STAT7016 Final Project

Age Prediction for Abalone by Bayesian Analysis

Zihao Li (u7031432)

Introduction

* 1. Preface

In the recent few years, abalones are a highly-valued and popular seafood in Asia. As the growth of demand increases rapidly, lots of gastronomes and seafood lovers are looking for those abalones with higher quality. One of the judgment criteria is to see the age of the abalone. The old-fashioned way of estimating the age of abalone is to count the rings on the shell and it’s obviously time-consuming and complicated (must remove the shell first). So, as an enthusiastic seafood lover (especially abalone), I wish to use the knowledge of Bayesian Inference I learned to contribute on assessing the age of the abalone purely on its physical measurements.

* 1. Objective

The objective of this project is to build an optimal model for predicting the age of abalone from their physical characteristics using the 1994 original data from “The population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H.rubra) from the North Coast and Islands of Bass Strait.” This dataset is sourced from UC Irvine Machine Learning Repository <https://archive.ics.uci.edu/ml/index.php>. The main goal of this project is to perform a model selection using Bayesian approach and assess the relative importance of the variables which will be stated below. Moreover, I’m interested to assess a model with all main effects and some two-way interactions included.

* 1. Data Sets

There are totally 4177 observations, 8 descriptive features and 1 response feature in the abalone dataset.

* + 1. Response feature

‘Rings’ is an integer type variable range from 1 to 29. It indicates the rings of the abalone which is used to describe the age of abalone. Adding 1.5 gives the age in years.

* + 1. Descriptive features

Name / Data Type / Measurement Unit / Description  
-----------------------------  
Sex / nominal / -- / M, F, and I (infant)  
Length / continuous / mm / Longest shell measurement  
Diameter / continuous / mm / perpendicular to length  
Height / continuous / mm / with meat in shell  
Whole weight / continuous / grams / whole abalone  
Shucked weight / continuous / grams / weight of meat  
Viscera weight / continuous / grams / gut weight (after bleeding)  
Shell weight / continuous / grams / after being dried  
Rings / integer / -- / +1.5 gives the age in years

Methodology

2.1 Variable Analysis

In the abalone dataset, we are going to use ‘Rings’ as the response variable to predict the age of abalone and rest of the variables as predictors. Since ‘Sex’ is a categorical variable with 3 levels (Female, Infant, Male), I need to transfer it to a quantitative variable with values 1,2,3 respectively. As a consideration of sex is not belong to the physical features, we will not add interaction terms involved sex into the implementation. For the response variable, ‘Rings’ is a discrete integer from 1 to 29, and the distribution of the values is not balance. Its values and the predictors’ values are also not in the same scale. So, I decide to apply min-max normalization to the dataset. In addition, all the interaction terms (total 21 terms) will also be taken into consideration.

2.2 Prior Distribution and Sampling Model