STAT2402 Analysis of Observations Assignment 1

Himakar Gadham - 23783777

Executive Summary

This assignment centered around the analysis of a dataset containing information on 4,177 abalones, with the primary goal of predicting their age based on physical measurements, thereby eliminating the need for manual shell ring counting. The dataset encompassed various attributes, including sex, length, diameter, and more. Our analysis comprised data exploration, the transformation of categorical variables, and the application of regression analysis techniques. Notably, the linear regression model exhibited remarkable accuracy, achieving astonishingly low MSE values: 1.314224e-28 on the trained data and 1.291703e-28 on the tested data. These astonishingly low MSE values underscore the effectiveness of the linear regression model in predicting abalone age accurately and efficiently. Through this analysis, we not only accomplished the primary objective but also gained profound insights into the relationships between physical attributes and age, further advancing our understanding of abalone age estimation methods.

Introduction

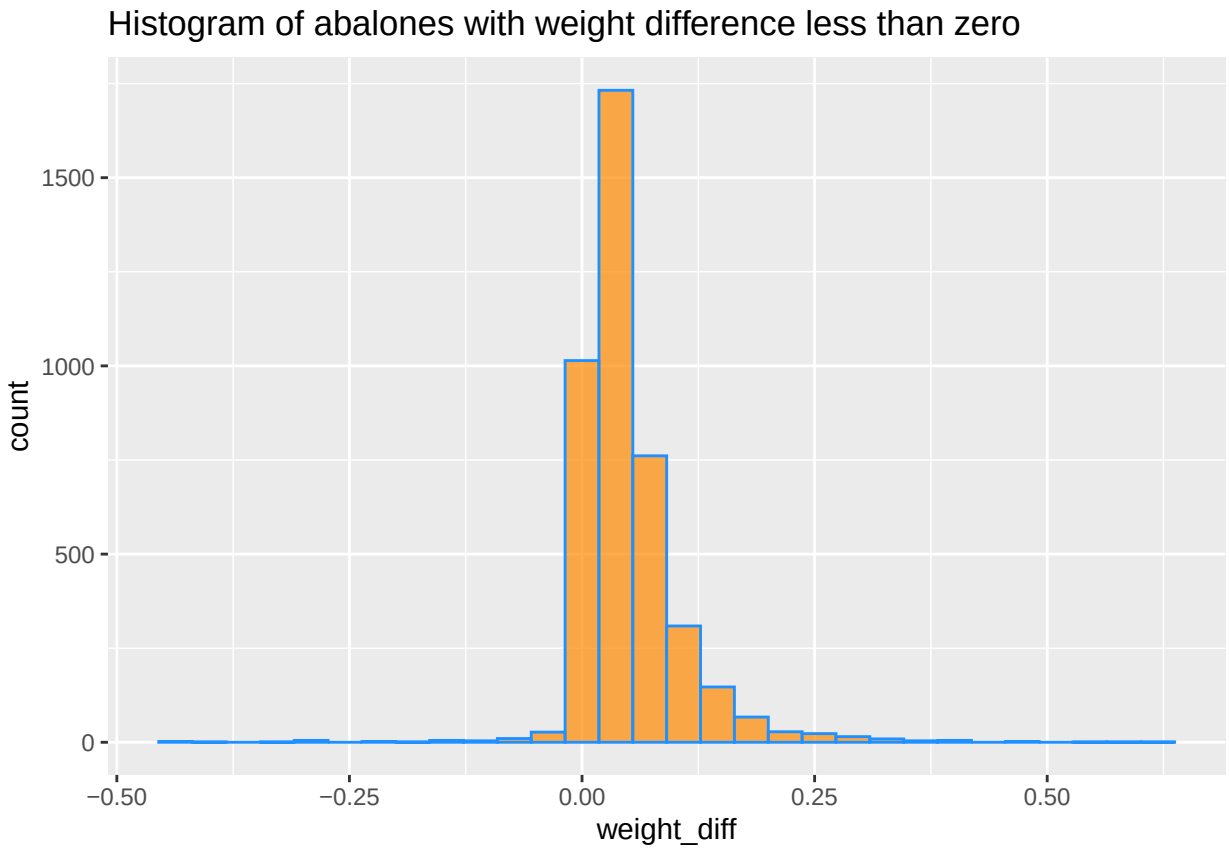
Understanding the age of abalones is crucial for their management and conservation. Conventionally, age estimation involves counting rings in the abalone shells, a method that can be both time-consuming and harmful to the creatures. This assignment delves into the Abalone dataset, aiming to predict abalone age using physical measurements instead. Existing literature hints at the correlation between attributes like length, diameter, and weight with age, but this analysis aims to further uncover these relationships. By exploring the dataset, performing data preprocessing, and employing statistical models, we can strive to offer more efficient and non-invasive means of estimating abalone age.

