Constructing Predictive Model for Abalone Age

Wenjia Wang

University of California, Davis

***Abstract*— The population of abalones, marine animals with high commercial and recreational values, have been threatened by the ongoing overfishing. A lay off on capturing young abalones could be considered as a measure to alleviate the issue. Estimating the ages of abalones will thus be necessary. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. In this project, I developed a predictive model for abalone age using physical characteristics relating gender, size and weight. With appropriate data manipulations and model selections, I arrived at a relatively decent predictive model.**

I. INTRODUCTION

Abalone is a type of marine snail. Unfortunately, the important commercial and recreational value of abalones put their populations at risk. Overfishing and poaching have caused some species of abalone to be endangered. To remedy the situation, putting regulations on harvesting young individuals would be a suitable approach, and elucidating the ages of abalones is critical for this task. The traditional method, counting the number of rings through a microscope to determine the age, is tedious and time-consuming. Thus, a predictive model based on physical characteristics, which are easier to obtain, is desirable. Linear models are my main candidate models in this report. The pipeline of my model constructing is data exploration, preliminary analysis, model selection and conclusion.

II. DATA EXPLORATION

This section is to figure out the data structure and roughly explored the potential relationship among variables. Appropriate transformation will also be implemented for future model construction.

A. Attribute Information

The data set contains nine attributes of abalones: Sex (male, female, or infant), length of the shell, diameter of the shell, height, whole weight, shucked weight, viscera weight, shell weight and the rings. These attributes were measured on 4177 abalones. In the future model constructing and selecting, the rings of abalones is response variable, which is integer, representing 28 different classes. Except Sex, the other attributes are continuous, and all of them are predictors in my future analysis. Additional, there is no missing value in the 4177 observations.

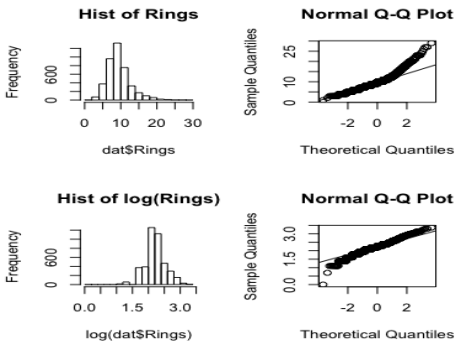


Fig. 1. The upper row is the hist gram and normal quantile plot of the original response. The second row is the hist gram and normal quantile plot of the response after logarithm transformation.

B. Distribution of Variables and Transformation

In multiple linear regression models, all continuous variables are expected to be normally distributed or at least symmetric. In particular, normal distribution of the response variable is assumed in linear regression.

Figure A1 (see Appendix 1) shows that the three different categories in the variable "Sex" is distributed evenly across the data set, which is desirable. However, the hist gram (Figure 1) of the response variable "Rings" in the original data skews to the right. Hence, a transformation is required to satisfy the normal error assumption of multiple linear model. The box-cox transformation suggests a logarithmic transformation for the response variable (Figure 2). The histogram and the normal probability plot (Figure 1) of the log transformation show no severe departures from normal distribution.

C. Correlation Among Variables

Correlation analysis first aims to check whether linear relationship exists and the reasonability of linear models, and subsequently guides the construction of model candidate pool.

Box plots are used to explore the relationship between the categorical variable "Sex" to each of the other variable. From the box plots (Figure 3), we can infer that the factor "Male" and "Female" show very similar distributions in other variables. This may suggest "Male" and "Female" may be merged as one category. The levels of "Sex" could be reduced to "Mature" and "Infant".

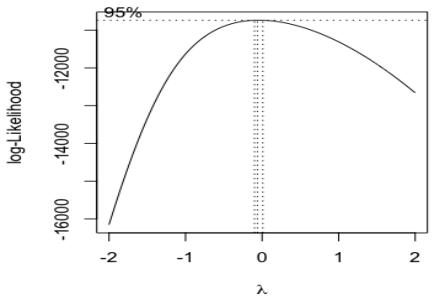


Fig. 2. "boxCox" chose $\lambda = 0$, suggesting logarithm transformation.

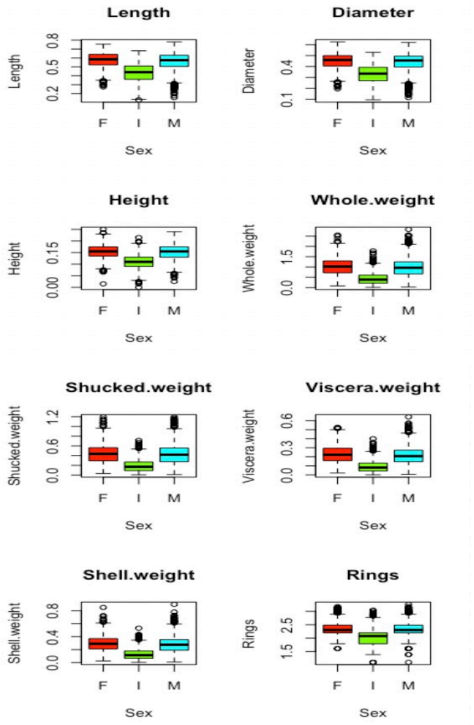


Fig. 3. "Female": red; "Infant": green; "Male": blue.

The scatter plots of all the pairs of the quantitative variables are presented in Figure A2 (see Appendix 1). The correlation matrix with exact correlation coefficients can be found in Figure 4. The two figures indicate a linearity among the response variable and the predictors. The scatter plots suggest there might be a curvilinear relationship between the variable "Rings" and the four weight variables. Moreover, high collinearity among the predictors is clearly presented. Multiple predictors have correlation coefficients over 0.9.

III. Preliminary Analysis

The preliminary analysis is based on the first order full model fitted on the whole data set (see Table A in Appendix 1 for the variable assignments):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_9 X_9 + \epsilon \qquad (1)$$

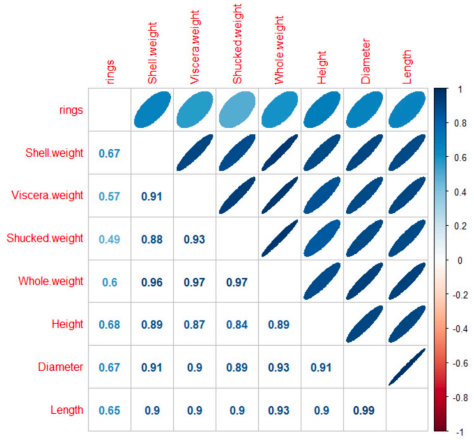


Fig. 4. Correlation coefficients of all pairs of quantitative attributes.

where X_1, X_2 are dummy variables from categorical attribute "Sex", and ϵ represents the random error.

