**Project proposal**

**Group Membership:**  Xueting Shao , Yingjie Li , Yibing Luo

**Dataset:** Abalone Age

**Question of Interest:靠你们了**

**Plan for Data analysis:**

***Step1: explore the data set***

1. Some statistics methods are applied to obtain a brief image of the data set. For quantitative variable, histogram, boxplot, summary statistics can be used to observe the structures and distribution. For qualitative variables, pie chart, frequency table can be used to do so. Correlation matrix and scatter plot matrix can used to observe the relationships among the variables.
2. Some transformations can be considered for specific variables according to #1(log transformation for right-skewed distribution).
3. Noticed that in data set, sex is a factor. Side-by-side boxplot is applied to decide whether a blocked experiment or merely the dummy variables are used.
4. Under normal error model, the response variable Y (Rings in this case) need to under normal distribution. Boxcox procedure is applied to decide whether Y need to be transformed.

***Step2: Model building***

1. First, the dataset need to be split into two folds, namely training data and test data.
2. Model selection:

* Consider subsets selection from the pool of first-order effects and the pool of first-order as well as second-order separately.
* Best subsets algorithms and stepwise regression procedure are applied to each pool to select the good model.

1. Model validation:

Internal validation:

* assumptions of linear model with normal errors need to be test
* , , , can be used to conduct internal validation of candidate models

External validation:

The test data is used to fit the model. MSPE can be considered as an indicator of the predictive ability of the model.

1. Outlying cases need to be identified: outlying Y observations are tested by Bonferroni’s procedure. And outlying X observations are identified by leverage. Further, influential cases need to be identified. (DFFITS: single fitted value; Cook’s distance: all fitted values; DFBETAS: regression coeffiients )
2. Remedial measures:

Ridge regression (bias-variance trade-off) and Robust regression.

Step3: Ability of model

“After review of the appropriateness of the model and completion of any necessary remedial measure and an evaluation of their effectiveness, inferences based on the model can be undertaken”

上面这段我是抄的书上的。因为我没想Question of Interest. 有了QI 这个应该挺好写的。我想了些大概就是求点C.I， 求点prediction。 我去写141 report了，这部分靠你们了。

=。=