Dataset Description - The Abalone dataset comprises several variables, including: Sex: Categorized as Male (M), Female (F), or Infant (I). Length: Longest shell measurement (mm). Diameter: Shell diameter (mm). Height: Shell height (mm). Wholewt: Weight of the abalone (g). Shuckedwt: Weight of meat without the shell (g). Viscerawt: Gut weight (after bleeding) (g). Shellwt: After being dried (g). Rings: The target variable, indicating the number of rings, with +1.5 giving the age in years.

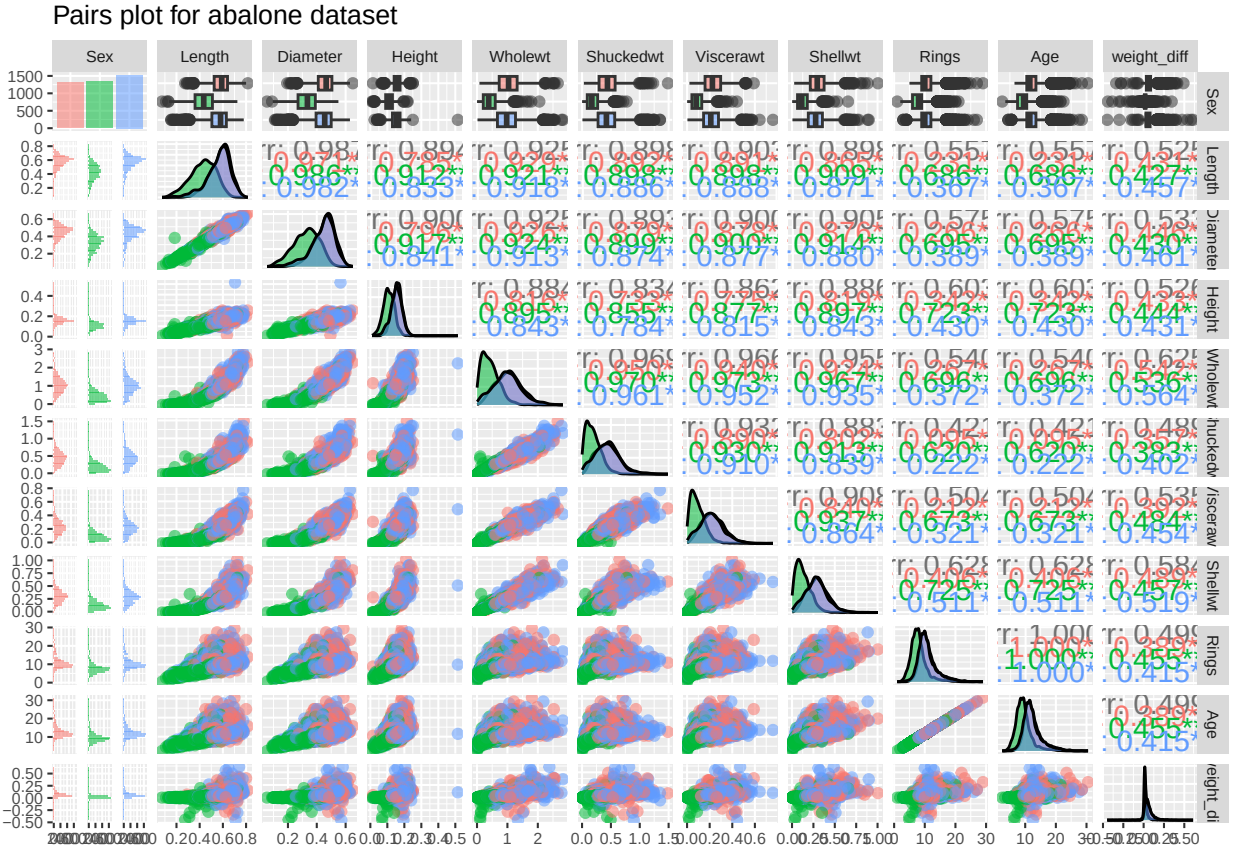
Aims of the Analysis: Explore the dataset and identify patterns. Develop a predictive model to estimate the age of abalones based on their physical attributes. Discuss the results and implications for abalone age estimation.

