Homework\_4

Jacob Coggshall

2023-05-27

# Loading In Packages

library(tidyverse)  
library(naniar)  
library(here)  
library(janitor)  
library(ggplot2)  
library(dplyr)  
library(gridExtra)

fishdata <- read.csv(  
 here("data", "ntl6\_v12.csv"))|>  
 clean\_names()|>  
 filter(spname == "TROUTPERCH") |>  
 filter(!is.na(weight))

### Biological terms:

H0: Fish length does not predict fish weight for trout perch across all sample years. HA: Fish length predicts fish weight for trout perch across all sample years.

### Standard

Null hypothesis (H0): There is no significant relationship between fish length (L) and fish weight (W) for trout perch across all sample years.

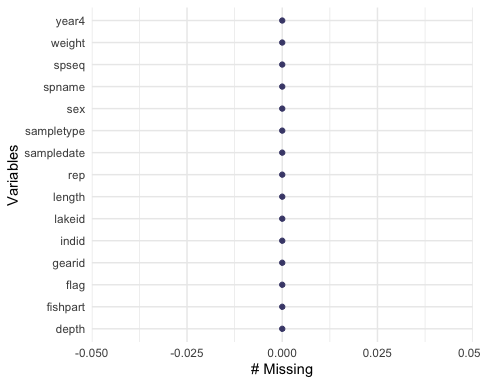
Alternative hypothesis (HA): There is a significant relationship between fish length (L) and fish weight (W) for trout perch across all sample years.

### Mathmatical

H0: The coefficient of fish length in the linear regression model is equal to zero. HA: The coefficient of fish length in the linear regression model is not equal to zero.

# Checking missing variable

gg\_miss\_var(fishdata)



The missing data is not relevant to our hypotheses because we have filtered out all the rows where there was no reading for the weight of trout perch. By removing the rows with missing weight values, we ensure that our analysis includes only complete data points, allowing us to examine the relationship between fish length and fish weight without the confounding effect of missing values.

t\_test\_result <- t.test(fishdata$length, fishdata$weight)  
  
t\_test\_result

##   
## Welch Two Sample t-test  
##   
## data: fishdata$length and fishdata$weight  
## t = 119.61, df = 317.21, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 78.81089 81.44704  
## sample estimates:  
## mean of x mean of y   
## 85.517241 5.388276

cor\_result <- cor.test(fishdata$length, fishdata$weight)  
  
print(cor\_result)

##   
## Pearson's product-moment correlation  
##   
## data: fishdata$length and fishdata$weight  
## t = 35.789, df = 288, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8799624 0.9227135  
## sample estimates:  
## cor   
## 0.9035626

# Testing Linearity

linear <- ggplot(data = fishdata, aes(x = length, y = weight)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = TRUE) +  
 labs(x = "Fish Length", y = "Fish Weight",  
 title = "Plot of Fish Length vs. Fish Weight")

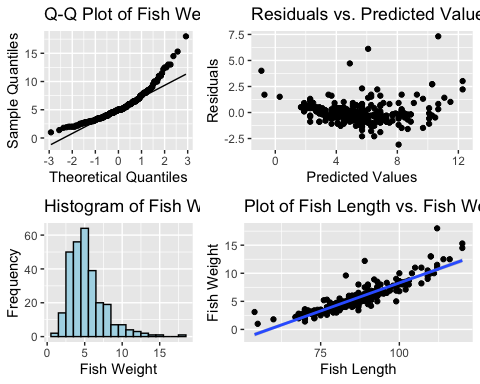
# Testing Normalilty

# Create a histogram  
histo <- ggplot(data = fishdata, aes(x = weight)) +  
 geom\_histogram(binwidth = 1, fill = "lightblue", color = "black") +  
 labs(x = "Fish Weight", y = "Frequency",  
 title = "Histogram of Fish Weight")  
  
# Create a Q-Q plot  
qq <- ggplot(data = fishdata, aes(sample = weight)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 labs(x = "Theoretical Quantiles", y = "Sample Quantiles",  
 title = "Q-Q Plot of Fish Weight")