A. Diagnostics

Diagnostics is to check whether the assumptions of the multiple linear regression model (for inference) are satisfied:

- The error or response variable is normally distributed. This has been achieved by applying logarithmic transformation on the response variable.
- The errors have a common variance. This can be confirmed by the scatter plot of residuals against fitted values.

The residual plots (Figure A4 in Appendix 1) display the residuals versus each of the predictors. In each plot, the points are roughly symmetrically and randomly distributed around zero, which implies constant variance holds. In addition, there is no significant pattern in the residual plots, so the quadratic or cubic terms of the predictors may not be helpful. The normal quantile-quantile plot (Figure A5 in Appendix 1) indicates no severe departure from normality.

B. Identify Outlying Cases

The plot of residuals versus leverage (see Figure 5) suggests that the influence of outlying cases exists.

Three steps to identify influential outlying cases: (1) identify outliers in the response by deleted studentized residuals; (2) identify outliers in the predictors by leverage values; (3) identify influential outlying cases by Cook's distance.

The deleted studentized residual of the ith case can be expressed as:

$$t_i = \frac{Y_i - \hat{Y}_{i(j)}}{\sqrt{MSE_{(i)}/(1 - h_{ii})}}, \qquad (2)$$

where $\hat{Y}_{i(j)}$ is the predicted value by the fitted regression function based on all data except the ith observation, $MSE_{(i)}$ is the MSE of the regression fit based on cases excluding case i, and h_{ii} is the i-i entry of the design matrix. In

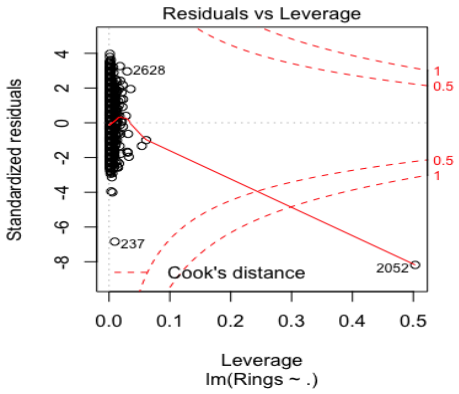


Fig. 5. Studentized residuals versus leverage of the first order full model based on the whole data set.

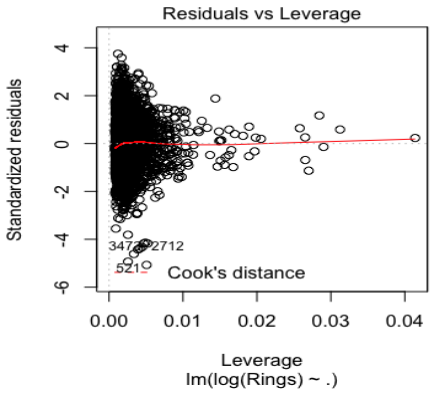


Fig. 6. Studentized residuals versus leverage of the first order full model based on the whole data excluding identified influential outlying cases.

fact studentized deleted residuals can be computed from the regression fit based on all cases:

$$t_i = (Y_i - \hat{Y}_i)\sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - (Y_i - \hat{Y}_i)^2}}, \quad (3)$$

where n is the number of the observations and $p-1$ is the number of predictors. Under null hypothesis: the model is correct and all cases follow the model, we have t_i follow t distribution with degree of freedom $n-p-1$. Given significance level $\alpha = 0.1$, the Bonferroni's procedure identifies cases with $\|t_i\| > t(1-\alpha/(2n); n-p-1)$ as outlying Y observations. In my data set, it turns out that the 237th and 2052nd observations are outlying in Y.

The outlying cases in terms of the predictors are identified by leverage values, ie. h_{ii}. In fact, h_{ii} reflects the Mahalanobis distance between the predictor values of the ith case and the sample mean of the predictor values. Thus, a large leverage value is an indication of outlying case. In my data, The outlying case is identified if its leverage value is twice larger than the mean leverage value h, equal to p/n. It turns out that 276 outlying cases containing the 237th and 2052nd observations.

The final step is to determine whether the outlying cases detected above are influential in determining the fitted regression model. Cook's distance measures the aggregate influence on all fitted values that is made by the omission of a single case in the fitting process. It can be computed from the regression fit based on all cases by expression:

$$D_i = \frac{(Y_i - \hat{Y}_i)^2}{p \times MSE} \frac{h_{ii}}{(1-h_{ii})^2}. \quad (4)$$

Note that the expectation of D_i is approximate to $h_{ii}/p(1-h_{ii})$. Since if case i follows the same regression relation as other cases, ie. $h_{ii} = p/n$, the the expectation will close to $1/n-p$. Thus, I used $4/n-p$ as a threshold for Cook's distance to identify influential cases. Eventually, 100 influential outlying cases are removed. Figure 6 shows a significant improvement compared to Figure 5.

IV. Model Selection

Model selections are executed to balance the variance bias trade-off. The optimal model should fit the data well (low bias) and have little model variance. That is, we need to include the necessary predictors while keeping the number of predictors low. The previous exploration of the data structure indicates severe multicollinearity, which generally incurs inflated variance and makes the model prone to overfitting. Therefore, model selections should be applied cautiously to remedy this possible issue.

The original data is split into two parts: 80 percentage of which were assigned to the training set, and 20 percentage of which were assigned to the validation set. The training set will be used for model fitting and model selecting, and the validation set will be used for validating the final model. Since the main purpose is to find a model to predict the age effectively, the predictive ability of the model is the main concern. Mean squared prediction errors (MSPE) of the validation data set will be used to show the predictive ability of the final model.

A. Best Subsets Procedure

Compared with stepwise model selection procedure, best subsets selection will exhaustively check and compare all subsets of the predictors pool. On the other hand, this method becomes not applicable when the pool of potential predictor variables is large. However, there are only eight potential attributes involved to build a predictive model. Therefore, best subsets method is feasible.

I started by using best subset selection to find the optimal option for each choice of the number of pre- dictors, denoted as k, which would be included in the model (see TABLE I). Then I compared the optimal options over k by using multiple criteria: R^2, adjR^2, sse/n, C_p, BIC, AIC and GCV (generalized leave one out cross validation).