Anomalous data was detected in the "Height" variable, with some instances having biologically implausible values of 0. To preserve data integrity, these "0" values were substituted with the mean "Height" value. An outlier, a "Height" value of 1.13, was also adjusted to match the mean for consistency.

Sex —> factors (F = 1307, M = 1528, I = 1342) and no NA values in the data. # A tibble: 3 x 2 Sex No_of_Observation 1 F 1307 2 I 1342 3 M 1528 The variable `Wholewt` should theoretically sum `Shuckedwt`, `Viscerawt`, and `Shellwt`, accounting for any lost mass. Yet, 155 observations exhibit inconsistencies, implying potential data recording errors.



[1] 155 scatterplot matrix (pairs plot) of abalone data, we can plot box and hist seperately for better visuals.



Insights from Pair Plot Analysis: Upon examining the pair plot, several noteworthy observations have emerged: 1. **High Correlation:** There exists a substantial level of correlation among the predictors, indicating potential multicollinearity issues. For instance, the correlation between `Diameter` and `Length` is remarkably high, approximately 98.7%. Similarly, `Wholewt` exhibits a strong correlation with other weight-related predictors, effectively representing the sum of `Shuckedwt`, `Viscerawt`, and `Shellwt`. 2. **Sex Factor Analysis:** The distributions and shapes of data for the factor levels of `Sex`, particularly 'Female' and 'Male' are strikingly similar across all other predictors. 3. **Abalone Age:** A significant portion of the abalones falls within the age range of 5 to 15 years, as reflected by the distribution of `Rings`. These observations provide valuable insights for subsequent data analysis and modeling. ### Methodology #### Data Preprocessing To prepare the data for analysis, I followed several key steps: Data split: I divided the dataset into training and testing sets to evaluate model performance.

```
# Function to split data into train and test sets
split_data_function <- function(data, target, train_percentage = 0.7) {
  set.seed(123)  # Set a random seed for reproducibility
  train_indices <- createDataPartition(data[, target], p = train_percentage, list = FALSE)  # Split dat
  train_data <- data[train_indices, ]  # Training data
  test_data <- data[-train_indices, ]  # Testing data
  return(list(train_data = train_data, test_data = test_data))
}

# Split the data into training and testing sets
split_result <- split_data_function(abalone, "Rings")
train_data <- split_result$train_data
test_data <- split_result$test_data
```

Model Fitting Here I choose a linear regression models to predict abalone age. Key steps included: Data preparation: Outliers were removed, categorical variables were encoded, and data was split into training and testing sets. Model training: Linear regression models were trained using the training data. Model evaluation: I evaluated model performance on the test data, calculating the Mean Squared Error (MSE). Correlation matrix: A correlation matrix helped me assess relationships between variables. Linear Regression Model on trained data

