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 in cm), *Length2* (Fork Length in cm), *Length3* (Total Length in cm), *Height in cm*, and *Width in cm*. *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 model risiduals.

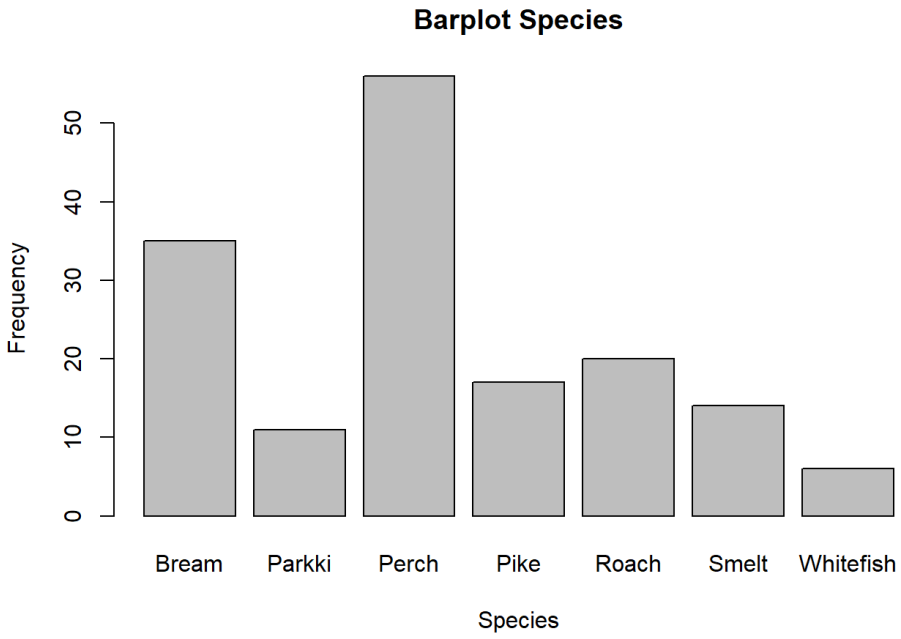

Figure 1: *A barplot with all seven fish spcecies. Perch has the largest frequency of samples while Whitefish and Parkki has very few samples.*

The qualitative variable *Species* is transformed into indicator variables so that it can be incorporated into the regression analysis. *Figure 1* shows that there are many more *Perch* samples than other fish species in the data set. There is a possibility that an indicator variable like *Perch* carries more statistical power (greater probability of a statistically significant result) over a low sample size indicator variable like *Whitefish*.

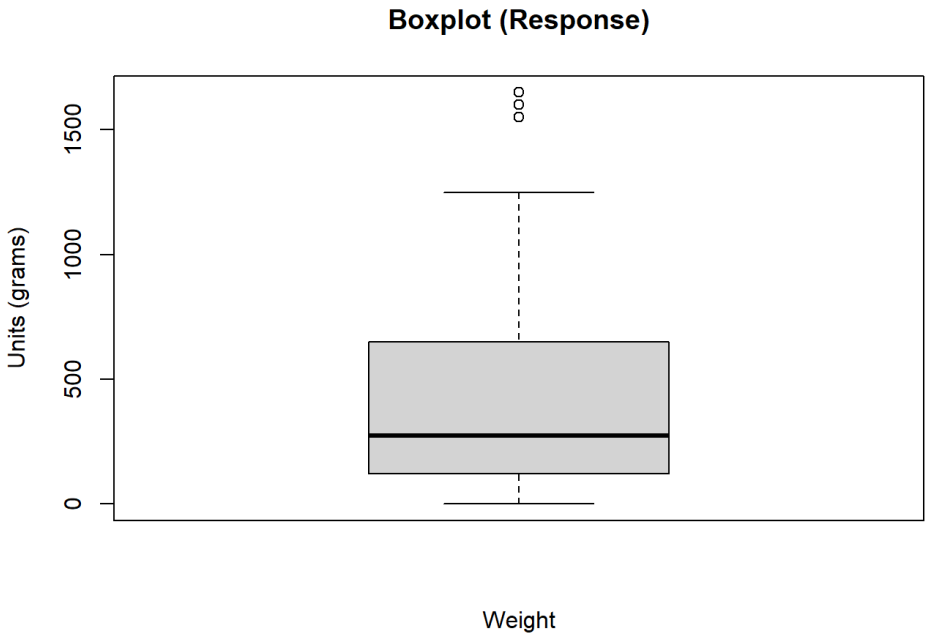

Figure 2: *A boxplot of for the response variable weight. The distribution is skewed to the right with three outliers.*