R^2, adjR^2 and sse/n can measure how much variation in Y which can be explained by the model. Mallows' C_p criteria is based on unbiased estimation of the L^2 risk while GCV

TABLE I

OPTIMAL OPTION FOR EACH K

Size k	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1					*				
2							*		*
3				*		*	*		
4				*	*	*	*		
5	*			*	*	*	*		
6	*			*	*	*	*	*	
7	*			*	*	*	*	*	*
8	*		*	*	*	*	*	*	*
9	*	*	*	*	*	*	*	*	*

TABLE II

COMPARISON OF OPTIMAL OPTIONS

Size k	R^2	**adj** R^2	sse/n	C_p	**BIC**	**AIC**	**GCV**
1	0.4510	0.4508	0.0559	1307.2104	-1940.846	-9607.365	0.0527
2	0.4913	0.4910	0.0518	973.7186	-2181.780	-9854.390	0.0488
3	0.5671	0.5667	0.0440	345.2236	-2700.285	-10378.986	0.0416
4	0.5842	0.5837	0.0423	204.8724	-2823.835	-10508.626	0.0400
5	0.5964	0.5958	0.0411	105.0972	-2913.227	-10604.109	0.0388
6	0.6077	0.6070	0.0399	12.9390	-2997.992*	-10694.964	0.0378
7	0.6084	0.6076	0.0398*	9.3190	-2995.528	-10698.591	0.0377*
8	0.6088*	0.6079*	0.0398	8.0064*	-2990.758	-10699.912*	0.0377
9	0.6088	0.6077	0.0398	10.0000	-2982.674	-10697.918	0.0377

criteria tries to find the model with the smallest predictive rist, ie. the expected L^2 loss of predicting a new observation. Note that if a model has no (in-sample) bias, then $E(C_p)$ is approximate to p, ie. the number of predictors including in the model, otherwise, $E(C_p)$ tends to be larger than p. AIC criteria aims to find the model that is closest to the true data generating mechanism where the closeness is defined in terms of the Kullback-Leibler divergence of the fitted model with respect to the true distribution of the response Y. In contrast, BIC imposes a prior distribution on the class of submodels, and aims to find the submodel that has the highest posterior probability. We should look for models with small AIC or BIC.

TABLE II shows the values of multiple criteria for the best subset over different sizes. Though the model including eight predictors minimizes the values of most criteria, I tended to choose a smaller model guided by BIC criteria since overfitting is the main concern. Thus, the selected model involves six predictors. Note that the categorical attribute "Sex" with three categories is modeled as two dummy predictors X_1 and X_2, while only one of them is kept in the selected model. This result is consistent with my proposition in the section of correlation analysis that the two levels "Female" and "Male" can be merged together. The fitted model is expressed as $\hat{Y} = 1.2438 + 0.0948X_1 + 1.7896X_4 + 2.5024X_5 + 1.0614X_6 - 2.2036X_7 - 1.2646X_8$. Apply the fitted model to validation data set, and the MSPE turns out to be 0.0353. Notice that the MSPE is smaller than sse/n=0.0399, which implies that there is no significant overfitting.

B. Ridge Regression

To remedy the severe multicollinearity in the predictors, some linear shrinkage procedures are implemented. Ridge

TABLE III

CORRELATION COEFFICIENT BETWEEN LEFT-SINGULAR VECTORS AND Y

	1st Vec	**2nd Vec**	**3rd Vec**	**4th Vec**	**5th Vec**	**6th Vec**	**7th Vec**	**8th Vec**	**9th Vec**	**10th Vec**
$\sqrt{\delta_j}$	94.4115	35.7612	24.2037	12.7193	3.1317	2.7197	1.7087	0.9843	0.8386	0.6615
corr	-0.5672	0.5156	-0.3116	0.0547*	0.4546	0.0727	0.0404	0.0820	-0.1131*	0.0153

regression is a penalized least squares regression procedure with a quadratic penalty on the regression coefficient vector β. Ridge regression estimator of β is given by:

$$\hat{\beta}(\lambda) = argmin[\|Y - X\beta\|^2 + \lambda\|\beta\|^2] \tag{5}$$

$\lambda \geq 0$ is a penalty parameter, and if $\lambda = 0$, the ridge regression estimator is the ordinary least squares estimate of β. This method is particularly suitable when there is a degree of multicollinearity in the predictors. Theoretically, the optimal choice of λ can be determined by minimizing the Expected Squared Error of the corresponding estimator ($MSE(\hat{\mu}(\lambda)) = \|\hat{\mu}(\lambda) - \mu\|^2$, where $\hat{\mu}(\lambda) = X\hat{\beta}(\lambda)$, $\mu = E(Y)$).

The range of λ I used is from 0 to 0.1. As a result, based on GCV criteria, the best choice of λ is zero. This result suggests ordinary least squares estimate of β and the penalized regression seems not to be helpful. To figure it out, let's recall the expression of $MSE(\hat{\mu}(\lambda))$:

$$\|\hat{\mu}(\lambda) - \mu\|^2 = \sigma^2 \sum_{j=1}^{p} \frac{\delta_j^2}{(\delta_j + \lambda)^2} + \sum_{j=1}^{p} \frac{\lambda^2}{(\delta_j + \lambda)^2} \theta_j^2, \tag{6}$$

where δ_j is the ith eigenvalue of X^TX, and $\theta = (\theta_1, \ldots, \theta_p)^T = \Delta V^T \beta$, Δ consists of the singular value of X, V consists of the right-singular vector of X.

In equation (6), the first term representing the variance of $\hat{\mu}(\lambda)$, will be smaller for larger value of λ, while the second term representing the bias of $\hat{\mu}(\lambda)$, will get larger for larger λ. Under strong multicollinearity, the potential benefits of ridge regression are on the condition that θ_j will close to zero if δ_j is a quite small eigenvalue. In addition, notice that $\mu = U\theta$, U is the left-singular vector of X. Thus, ridge regression can be quite advantageous compared to standard least squares regression, in terms of l^2 risk of estimating μ, provided (1) there is strong multicollinearity among the predictor variables;(2) μ essentially lies in the subspace spanned by the left- singular vectors associated with the relatively large singular values of X. In my data set, the multicollinearity among predictors is obvious, and thus the second condition may be violated (verified by TABLE III).

C. Principle Components Regression

Principal component regression (PCR) is a method of fitting the regression model through an orthogonal transformation of the columns of the design matrix. This orthogonal transformation is done with respect to the eigenbasis of X^TX, or equivalently, through the singular value decomposition of X. PCR is also closely related to the ridge regression, and in fact can be viewed as a discrete form of regularized regression, where the number of principal

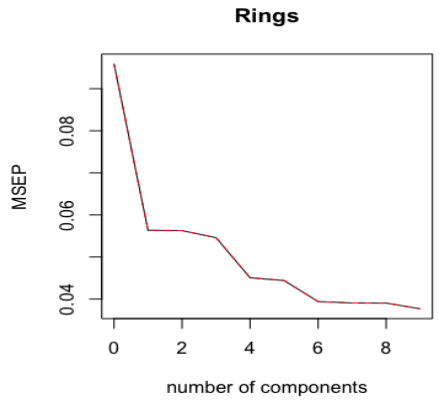


Fig. 7. MSEP based on 10-fold cross validation versus the number of components involved.