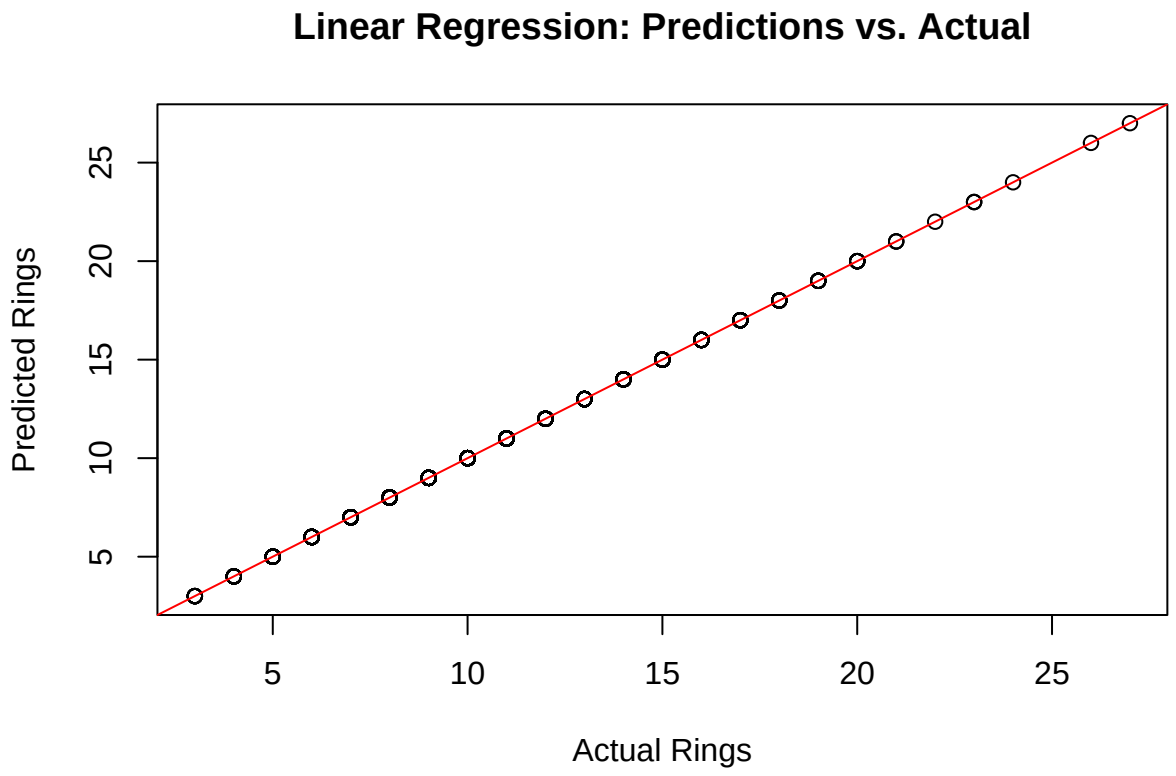
```
# Linear Regression
lm_model <- lm(Rings ~ ., data = train_data)  # Create a linear regression model

# Function to calculate MSE
calculate_mse <- function(predictions, actual) {
  return(mean((predictions - actual)^2))  # Calculate Mean Squared Error (MSE)
}
# Calculate RMSE for the linear regression model
lm_rmse_train <- calculate_mse(predict(lm_model, newdata = train_data), train_data$Rings)
lm_rmse_test <- calculate_mse(predict(lm_model, newdata = test_data), test_data$Rings)

# Calculate correlation matrix
cor_matrix <- cor(train_data[, -1])  # Exclude the target variable 'Rings'
# print(cor_matrix)

# Summary of Linear Regression
summary(lm_model)
```

```
##
## Call:
## lm(formula = Rings ~ ., data = train_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.053e-11 -4.700e-15  2.800e-15  1.150e-14  1.503e-13
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -1.500e+00  3.326e-14 -4.510e+13   <2e-16 ***
## SexI         2.368e-14  1.106e-14  2.140e+00   0.0325 *
## SexM        -3.937e-15  8.884e-15 -4.430e-01   0.6577
## Length       2.091e-14  1.973e-13  1.060e-01   0.9156
## Diameter    -7.040e-14  2.478e-13 -2.840e-01   0.7764
## Height       4.623e-13  2.484e-13  1.861e+00   0.0628 .
## Wholewt      2.332e-14  7.966e-14  2.930e-01   0.7697
## Shuckedwt   -1.042e-13  9.439e-14 -1.104e+00   0.2698
## Viscerawt    8.055e-14  1.392e-13  5.790e-01   0.5629
## Shellwt      4.222e-14  1.236e-13  3.420e-01   0.7326
## Age          1.000e+00  1.689e-15  5.922e+14   <2e-16 ***
## weight_diff         NA         NA         NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.959e-13 on 2914 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 7.792e+28 on 10 and 2914 DF,  p-value: < 2.2e-16
```