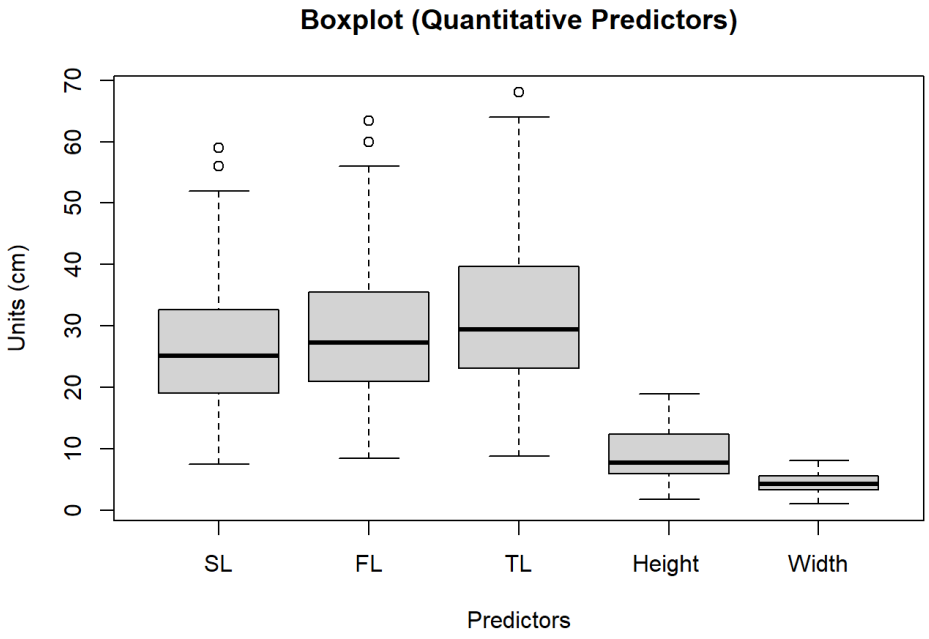

Figure 3: *A boxplot of for the quantitative predictors. The distributions are fairly normal with SL (Standard Length), FL (Fork Length), and TL (Total Length) having some outliers.*

The boxplot's (*Figure 2* and *Figure 3*) confirm that the predictors and response are on a different scale (units of measure). It is important to note that SL, FL, and TL have almost identical distributions, but are centred around slightly different means. This makes sense since these are all three different ways of measuring length.

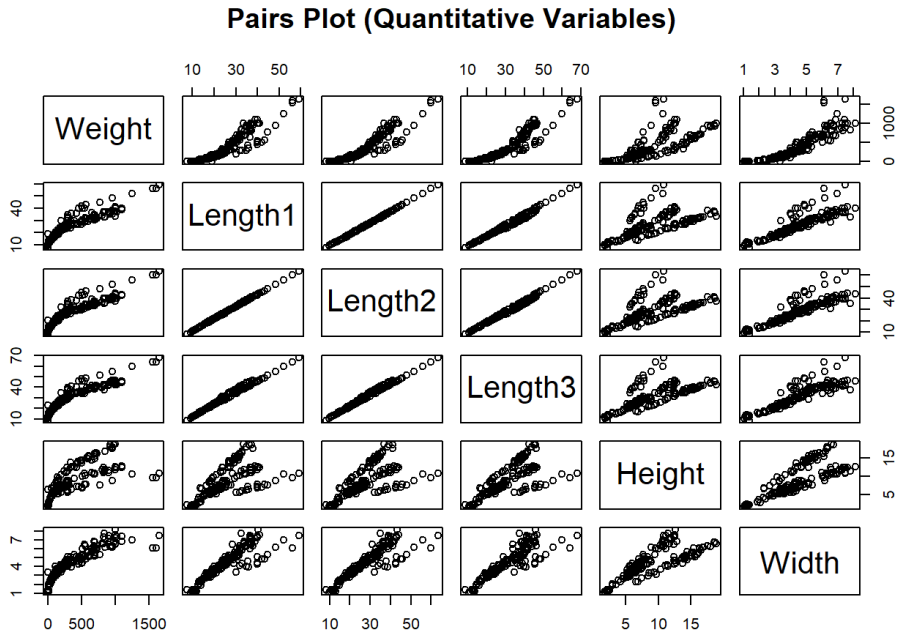

Figure 4: *A pairs plot for quantitative variables. A plot's x and y axis are the intersection of two diagonal elements.*