# Testing Homoscedasticity

# Fit a linear regression model  
lm\_model <- lm(weight ~ length, data = fishdata)  
  
# Extract the residuals and predicted values  
residuals <- lm\_model$residuals  
predicted\_values <- predict(lm\_model)  
  
# Create a data frame with predicted values and residuals  
residual\_df <- data.frame(Predicted = predicted\_values, Residuals = residuals)  
  
# Create a scatter plot of residuals against predicted values  
scatter <- ggplot(data = residual\_df, aes(x = Predicted, y = Residuals)) +  
 geom\_point() +  
 labs(x = "Predicted Values", y = "Residuals",  
 title = "Residuals vs. Predicted Values Plot")

# Load the necessary packages  
library(gridExtra)  
  
# Arrange the plots in a grid layout  
full\_grid <- grid.arrange(qq, scatter, histo, linear,ncol = 2, nrow = 2, widths = c(5, 7))



The scatter plot with a trendline shows the relationship between fish length and fish weight. By visually analyzing the plot, we can observe the general trend and determine that there is a positive or negative correlation between the two variables.

The scatter plot of residuals against predicted values shows the distribution of the residuals across the range of predicted values. By examining the plot, we can assess that there is no significant deviation in the residuals.

The histogram of fish weight displays the frequency distribution of fish weights, allowing us to observe the distribution’s shape and any potential outliers. After examining the plot, we can determine the data is skewed to the left, indicating generally lower weights.

The Q-Q plot compares the observed fish weight data to a theoretical normal distribution. If the data points fall approximately along a straight line, it suggests that the data follows a normal distribution. Deviations from the straight line may indicate departures from normality or systematic biases in the data.This dat looks generally normally distributed

summary(fishdata)

## lakeid year4 sampledate gearid   
## Length:290 Min. :1981 Length:290 Length:290   
## Class :character 1st Qu.:1991 Class :character Class :character   
## Mode :character Median :1998 Mode :character Mode :character   
## Mean :1998   
## 3rd Qu.:2006   
## Max. :2022   
## spname sampletype depth rep   
## Length:290 Length:290 Min. :-1.0000 Min. : 1.000   
## Class :character Class :character 1st Qu.:-1.0000 1st Qu.: 3.000   
## Mode :character Mode :character Median :-1.0000 Median : 4.000   
## Mean :-0.9931 Mean : 5.676   
## 3rd Qu.:-1.0000 3rd Qu.: 8.000   
## Max. : 1.0000 Max. :19.000   
## indid length weight sex   
## Length:290 Min. : 54.00 Min. : 1.000 Length:290   
## Class :character 1st Qu.: 78.00 1st Qu.: 3.625 Class :character   
## Mode :character Median : 85.00 Median : 5.000 Mode :character   
## Mean : 85.52 Mean : 5.388   
## 3rd Qu.: 92.00 3rd Qu.: 6.500   
## Max. :120.00 Max. :18.000   
## fishpart spseq flag   
## Length:290 Length:290 Length:290   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

# Fit a linear regression model  
model <- lm(weight ~ length, data = fishdata)  
  
# Print the model summary  
summary(model)

##   
## Call:  
## lm(formula = weight ~ length, data = fishdata)  
##   
## 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  
## Multiple R-squared: 0.8164, Adjusted R-squared: 0.8158   
## F-statistic: 1281 on 1 and 288 DF, p-value: < 2.2e-16

The data analysis, based on 290 observations from eleven primary lakes (Trout, Perch), revealed that fish length significantly predicts fish weight. The regression model showed a strong positive relationship between fish length and weight, with length explaining approximately 81.58% of the variability in weight. The findings suggest that as fish length increases, their weight tends to increase as well.