TABLE IV
SUMMARY OF "PCR"

	Intercept	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps
CV-MSEP	0.0959	0.0563	0.0563	0.0546	0.0451	0.0444	0.0394	0.0391	0.0390	0.0377*
% Variance (X)	NA	77.24	90.05	94.44	96.7	98.20	99.14	99.81	99.95	100.00
% Variance (Y)	NA	41.28	41.38	43.15	53.1	53.86	59.00	59.41	59.47	60.88

components used in fitting the regression model becomes the regularization parameter.

One approach to formulate PCR is:

$$Y = U\theta + \epsilon, E(\epsilon) = 0, Var(\epsilon) = \sigma^2 I_n \tag{7}$$

This actually uses U, ie. the matrix of left singular vectors of X as the basis for representing $\mu = X\beta$. The least squares estimator of μ, using first k principal components, becomes $\hat{\mu}_{PC}(k) = \sum_{j=1}^{k} \hat{\theta}_j U_{.j}$.

I implemented PCR and choose the optimal k by using 10-fold cross validation. TABLE IV shows the MSEP of the model involving the first k components and accumulative variation that the first k components are able to explain. Figure 7 directly suggests that the MSPE will become minimum when all components are involved, which is equivalent to the standard least squares regression.

The consistent results (suggesting standard lest squares regression) of PCR and ridge regression indicates that instead of overfitting, the first order full model may be unable to sufficiently explain the response variable. Therefore, higher order model should be considered.

D. Models with Interactive Terms

In the section of preliminary analysis, the residuals plots (residuals versus each predictor) did not show significant quadratic pattern, so I mainly consider the potential interactive effects in the second order model. Since the interaction of all pairs of predictors will be considered, the best subset selection is no longer applicable. Stepwise regression procedures will be performed to arrive at a suboptimal model.

Backward stepwise selection is applied with "direction='both'". The initial model includes the six predictors chosen in best subset procedures, which are X_1, X_4, X_5, X_6, X_7, X_8, as well as their pair-wise interactive terms. Based on BIC criteria, the final model includes ten variables: X_1, X_4, X_5, X_6, X_7, X_8, $X1:X7$, $X4:X5$, $X4:X6$, $X4:X7$; The corresponding values of R^2, adjR^2, sse/n, BIC, AIC and GCV are displayed in TABLE V.

TABLE V
SUMMARY OF SELECTED HIGHER ORDER MODEL

R^2	adj R^2	sse/n	BIC	AIC	GCV
0.646	0.645	0.0339	-10955.22	-11022.22	0.0342

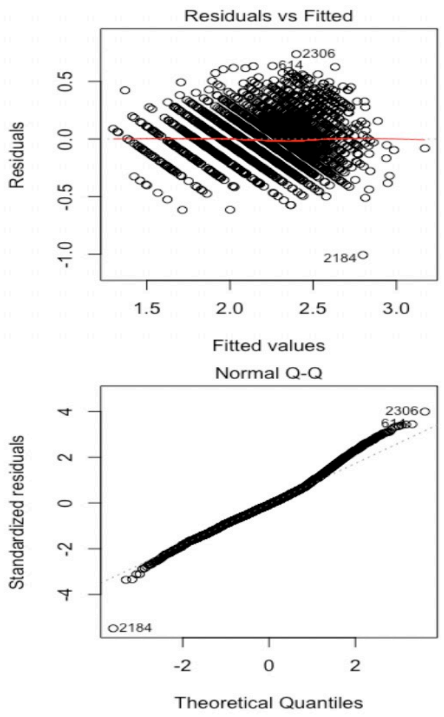


Fig. 8. The upper plot displays the residuals versus fitted values. The bottom is normal quantile-quantile plot.

Compared to TABLE II, the values of all criteria become better. The residuals plot and Q-Q plot of the selected model (see Figure 8) show no objection to the assumptions for linear regression model. Since the purpose is to make prediction based on the selected model, I will not bother to interpret the specific associations between the response and the predictors. Fit model with the selected variables and apply it on validation set, then the MSPE becomes 0.0320, which is smaller than the MSPE obtained by best subsets procedure. Again, the MSPE is smaller than sse/n=0.0339, suggesting no sign of overfitting, and thus more complicated interactions can also be tried.

V. CONCLUSIONS

By leveraging a multiple linear regression, We have built a predictive model for rings of abalones, which can be used to estimate the ages. It turns out that being a male or a female does not influence other characteristics of a abalone; the

only thing we need to consider is its maturity, and this also interacts with "Shucked.weight" in the final model. Besides the original variables, the final model also included some interactive terms. Since our objective is prediction, we will not go at length to interpret the associations. In spite of utilizing a variety methods, the R^2 is still less than 0.65. Here are some limitations of the data and the analysis that we have reasoned:

- High multicollinearity in predictors.
  The correlation coefficient among multiple predictors are over 0.9 (the predictors are more correlated with each other than the response), which reduced the overall ability of the set of predictors to sufficiently explain the response "Rings". In spite of multicollinearity, neither ridge nor PCR shows effectiveness, which also suggests that the set of predictors is not influential enough.
- Not enough influential variables.
  Considering the final model only explaining less than 65% variation in "Rings", more robust predictive model may requires more influential predictors in the dataset. Intuitively, predicting age from these variables is inherently difficult. This situation would be analogous to predicting human ages by size and weight alone.

There is likely room to improve the model. One direction is that more complicated interactions can be involved in the model. In addition, consider the response actually represents various classes, other than using linear regression models, some non-linear classifiers such as classification trees, are worth to attempting.

REFERENCES

[1] Lecture notes in class.

[2] Michael H. Kutner (2004) Applied linear regression models. McGraw-Hill/Irwin 2004.

[3] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2014) An Introduction to Statistical Learning: with Applications in R. SBN:1461471370 9781461471370.

[4] R.Naylor, Literature review of abalone ageing tech- niques, New Zealand Fisheries Assessment Report, 2015.