The pairs plot (*Figure 4*), shows some challenges. Multicollinearity will be an issue with SL (Length1), FL (Length2), and TL (Length3) since these predictors are all colinear with each other. Height and Width may also have issues with colinearity. On a positive note, the response variable Weight has a relationship with all of the predictors. Since it is not a linear relationship, the boxplot (*Figure 2*) indicates that a square root, cube root, or log transformation may be necessary on the response given the right skewness of the distribution.

3 Methods

3.1 Multicollinearity

Multicollinearity is defined as a linear or near linear dependence between predictors. Large multicollinearity in the model can reduce and undermine the statistical significance of the predictors. *VIF* (Variance Inflation Factors) are one way to measure and establish the presence of multicollinearity. Using an inital linear model with unit normal scaling (see linear model output in *Appendix - Checking for Multicollinearity*), two main symptoms of multicollinearity arise. First, the coefficients for *Length2* and *Length3* are negative. This is counter-intutitive because you would expect length to be postively related to weight. Second, these predictors have lower statistical signifcance than expected.

Computing the VIF for each predictor confirms this issue. A VIF of over 10 is generally considered high, and lengths have a VIF of 908.204, 2894.694 and 958.560 respectively. Two of these lengths are redundant and are reducing the explanatory power of the model. Therefore, any two is the best option. After leaving Length1 (SL) in the model, the VIF on the predictors drops to a normal range. Note that Length1 and Height still have high VIF (22.433 and 21.136 respectively), but since both variables seem important, they can be left in until the variable selection process and are not an immediate concern.

3.2 Risidual Analysis

A risidual analysis needs to be conducted so that the model assumptions hold (see *see Appendix - Residual Analysis* for a list of the model assumptions) and the model can be useful for hypothesis testing and prediction. The main method for assessing the appropriateness of the linear regression model is done through *risidual plot graphs* (see *Figure 5*).

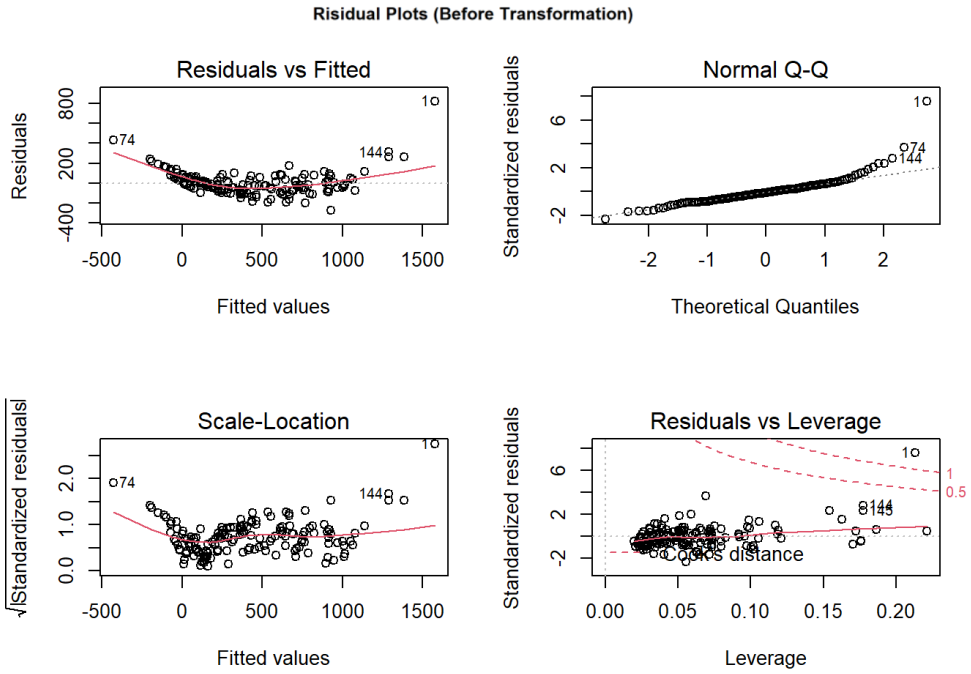

Figure 5: *Residual plots used for determining whether model assumptions have been met.*

3.2.1 Residuals vs Fitted

Residuals vs Fitted describes the relationship between the response (wight in this case) and the residuals. Since the relationship is not linear, the model assumption that states *the relationship between the response y and the regression must be linear*, does not hold. A square root transformation on the response resolves this issue. The R^2 increasing from 0.931 to 0.9873 confirms an improvement in the linear relationship. *Figure 6* shows that drawing a horizontal line at zero seems reasonable and that there is no longer a pattern indicating correlated residuals.

