Assignment 17: Predict Fish Weight Using Multiple Linear Regression

100 Points scaled to 20 Points

Introduction

This assignment explores multiple linear regression to predict the weight of a fish using multiple measures of length and measures of height and width. The species of fish will be represented as dummy variables. These data ("Fish.csv") are provided on Kaggle: https://www.kaggle.com/datasets/aungpyaeap/fish-market.

Objectives

- *Prepare data for analysis and prediction using multiple linear regression*
- *Explore correlation between continuous variables and the difference in distribution of a continuous variable amongst different categories*
- *Create and assess multiple linear regression models*
- *Assess and investigate multiple linear regression assumptions*

Deliverables

- *Jupyter Notebook (Python) or R Markdown file (R) with all code and results embedded. Files can be rendered to HTML webpages if your instructor requires this. Questions should be stated and answered within Markdown cells.*

Background Questions

Question 1: Provide a general multiple linear regression equation and explain all the terms. (5 Points)

Question 2: Explain the concept of a dummy variable and how such variables are calculated. Why is it necessary to convert nominal variables to dummy variables when performing linear regression? (5 Points)

Question 3. State all the assumptions of multiple linear regression. (10 Points)

Question 4. Explain the concept of homoscedasticity as it relates to multiple linear regression. (5 Points)

Question 5. Explain the purpose and interpretation of a QQ-Plot. What is a QQ-Plot used for when assessing assumptions of multiple linear regression? (5 Points)

Question 6. Explain the difference between R-Squared and Adjusted R-Squared. Why is Adjusted R-squared required when assessing multiple linear regression models? (5 Points)

Tasks and Questions

This assignment can be conducted using either Python or R, whichever you prefer or whichever you instructor requires. Generate code to perform the following analyses and answer the stated questions.

Task 1: Create separate scatterplots to compare the fish weight ("Weight") to each of the continuous independent variables: "Length1", "Length2", "Length3", "Height", and "Width". (5 Points)

Task 2: Create a grouped box plot to compare the distribution of weight amongst the seven different species included in the dataset. (5 Points)

Task 3: Recreate the scatterplots from Task 1. However, this time differentiate the different species using the color of each data point. (5 Points)

Question 7. Discuss the relationships between the dependent variable and each independent variable based on the scatterplots and grouped boxplot obtained. Does each independent variable appear to be correlated with fish weight? If so, is the correlation positive or negative? Is the correlation linear? Does the distribution of weight seem to vary by species? Does the relationship between weight and each continuous variable appear to be different between species? (10 Points)

Task 4. Create dummy variables for the species variable ("Species"). (5 Points)

Task 5. Fit a multiple linear regression model that predicts "Weight" using the available continuous variables ("Length1", "Length2", "Length3", "Height", and "Weight") and the dummy variables representing the species ("Species"). (5 Points)

Task 6. Calculate or obtain the Adjusted R-Squared and RMSE metrics for the model. (5 Points)

Question 8. Describe the model performance using the obtained Adjusted R-Squared and RMSE values. Also, discuss the coefficients obtained. What variable coefficients were found to be statistically significant in the model? (10 Points)

Task 7. Calculate a QQ-Plot to assess the model assumptions. (5 Points)

Task 8. Create a plot with the “Weight” values mapped to the x-axis and the residuals mapped to the y-axis to assess homoscedasticity. (5 Points)

Question 9. Discuss the graphs obtained in Task 7 and 8. Do these graphs suggest issues of normality of the residuals and/or homoscedasticity in the model? (5 Points)