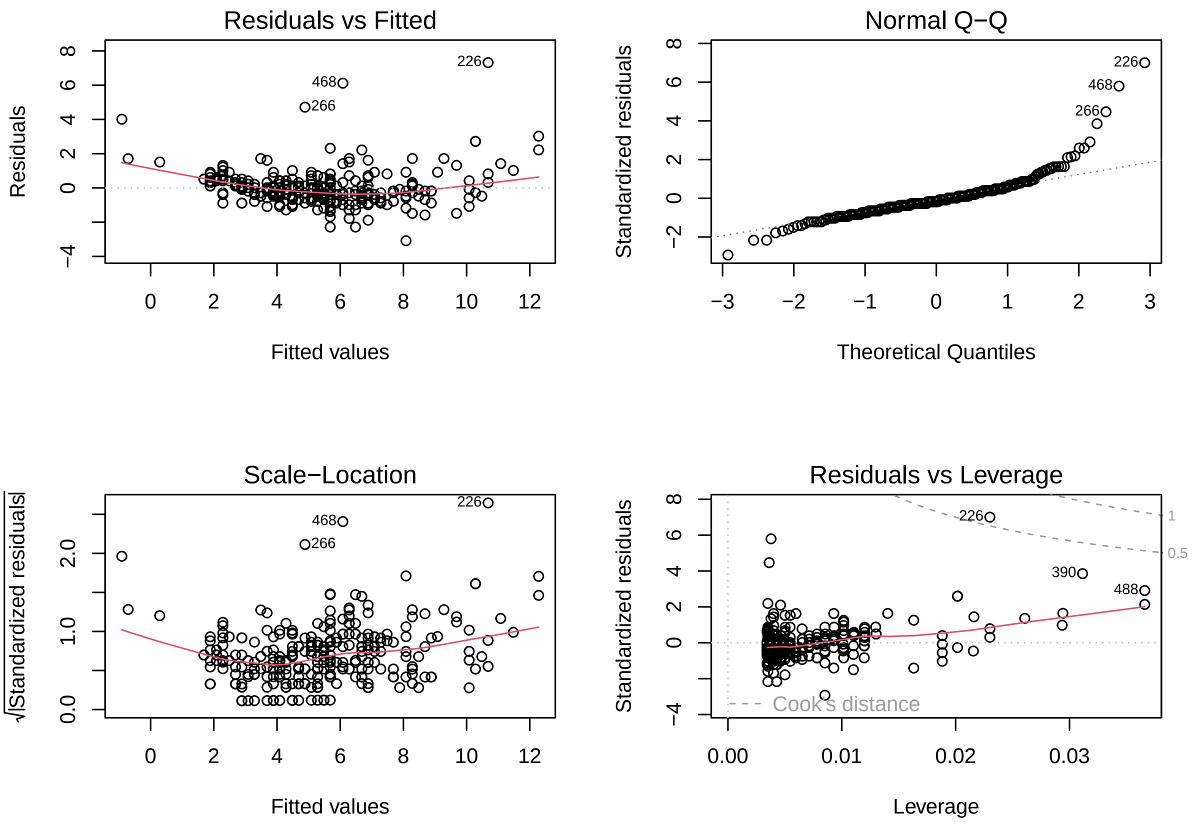

Figure 2: Diagnostic plots for linear model.

5. Descriptions of diagnostic plots

- Residuals vs Fitted: Shows residuals and fitted line to visualize linearity and constant variance. Points appear to be evenly and randomly distributed around the line, although there are a few outliers.
- QQ: Shows both data sets against one another. Data appears to be normally distributed.
- Scale Location: Similar to residuals vs fitted but is more accurate for showing homoscedasticity of variance. Data appears to be evenly and randomly distributed around the line.
- Residuals vs Leverage: Shows which data points are influential in the model. There are a few outliers identified that could be influential.

