Project Proposal

Hongyi Deng Chengkai Shi

November 27, 2023

1 Group Membership

There are two members in our group.

Table 1: Group Members

Name	UCD ID Number	Email Address
Hongyi Deng	922850710	hdeng@ucdavis.edu
Chengkai Shi	922813391	CKShi@ucdavis.edu

2 Data Set to Analyze

We choose one of the data sets posted on Canvas: Abalone Age. The data file is "abalone.txt"

3 Questions of Interest

1. Which of these measurements are significantly related to the number of rings(age)? That is, taking "rings" as the response variable, which X variables should be included into the model?

2. The multicollinearity patterns within the X variables.

3. How to interpret the derived model? That is, how to understand the relationship between age and other variables?

4 Plan for Data Analysis

4.1 General Procedures

1. Exploratory data analysis: Taking "rings" as the response variable, we examine the types of X variables, and find their rough distributions through histograms and pie charts. Moreover, we examine the relationships among X variables through the scatter plot matrix and Box plots. Besides, we detect possible outliers.

2. Model investigation: For one thing, we focus on deciding whether transformations are needed. For another, we examine if there should be interaction terms or high-order power terms. Residual plots, Q-Q plots and Box-Cox tests are used in this part.

3. **Model selection:** Apply stepwise regression to select proper variables, supplemented with significance tests. Criterion such as C_p, AIC_p and BIC_p are considered in this procedure.

4. **Model validation:** Check the model both internally and externally. Criterion like C_p and $MSPE$ are considered. The model is going to be re-fitted if necessary.

4.2 Model Type

Due to the inconsistency of units, we mainly consider standardized linear regression model with 1 categorical variable, and use ordinary least-squares estimation. One important problem is how to deal with the categorical variable "sex". Moreover, interaction or high-order power terms should be added if necessary.

4.3 Potential Pitfalls

1. **Multicollinearity within the X variables.** For example, intuitively, there are likely to be inner correlations between shucked weight and viscera weight. To deal with multicollinearity, we may be more careful in variable selection, or apply biased methods like ridge regression.

2. **Outliers.** We must deal with different kinds of outliers appropriately.

3. **Heteroscedasticity.** The model we derive might contain heteroscedasticity, which contradicts the basic model assumptions. For example, residuals from male samples can have different variances from those of female ones. The weighted least squares method may be useful to deal with this.