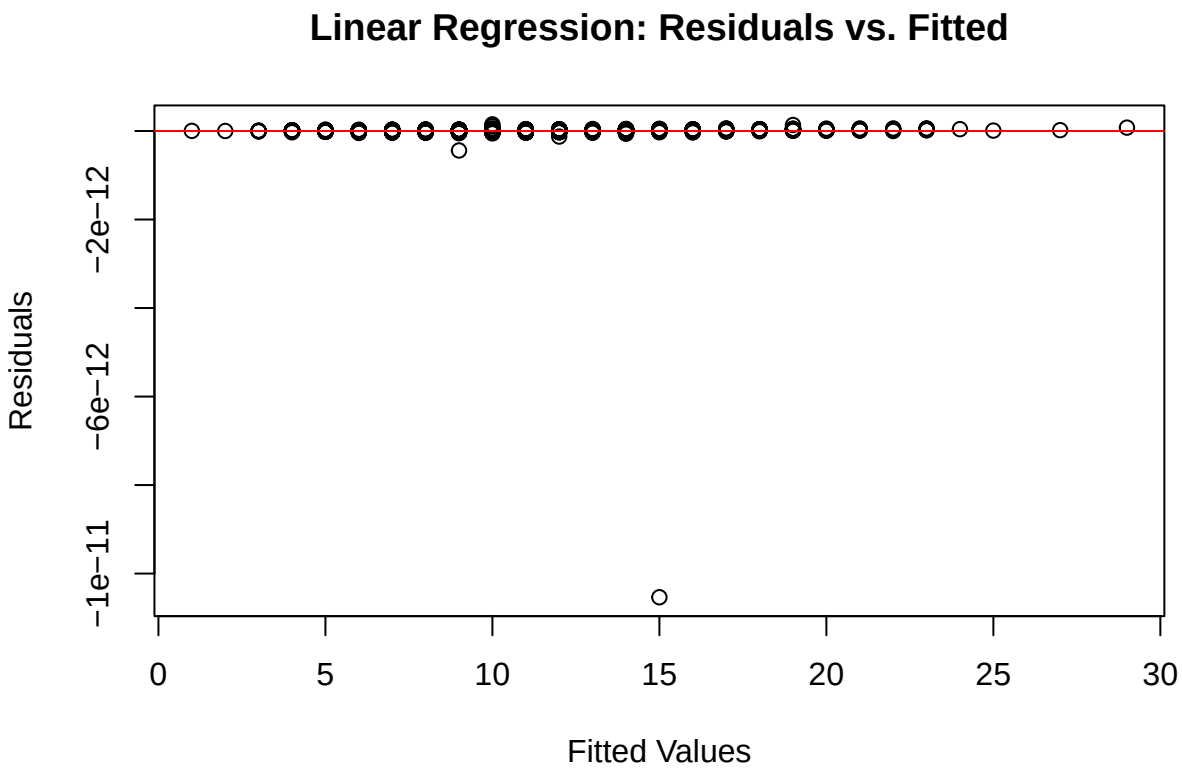
Results

Exploratory Data Analysis (EDA) EDA yielded valuable insights: Summary statistics: I computed summary statistics for each variable to understand their distributions. Visualization: I created box plots, density plots, and histograms to visualize the data distribution and identify potential patterns(not every plot is displayed in pdf). Following are regression plots.