3.2.2 Normal Q-Q

The Normal Q-Q plot shows a huge outlier (6+ standard deviations from the mean residual) which needs to be scrutinized. Observation 1 is the data point that was artificially added to the model. It is a Perch that has for example, a weight to length ratio that is over two times greater than the next largest Perch. There are many possible explanations for why such an observation could occur, ranging from a fisherman wanting to get more money for his catch to a simple

data entry error. Other than this outlier, the residuals are fairly normal and lie close to the $x = y$ diagonal. The one downside of the square root transformation is that the residual of observation 15 was magnified due to it's relatively low weight.

3.2.3 Scale-Location

The Scale-Location plot shows whetther the risiduals ϵ have some contstant variance σ^2. Before the transformation (*Figure 5*), there is a definite non-linear pattern. After the transformation (*Figure 6*), no definite pattern exists and the error terms have constant variance.

3.2.4 Residuals vs Leverage

The Residuals vs Leverage plot shows which observations are negatively impacting the explanatory power of the model due to high leverage and residuals. Here, the high Cook's D value (greater than one) on observation 1 makes it an influential point. This means that it is significantly moving the regression line away from the best fit for the majority of the data points. For this reason, as well as the reason mentioned in 3.2.2, observation 1 will be removed.

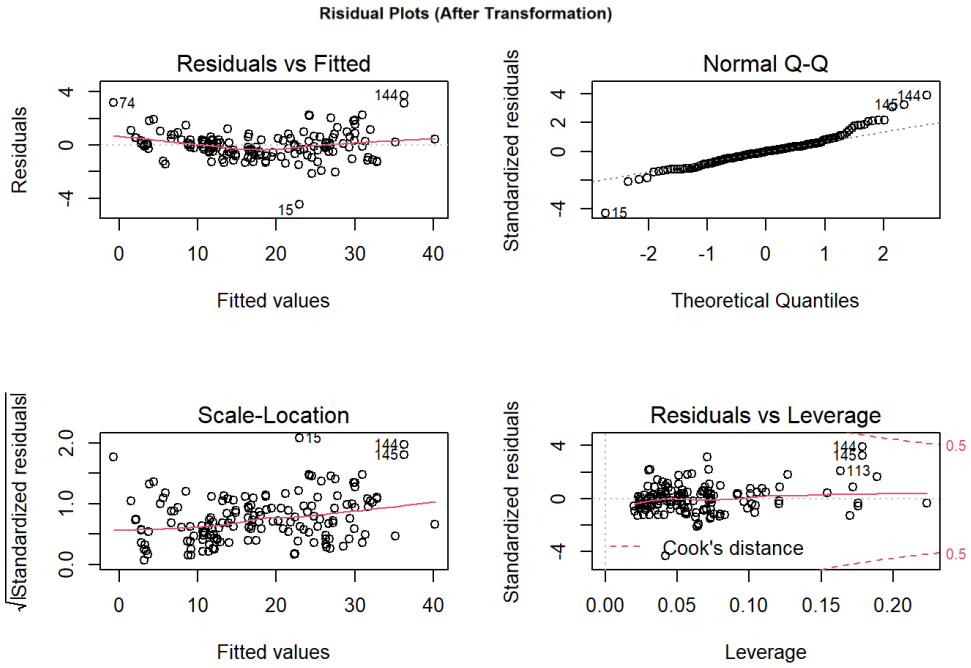

Figure 6: *Residual plots after model assumptions have been met.*

3.2.5 Risidual vs Index

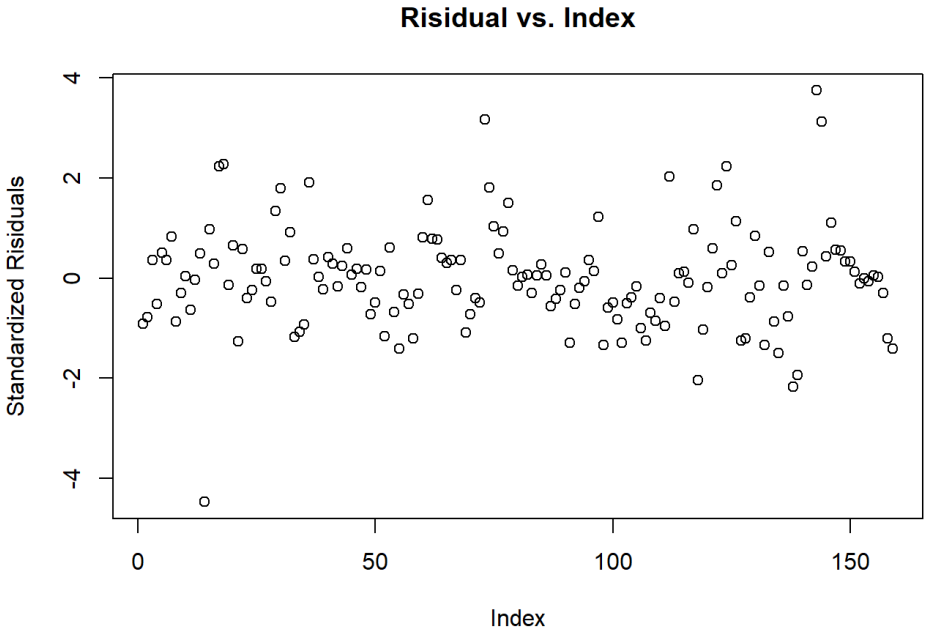

Figure 7: *Risidual vs Index is the risiduals plotted against each observations index (after the square root transformation on Weight).*

In the Risidual vs Index plot, the risiduals are randomly scattered around around $\epsilon = 0$. This indicates that there is no correlation between the errors.

3.3 Variable Selection

The objective with variable selection is to balance the best possible model (the most variables is the most realistic) and the simplest possible model (including only the variables needed). Interaction terms have been omitted from the model from the start due to the fact that it adds unnecessary complexity for only a marginal gain (R^2 is already 0.9873). The initial model is still fairly complex (many indicator variables), and the goal here is to further reduce the complexity of the model without sacrificing explanatory power. The method choosen to do perform variable selection is *forward selection* implemented using p-values and a at a critical value of $\alpha = 0.05$. R^2 is the main criterion, but others like AIC (*Akaike Information Criterion*) and RMSE (*Root Mean Square Deviation*) are also considered.