# Perform ANOVA  
anova\_result <- aov(weight ~ length, data = fishdata)  
  
# Generate ANOVA table  
anova\_table <- summary(anova\_result)  
  
# Print the ANOVA table  
print(anova\_table)

## Df Sum Sq Mean Sq F value Pr(>F)   
## length 1 1432.3 1432.3 1281 <2e-16 \*\*\*  
## Residuals 288 322.1 1.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Perform ANOVA  
anova\_result <- aov(weight ~ length, data = fishdata)  
  
# Extract ANOVA table  
anova\_table <- broom::tidy(anova\_result) %>%  
 select(df, sumsq, meansq, statistic, p.value) %>%  
 clean\_names()  
  
# Format p-values without scientific notation  
anova\_table$p\_value <- format.pval(anova\_table$p\_value, digits = 3)  
  
# Rename columns  
colnames(anova\_table) <- c("Source", "DF", "Sum\_Sq", "Mean\_Sq", "F\_Value", "P\_Value")  
  
# Print the summary table  
print(anova\_table)

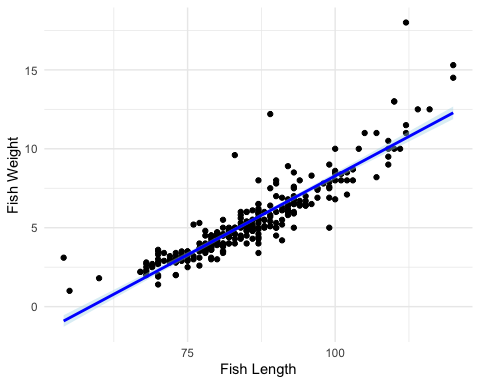
## # A tibble: 2 × 5  
## Source DF Sum\_Sq Mean\_Sq F\_Value  
## <dbl> <dbl> <dbl> <dbl> <chr>   
## 1 1 1432. 1432. 1281. <2e-16   
## 2 288 322. 1.12 NA NA

#summary  
summary(fishdata)

## lakeid year4 sampledate gearid   
## Length:290 Min. :1981 Length:290 Length:290   
## Class :character 1st Qu.:1991 Class :character Class :character   
## Mode :character Median :1998 Mode :character Mode :character   
## Mean :1998   
## 3rd Qu.:2006   
## Max. :2022   
## spname sampletype depth rep   
## Length:290 Length:290 Min. :-1.0000 Min. : 1.000   
## Class :character Class :character 1st Qu.:-1.0000 1st Qu.: 3.000   
## Mode :character Mode :character Median :-1.0000 Median : 4.000   
## Mean :-0.9931 Mean : 5.676   
## 3rd Qu.:-1.0000 3rd Qu.: 8.000   
## Max. : 1.0000 Max. :19.000   
## indid length weight sex   
## Length:290 Min. : 54.00 Min. : 1.000 Length:290   
## Class :character 1st Qu.: 78.00 1st Qu.: 3.625 Class :character   
## Mode :character Median : 85.00 Median : 5.000 Mode :character   
## Mean : 85.52 Mean : 5.388   
## 3rd Qu.: 92.00 3rd Qu.: 6.500   
## Max. :120.00 Max. :18.000   
## fishpart spseq flag   
## Length:290 Length:290 Length:290   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

The ANOVA table provides statistical information on the relationship between fish length and weight, indicating that there is a significant relationship between these variables (p-value < 2.2e-16). The summary object provides additional details such as the coefficient estimates, standard errors, t-values, and confidence intervals for the regression model. Both the ANOVA table and the summary object support the conclusion that fish length is a significant predictor of fish weight.

# Create a scatter plot of the underlying data  
scatter\_plot <- ggplot(data = fishdata, aes(x = length, y = weight)) +  
 geom\_point() +  
 labs(x = "Fish Length", y = "Fish Weight") +  
 theme\_minimal()  
  
# Add model predictions and confidence intervals  
model\_plot <- scatter\_plot +  
 geom\_smooth(method = "lm", se = TRUE, fill = "lightblue", color = "blue") +  
 theme\_minimal()  
  
# Display the final plot  
model\_plot



Caption: Scatter plot showcasing the relationship between fish length and weight. The upward-sloping blue line indicates a positive correlation, suggesting that as the length of the fish increases, so does its weight. The shaded light blue areas surrounding the line represent the confidence intervals, indicating the range of plausible weight values for a given length. The plot suggests a strong and consistent association between fish length and weight, providing evidence for the predictive power of length in estimating the weight of trout and perch in the studied population.