Final Project - STAT 350

Michael Zaghi

December 08, 2020

Abstract

The purpose of this analysis is to construct a multiple linear regression model that best predicts fish *weight* using the provided data set '*Fish Market*'. The regression analysis begins with examining descriptive statistics and cleaning the data. This is followed by residual analysis and check for multicollinearity. Once and adequate model has been established, model selection is performed using forward selection. Finally, the model's predictive performance is measured using cross-validation. The conclusion is that fish *weight* for the species in the data set can be explained almost completely by the selected model with an R^2 of 0.9866.

Contents

1 Introduction

The question of interest is determining how well fish *weight* can be predicted given the data set '*Fish Market*'. This prediction is only valid for the fish species in the data set. Since there are several variables, a multiple linear regression (MLR) model is appropriate. *Weight* is chosen as the response variable since it seems like a metric one would naturally want to predict. Fish are priced by weight, and the quality of a catch is often based on the weight of the fish more than anything else.

2 Data Description

2.1 Overview

The data consists of one qualitative variable fish species (categories *Bream*, *Roach*, *Whitefish*, *Parkki*, *Perch*, *Pike*, and *Smelt*), and six qualitative variables *Weight* (grams), *Length1* (Standard Length), *Length2* (Fork Length), *Length3* (Total Length), *Height*, and *Width*. *Weight* is the response verable and the remaining variables are the predictors. Note that the provider does not explicitly give any units of measure for the data set. These units are an educated guess in order to give context to the data and discussion.

2.2 Data Cleaning

The provided data set is clean with the exception of observation 42 with a weight of 0. Since the remaining data on this observation seemed fine, linear interpolation was used to estimate its weight. No other incomplete, corrupt, or otherwise incorrect data was present.

2.3 Descriptive Statistics

The purpose of analyzing descriptive statistics is to get a better understanding of the raw data before the regression analysis is conducted. This allows for the determination of any possible patterns or inconsistencies that may impact multicollinearity or the models risiduals.

3 Methods

4 Results

5 Conclusion

6 Appendix