Appendix 1

Figure A1

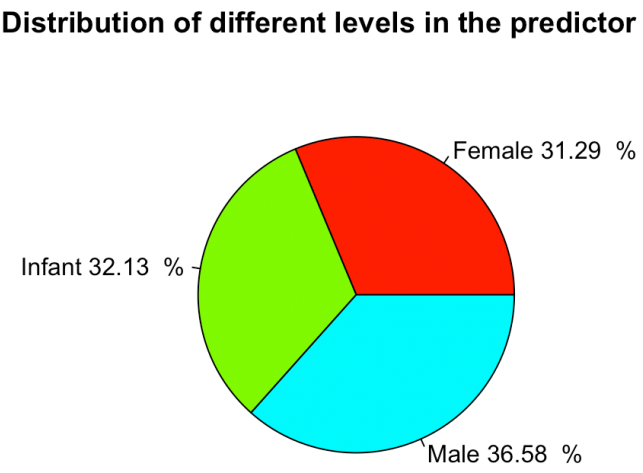


Figure A2

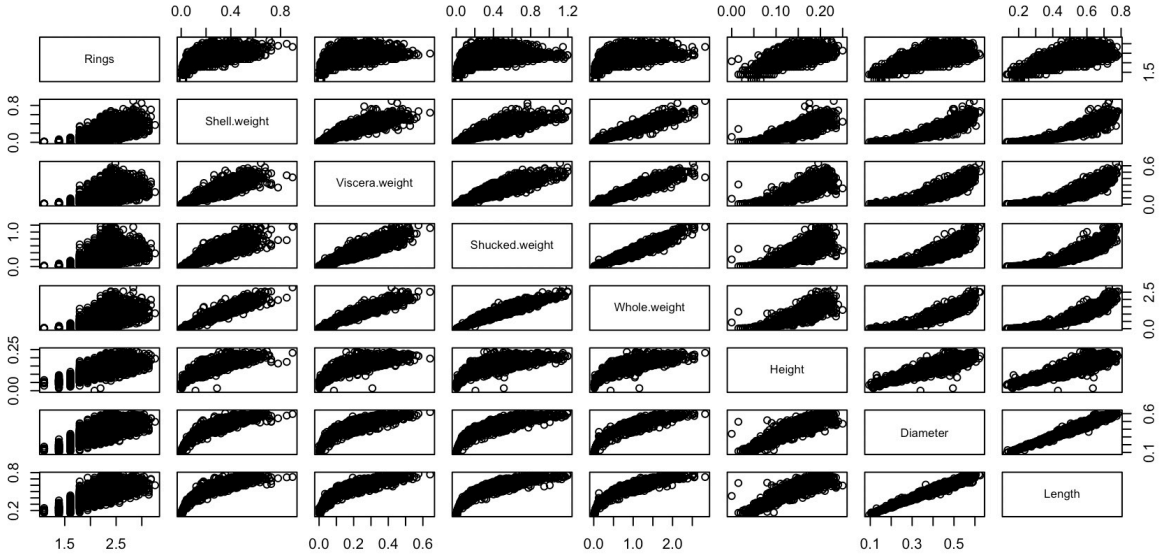


Table A

Name	Variable	Data Type	Description
Rings	Y	Integer	+1.5 gives the age in years
Sex	$X_1(I), X_2(M)$	Categorical	M, F, and I (infant)
Length	X_3	Continuous	Longest shell measurement
Diameter	X_4	Continuous	perpendicular to length
Height	X_5	Continuous	with meat in shell
Whole.weight	X_6	Continuous	whole abalone
Shucked.weight	X_7	Continuous	weight of meat
Viscera.weight	X_8	Continuous	gut weight (after bleeding)
Shell.weight	X_9	Continuous	after being dried