3.4 Model Validation

The objective in model validation is to evaluate the model performance. Cross validation was used with an 80% training set and a 20% test set. Since the data set is not large, there is a risk that observations with a large absolute values could have been included only in the test set or training set (creating bias). To confirm the results of the initial cross validation, it is iterated on 20 times using the same weights and random observations. The criterion for measuring the performance of the predictions are R^2, MAE (mean absolute error), and RMSE (root mean squared error).

4 Results

4.1 Square Root Transformation

Model adequacy was establed through a square root transformation on the response Weight during the residual analysis. Even though the residual on observation 15 increased because of the transformation, it's Cook's D value was still acceptable since it was not influential. The square root transformation had the added benefit of increasing the initial model's R^2 from 0.931 to 0.9873.

4.2 Variable Selection

In general, all of these criterion track each other over each iteration. The final model includes *Width*, *Length1*, *Height*, *Pike*, *Smelt*, and *Whitefish* (can be found in *Appendix - Variable Selection*). Recall that the VIF of Width and Length1 had high VIF. This is still the case, but it is clearly not causing issues for the model. The resulting R^2 of 0.9871 is basically the same as before the variable selection, but the complexity of the model has been reduced.

4.3 Model Validation

Using the training set model to predict the test set data resulted in an R^2 of 0.988, and RMSE of 1.069, and an MAE of 0.843. This is only a difference of 0.001 for the R^2 and a 0.011 difference on the RMSE compared to the model. This is a good result because it means the model was able to accurately predict the Weight of fish for observations it did not train on. The prediction error of the model is 0.059 grams on average. The performance of the model was confirmed by iterating over random test and training sets 20 times. The results were a similar a similar R^2 and RMSE, but a significantly lower MAE of 0.758.

5 Conclusion

6 Appendix