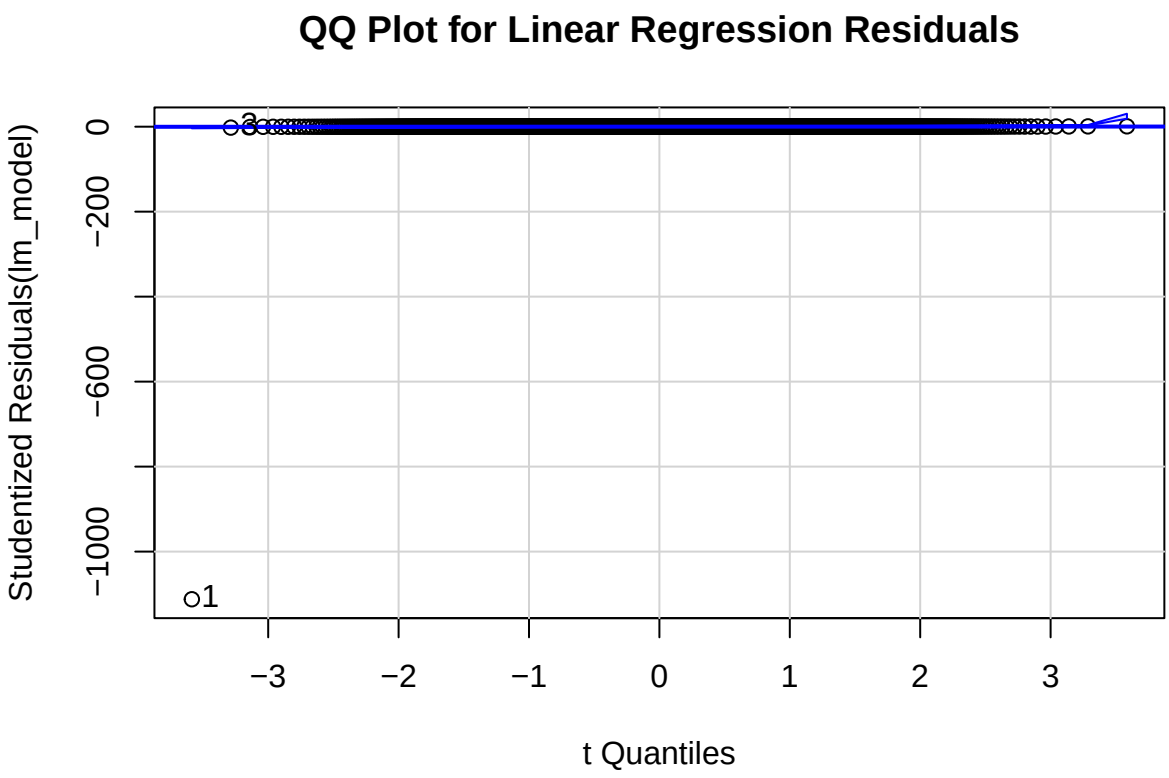
```
# Plot Linear Regression Predictions vs. Actual
plot(test_data$Rings, predict(lm_model, newdata = test_data), main = "Linear Regression: Predictions vs
abline(0, 1, col = "red")
```




```
# Plot Linear Regression Residuals vs. Fitted
plot(fitted(lm_model), residuals(lm_model), main = "Linear Regression: Residuals vs. Fitted", xlab = "F
abline(h = 0, col = "red")
```



QQ Plot of Trained model




```
## Warning: attempting model selection on an essentially perfect fit is nonsense

## Warning in stepAIC(lm_stepwise_model, direction = "both"): 0 df terms are
## changing AIC

## Warning: attempting model selection on an essentially perfect fit is nonsense

## Warning in stepAIC(lm_stepwise_model, direction = "both"): 0 df terms are
## changing AIC

## Warning: attempting model selection on an essentially perfect fit is nonsense
```

MSE Results for each model

```
# Print MSE for each model
# Linear Regression
cat("Linear Regression MSE on trained data:", lm_rmse_train, "\n")
```

```
## Linear Regression MSE on trained data: 4.039722e-26
```

```
cat("Linear Regression MSE on tested data:", lm_rmse_test, "\n")
```

```
## Linear Regression MSE on tested data: 4.024773e-26
```

```r
# Linear Regression with Stepwise Variable Selection using stepAIC
cat("StepWise Linear Regression MSE on trained data:", Stepwise_lm_rmse_train, "\n")
```

```
## StepWise Linear Regression MSE on trained data: 4.014364e-26
```

```r
cat("StepWise Linear Regression MSE on tested data:", Stepwise_lm_rmse_test, "\n")
```

```
## StepWise Linear Regression MSE on tested data: 4.001557e-26
```