Table: Description of Variables

Figure A3

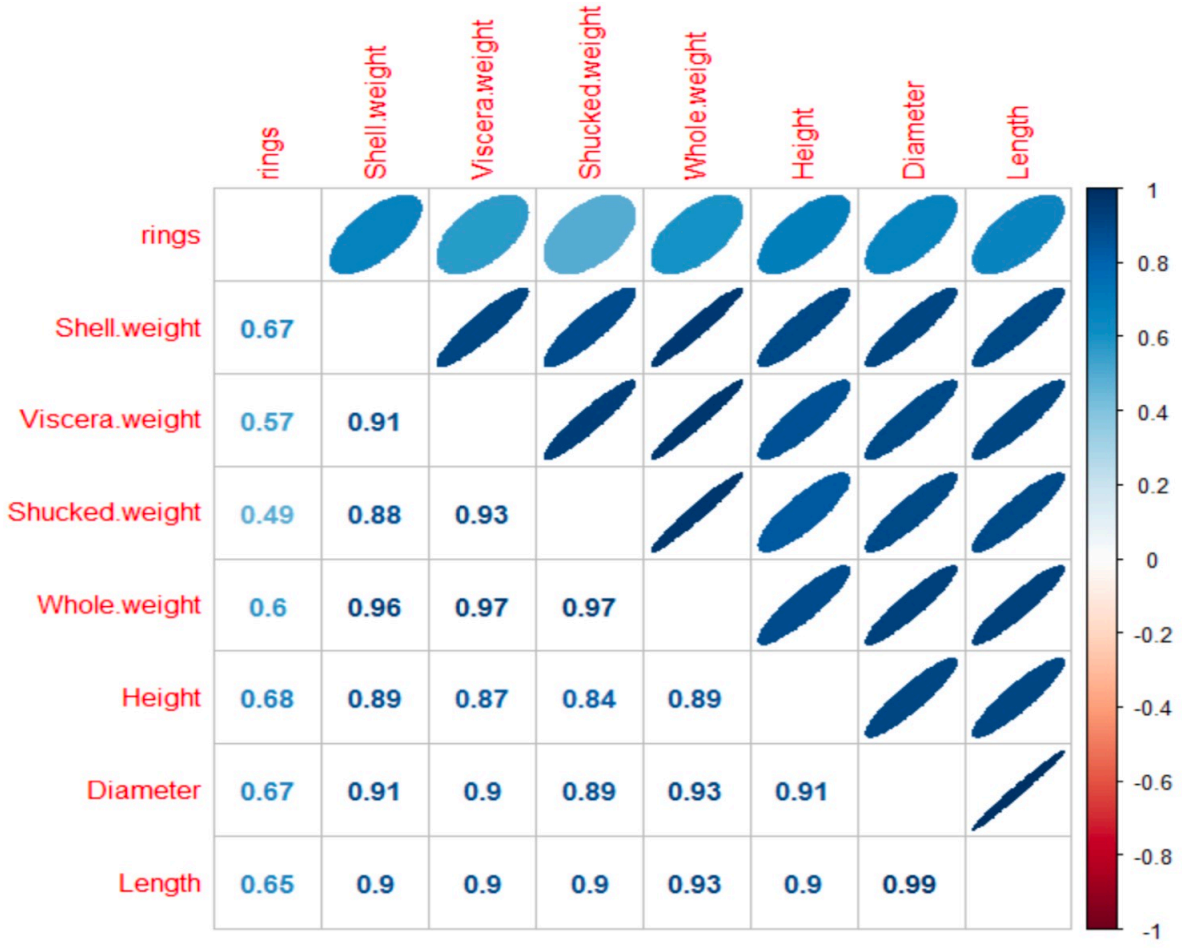


Figure A4

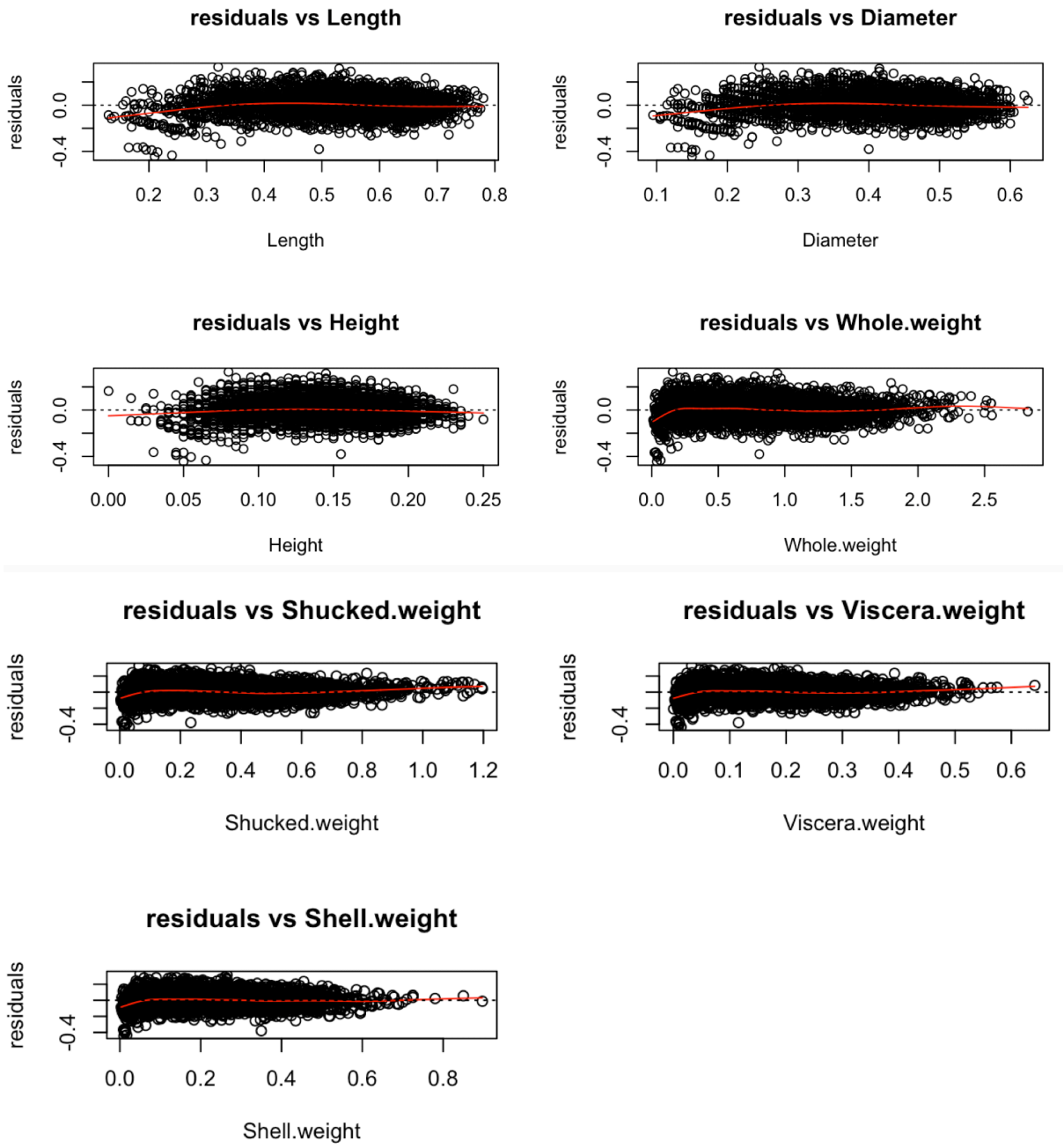
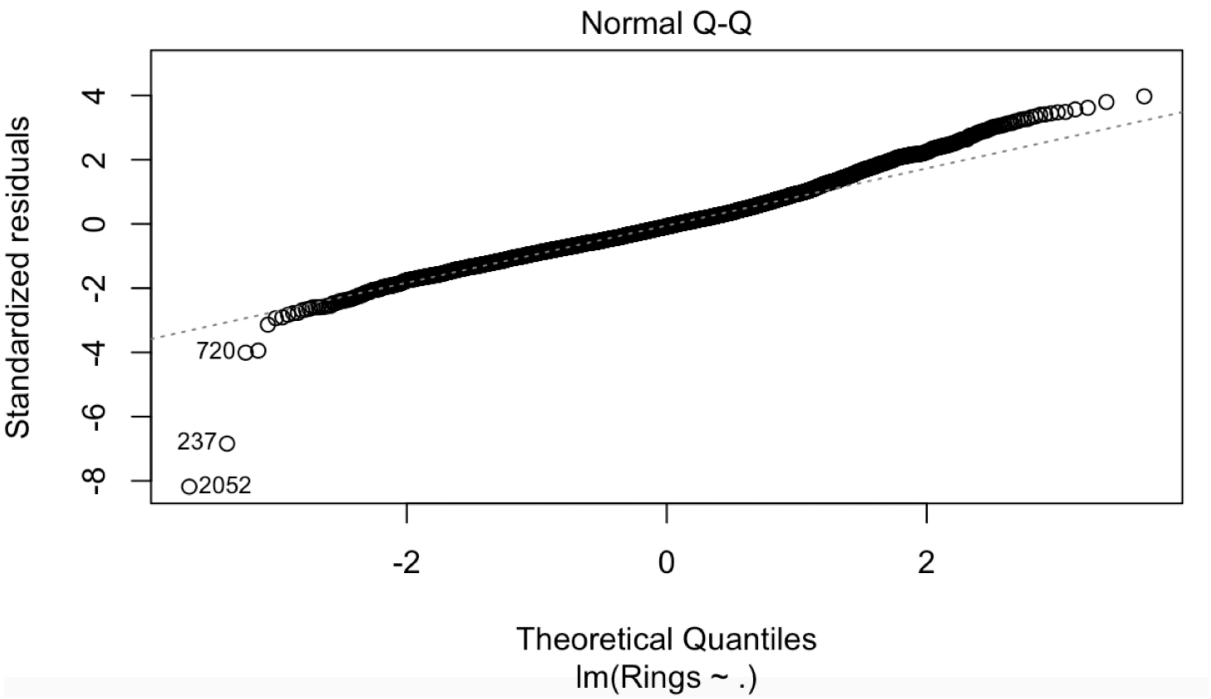