6. Display summary of model object

```r
#sumary of model object
summary(fish_lm)
```

```

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
  (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. Create summary ANOVA table

```r
#store ANOVA table as object
fish_squares <- anova(fish_lm)

fish_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
```

```r
#create table
fish_squares_table <- tidy(fish_squares) |>
                      #round the sum of squares and mean squares columns to have 5 digits
                      mutate(across(sumsq:meansq, ~ round(.x, digits = 5))) |>
                      #round the F-statistic to have 1 digit
                      mutate(statistic = round(statistic, digits = 1)) |>
                      #replace the very very very small p value with < 0.001
                      mutate(p.value = case_when(
                             p.value < 0.001 ~ "< 0.001")) |>
                      #rename the stem_length cell to be meaningful
                      mutate(term = case_when( term == "length" ~ "Fish Length (mm)",
                             TRUE ~ term)) |>
                     # format(scientific = T) |>
                      #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")

fish_squares_table
```

term	Degrees of Freedom	Sum of squares	Mean squares	F-statistic	p-value
Fish Length (mm)	1	1,432.2877	1,432.28769	1,280.8	< 0.001
Residuals	288	322.0525	1.11824		

8. Describe how the ANOVA table relates to the information given from summary object.

The information from this table is the relevant information about where the p-value and R^2 come from.

9. Summarize results with in-text references to test results.

After checking the assumptions for a linear model (step 4), I performed the linear regression model of length and weight of trout that showed that length is a significant predictor of weight in trout (part 6). It also showed that 81.6% the variance in weight of trout can be explained by length. This is a high percentage, meaning that this model is a good fit.

10. Visualize the model predictions and confidence intervals.

```
#generate predictions
predictions <- ggpredict(fish_lm, 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]
```

```
#create visualization
plot_predictions <-
  #set global ggplot parameters
  ggplot(data = fish_data, aes(x = length, y = weight)) +
  #plot normal points
    geom_point() +
  #plot regression line
    geom_line(data = predictions, aes(x = x, y = predicted), color = 'red',
              linewidth = 1) +
  #plot confidence interval and make it transperent
    geom_ribbon(data = predictions, aes(x = x, y = predicted,
                                        ymin = conf.low, ymax = conf.high), alpha = 0.2) +
  #set theme
    theme_bw() +
  #create labels and captions
```

```
    labs(x = 'Fish Length (mm)', y = 'Fish Weight (g)',
         title = 'Fish Length as a Predictor of Fish Weight',
         caption = "Figure 3: Shows fish lengths' and weights' plotted against the predict
                interval shaded around the regression line.") +
  #change font type and caption loaction
    theme(plot.caption = element_text(hjust = 0),
          text = element_text(family = 'Times'))

#show plot
plot_predictions
```

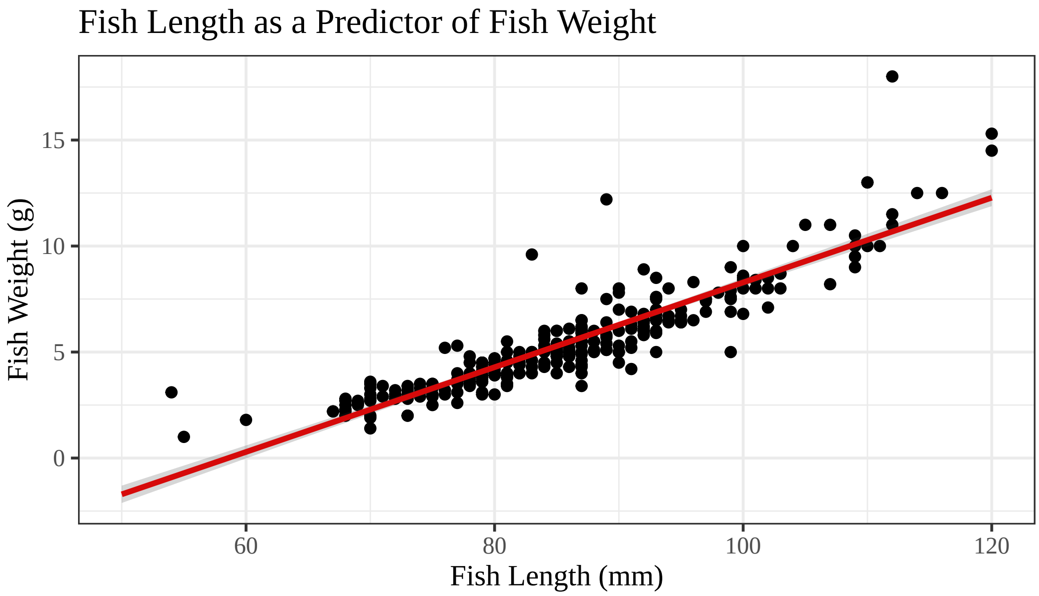

Figure 3: Shows fish lengths' and weights' plotted against the predictions with an accopanying confid
interval shaded around the regression line.

Repo Link