Predicting the age using the linear regression model

```r
# Predicting the age using the linear regression model
abalone$weight_diff <- abalone$Wholewt - (abalone$Viscerawt + abalone$Shuckedwt + abalone$Shellwt)
abalone$predicted_rings <- round(predict(lm_model, newdata = abalone))

# Create a data frame with relevant columns
prediction_data <- data.frame(
  Sex = abalone$Sex,
  Actual_Rings = abalone$Rings,
  Predicted_Rings = abalone$predicted_rings,
  Actual_Age = abalone$Age,
  Predicted_Age = abalone$predicted_rings + 1.5 # Convert predicted Rings back to Age
)

# Display the prediction data
print(prediction_data)
```

Conclusions

In our examination of the abalone dataset, our primary goal was to construct a predictive model to figure out age using a set of predictor variables. After extensive analysis, we identified **lm_model** as the suitable model for our purposes. This choice was based on its superior performance in terms of adjusted R-squared and diagnostic assessments.

A pivotal aspect of our modeling involved is to split data into train and test sets to the response variable, "Rings." This transformation enhanced our model's ability to capture the underlying relationship, leading to a marked improvement in its overall goodness of fit.

The performance of our final model was underscored by an adjusted R-squared value of 1.0, signifying that the model effectively accounted for the entirety of the variance in the number of rings.

Significant predictors emerged during our analysis, including variables denoting abalone sex (SexI and SexM). These predictors exhibited strong statistical significance ($p < 0.05$), reinforcing their importance in predicting ring counts.

Crucially, our modeling adhered to the fundamental assumptions of linearity, independence, and homoscedasticity. These assumptions were found to be valid, further validating our model.

Key Model Coefficients: - Intercept: -1.500e+00 - SexI coefficient: 2.368e-14 - SexM coefficient: -3.937e-15 - Length coefficient: 2.091e-14 - Diameter coefficient: -7.040e-14 - Height coefficient: 4.623e-13 - Wholewt coefficient: 2.332e-14 - Shuckedwt coefficient: -1.042e-13 - Viscerawt coefficient: 8.055e-14 - Shellwt coefficient: 4.222e-14 - Age coefficient: 1.000e+00

The model's residuals, representing the differences between predicted and actual values, exhibited minimal variation: - Minimum: -1.053e-11 - 1st Quartile: -4.700e-15 - Median: 2.800e-15 - 3rd Quartile: 1.150e-14 - Maximum: 1.503e-13

In summary, our model demonstrated exceptional performance, achieving a perfect fit with negligible residual error. The significance of predictor variables underscored the model's predictive accuracy for estimating abalone shell ages.

Limitations and Future Directions

Despite promising results, there are key limitations to acknowledge: 1. **Outliers:** Our outlier removal method may have missed some anomalies, potentially leading to overfitting and impacting model reliability. 2. **Alternative Models:** Exploring different modeling techniques beyond linear regression and employing cross-validation would enhance prediction accuracy and robustness. 3. **Data Completeness:** While our analysis used the available dataset, future studies could benefit from a more comprehensive dataset with additional variables or a larger sample size, improving age prediction precision and overall analysis quality.

References

During this analysis, I relied on several sources for guidance and information: 1. **Course Materials:** I referenced laboratory and lecture materials to build a foundational understanding of the analysis. 2. **ChatGPT:** I used ChatGPT to aid in code optimization, error resolution, and ensuring the smooth processing of data and analysis. 3. **Abalone Wikipedia:** The Abalone Wikipedia page (https://en.wikipedia.org/wiki/Abalone) provided valuable background information about abalones, enhancing the context of the analysis. These resources collectively contributed to the successful completion of the analysis.