residuals vs Length
residuals
-0.4
0.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Length
residuals vs Diameter
residuals
-0.4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Diameter
residuals vs Height
residuals
-0.4
0.0
0.00
0.05
0.10
0.15
0.20
0.25
Height
residuals vs Whole.weight
residuals
-0.4
0.0
0.0
0.5
1.0
1.5
2.0
2.5
Whole.weight
residuals vs Shucked.weight
residuals
-0.4
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Shucked.weight
residuals vs Viscera.weight
residuals
-0.4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Viscera.weight
residuals vs Shell.weight
residuals
-0.4
0.0
0.2
0.4
0.6
0.8
Shell.weight

Figure A5


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
-6
-8
720
237
2052
-2
0
2
Theoretical Quantiles
lm(Rings ~ .)

Appendix 2

```
library(MASS)
library(corrplot)
library(caret)
library(tidyverse)
library(leaps)
library(glmnet)
library(pls)


dat=read.csv("Downloads/abalone.csv")
head(dat)
dat$Sex=as.factor(dat$Sex)
sapply(dat,class)
# Sex       Length      Diameter        Height   Whole.weight
#"factor"     "numeric"     "numeric"     "numeric"     "numeric"
#Shucked.weight Viscera.weight   Shell.weight         Rings
#"numeric"     "numeric"     "numeric"     "integer"
apply(is.na(dat), MARGIN = 2, sum)
apply(dat==0,2,sum) #two observations with height=0
summary(dat)
n=nrow(dat)

dat %>% group_by(Sex) %>% summarize(percentage = n()/nrow(dat))
lbs <- c("Female", "Infant","Male")
pct <- round(100 * table(dat$Sex)/4177, 2)
lab <- paste(lbs, pct, " %")
pie(table(dat$Sex), labels = lab, col = rainbow(4), main = "Distribution of different levels in
the predictor 'Sex'")
par(mfrow=c(2,2))
hist(dat$Length,main = "Hist gram of 'Length'")
hist(dat$Diameter,main = "Hist gram of 'Diameter'")
hist(dat$Height,main = "Hist gram of 'Height'")
hist(dat$Whole.weight,main = "Hist gram of 'Whole.weight'")
hist(dat$Shucked.weight,main = "Hist gram of 'Shucked.weight'")
hist(dat$Viscera.weight,main = "Hist gram of 'Viscera.weight'")
hist(dat$Shell.weight,main = "Hist gram of 'Shell.weight'")

lm.fit0=lm(data=dat,Rings~.)
par(mfrow=c(2,2))
hist(dat$Rings, main = "Hist of Rings")
qqnorm(dat$Rings)
qqline(dat$Rings)
boxcox(lm.fit0) #logtransformation
hist(log(dat$Rings), main = "Hist of log(Rings)")
qqnorm(log(dat$Rings))
```

```
qqline(log(dat$Rings)) #at least symatric

dat_trans=dat
dat_trans$Rings=log(dat$Rings)

lm.log=lm(Rings~.,data=dat_trans)
#studentized deleted residuals
stu_res_del = studres(lm.log)
head(sort(abs(stu_res_del), decreasing=TRUE))
# 2052     237     720     629    2184    2306
#8.248828 6.879535 4.014811 3.980955 3.952881 3.802031
qt(1-0.1/(2*n), n-10-1)
# [1] 4.229344
#for alpha = 0.1. In this case Bonferroni's threshold counts 2 observations
#as outliers: 2052,237

#Outlying X Observations
h = as.vector(influence(lm.log)$hat)
index = which(h>2*10/n)
#[1]   5  31  36  39  41  42  53  84 103 156 171 201 211
#In terms of X observations, the leverage method has detected 276 observations as
#outliers (including 2052 and 237).

#Cook's distance
e = lm.log$residuals
mse = anova(lm.log)['Residuals',3]
cook_d = e^2*h/(10*mse*(1-h)^2)
head(sort(abs(cook_d), decreasing=TRUE))
#2052       237      2628      1529      3519       164
#6.77880030 0.04484218 0.02701689 0.01962696 0.01397598 0.01293020
influ = names(subset(cook_d[index], cook_d[index]>4/(n-10)))
#100 cases may potentially have some influence on the ftted values (including 2052 and
237).
table(dat$Rings[rownames(dat)%in%influ])
#all observations of class 1,2,25,27,29 are influential outliers
table(dat$Sex[rownames(dat)%in%influ])
#39 19 42
#delete influential cases
dat_re = dat_trans[!rownames(dat)%in%influ, ]


lm.re=lm(log(Rings)~.,data=dat_re)
summary(lm.re)
summary(lm.log)
par(mfrow=c(1,2))
plot(lm.log,which=5)
plot(lm.re,which = 5)
```

```
bx_plot <- function(V) {
  boxplot(dat_re[[V]] ~ dat_re$Sex, main = V, ylab = V, xlab = "Sex", col = rainbow(4))
}
par(mfrow = c(2, 2))
invisible(sapply(names(dat_re)[2:9], FUN = bx_plot))

par(mfrow=c(1,1))
pairs(rev(dat_re[,-1]))
cor(rev(dat_re[,-1]))

res_plot=function(V) {
  plot(dat_re[[V]],lm.re$residuals,main=paste0("residuals vs ",V),ylab="residuals",xlab=V)
  smooth=smooth.spline(dat_re[[V]],lm.re$residuals,spar=0.9)
  lines(smooth,col="red")
  abline(h=0,lty=3)
}
par(mfrow=c(2,2))
invisible(sapply(names(dat_re)[2:8], FUN = res_plot))
par(mfrow=c(1,1))
plot(lm.log,which=2)

set.seed(232)
train.index <- createDataPartition(1:nrow(dat_re), times = 1, p = 0.8)
saveRDS(train.index$Resample1,"resample.rds")
#resample=readRDS("resample.rds")
#train <- dat_re[resample, ]
#test <- dat_re[-resample, ]
train <- dat_re[train.index$Resample1, ]
test <- dat_re[-train.index$Resample1, ]
par(mfrow=c(1,2))
hist(train$Rings)
hist(test$Rings)
n_tr=nrow(train) #3264

lm.full=lm(data=train,Rings~.)
summary(lm.full)
par(mfrow=c(1,2))
plot(lm.full,which=1)
plot(lm.full,which = 2)

lm.subset=regsubsets(Rings~.,data=train,nbest=1, nvmax=9,method="exhaustive")
subset.summary=summary(lm.subset)
subset.display = as.data.frame(subset.summary$outmat)
press_GCV=n_tr/(n_tr-(2:10))^2*subset.summary$rss
aic=n_tr*log(subset.summary$rss/n_tr)+2*(2:10)
```

```
subset.display =
cbind(subset.display,round(subset.summary$rsq,4),round(subset.summary$adjr2,4),round(
subset.summary$rss/3075,4),round(subset.summary$cp,4),round(subset.summary$bic,4),
                round(aic,4),round(press_GCV,4))
varnames = c("X1","X2","X3","X4","X5","X6","X7","X8","X9")
names(subset.display)=c(varnames,"R^2","adj R^2","sse/n","Cp","BIC","AIC","PE_GCV")
subset.display
#         X1 X2 X3 X4 X5 X6 X7 X8 X9    R^2 adj R^2  sse/n       Cp      BIC       AIC   PE_GCV
#1  ( 1 )           *          0.4510  0.4508 0.0559 1307.2104 -1940.846  -9607.365   0.0527
#2  ( 1 )              *    * 0.4913  0.4910 0.0518  973.7186 -2181.780  -9854.390   0.0488
#3  ( 1 )        *    * *      0.5671  0.5667 0.0440  345.2236 -2700.285 -10378.986   0.0416
#4  ( 1 )        * * * *      0.5842  0.5837 0.0423  204.8724 -2823.835 -10508.626   0.0400
#5  ( 1 ) *      * * * *      0.5964  0.5958 0.0411  105.0972 -2913.227 -10604.109   0.0388
#6  ( 1 ) *      * * * * *    0.6077  0.6070 0.0399   12.9390 -2997.992* -10694.964   0.0378
#7  ( 1 ) *      * * * * *  * 0.6084  0.6076 0.0398    9.3190 -2995.528 -10698.591   0.0377*
#8  ( 1 ) *    * * * * * * * 0.6088  0.6079 0.0398    8.0064 -2990.758 -10699.912   0.0377
#9  ( 1 ) * * * * * * * * * 0.6088  0.6077 0.0398   10.0000 -2982.674 -10697.918   0.0377

train_reSex=train
train_reSex$Sex=as.character(train_reSex$Sex)
train_reSex$Sex[(train$Sex=="M" | train$Sex=="F")]="M"
train_reSex$Sex=as.factor(train_reSex$Sex)
levels(train_reSex$Sex)

lm.bestsub=lm(Rings~.,data=train_reSex[,-c(2,8)])
summary(lm.bestsub)
round(lm.bestsub$coefficients,2)
par(mfrow=c(1,3))
plot(lm.bestsub,which=1)
plot(lm.bestsub,which = 2)
plot(lm.bestsub,which=5)

test_reSex=test
test_reSex$Sex=as.character(test$Sex)
test_reSex$Sex[(test$Sex=="M" | test$Sex=="F")]="M"
test_reSex$Sex=as.factor(test_reSex$Sex)
levels(test_reSex$Sex)

pred_subset=predict(lm.bestsub,newdata = test_reSex)
mean((test$Rings-pred_subset)^2) #0.03532405, in-sample sse/n=0.0399

lambdaVec=seq(0,0.1,by=0.001)
X_tilta=model.matrix(Rings~.,data=train)
Y_tr=train$Rings
ridge.mod=glmnet(X_tilta[,-1],Y_tr,alpha=0,lambda = lambdaVec) #Scale all variables by
default
fitValue=predict(ridge.mod,newx=X_tilta[,-1],s=lambdaVec)
```

```
trainSSE=apply(fitValue,2,function(col) {sum((Y_tr-col)^2)})
plot(lambdaVec,trainSSE)
df_Vec=sapply(lambdaVec,function(lambda)
{sum(diag(X_tilta%*%solve(t(X_tilta)%*%X_tilta+lambda)%*%t(X_tilta)))})
GCV_vec=trainSSE/n_tr/(1-df_Vec/n_tr)^2
plot(lambdaVec,GCV_vec,xlab="lambda",ylab = "GCV",pch=16,main="GCV vs lambda")

svdX=svd(X_tilta)
corSingY=round(apply(svdX$u,2,function(col) {cor(col,Y_tr)}),4)
corSingY
#-0.56719155  0.51559080 -0.31159386  0.05470943  0.45458406  0.07268949  0.04041274
#0.08202615 -0.11309926  0.01532834

set.seed(232)
pcr.fit=pcr(Rings~.,data=train,scale=TRUE,validation="CV")
summary(pcr.fit)
validationplot(pcr.fit,val.type = "MSEP")
mean(pcr.fit$residuals[,1,9]^2) #0.03749019

X_val=model.matrix(Rings~.,data=test)
pcr.pred=predict(pcr.fit,X_val[,-1],ncomp = 9)
mean((test$Rings-pcr.pred)^2) #0.03511708


lm.full2=lm(Rings~.^2,data=train_reSex[,-c(2,8)])
summary(lm.full2)
stepwise=stepAIC(lm.full2,direction="both",k=log(n_tr),test="F")
stepwise$anova

lm.step=lm(Rings ~ Sex + Diameter + Height + Whole.weight + Shucked.weight +
        Viscera.weight + Sex:Shucked.weight +
        Diameter:Height + Diameter:Whole.weight +
Diameter:Shucked.weight,data=train_reSex[,-c(2,8)])
summary(lm.step)
mean(lm.step$residuals^2) #0.03392356
par(mfrow=c(1,2))
plot(lm.step,which = 1)
plot(lm.step,which=2)
plot(lm.step,which=5)
plot(lm.step,which=4)
h = as.vector(influence(lm.step)$hat)
GCV=mean(lm.step$residuals^2)/(1-mean(h))^2
AIC=n_tr*log(mean(lm.step$residuals^2))+2*sum(h)

pred.step=predict(lm.step,newdata=test_reSex[,-c(2,8)])
mean((test$Rings-pred.step)^2) #0.03202449
```