Prediction Model of the Age of Abalone

Group Membership:
Xueting Shao sxshao@ucdavis.edu
Yingjie Li yyjli@ucdavis.edu
Yibing Luo ybluo@ucdavis.edu

Abstract:

In this report, a dataset of abalone's age and other measurements is explored. It is supposed to analysis the potential pattern between the specified response variable (age) and other variables. On this purpose, a relatively optimal model with good predictive ability, which satisfies the assumption of linear regression, is established to measure and explorer the variables. Model diagnosis is applied to give evidence for model building.

Introduction:

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. The data shows 9 variables that may use to predict the age of abalone. They are sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight and rings.
The project is interested in three aspects: (i) The correlation between rings (age) and other explanatory variables; (ii) The potential multicollinearity among explanatory variables; (iii) Relatively optimal model to predict. Therefore, basic exploratory of data, data transformation, ridge regression, model selection procedure, and cross validation are applied.

Methods and Results:

1. Exploratory Data Analysis

    1.1. Type of each variable: quantitative or qualitative

    There are 10 variables and 4177 observations in the Abalone Dataset, including the response variable Rings, which is 1.5 less than the corresponding age in years. The variable Sex has three categories: female, male and infant. Meanwhile, variable Length, Diameter, Height, Whole, Shucked weight, Viscera weight, and Shell weight are all quantitative variables.

    1.2. Distribution of Each Variable:

    1.2.1. Quantitative Variables

    From Figure 1, the distributions of length and diameter are left-skewed, while those of height, whole weight, sucked weight, viscera weight and shell weight are right-skewed. The distribution of response variable rings is a little right-skewed. Therefore, it is reasonable to consider apply logarithm transformation.

1.2.2. Qualitative Variable

From Figure 2, it is easy to see variable sex has three categories: M, F, and I. The proportion of the three categories is similar.

1.3. Relationship among variables

From Table 1, there are high correlations among quantitative explanatory variables. Meanwhile, the correlation variance between those explanatory variables and rings are approximately equaled to 0.6, which is not very high.

From Figure 3, the distribution of quantitative explanatory variables is different with different category of Sex.

2. Data split

To find and validate the relative optimal model for prediction, the original data is randomly partitioned into test data and training data with 1392 and 2785 observations separately.

From Figure 4 and Figure 5, variables in test data and training data have similar distributions.

3. Model Selection

3.1. Preliminary investigation

3.1.1. First-order model with all variables

Fitted Model 1:

$$\begin{aligned} Y = 4.19888 - 0.84239\text{sexI} + 0.08123\text{sexM} + 1.47141\text{length} \\ + 8.23207\text{diameter} + 7.50919\text{height} + 7.82569\text{whole} \\ - 18.32843\text{sucked} - 8.69840\text{viscera} + 10.33840\text{shell} \end{aligned}$$

Model Assumptions: $\varepsilon_i \sim_{i.i.d} N(0, \sigma^2)$

The Adjusted R-squared is 0.5368, not very large which means this model can only describe 53.68% of the whole data set.
The F-statistic equals to 359.4 with p-value smaller than 0.05, so the model passes the F-test for regression relation.
For coefficients, only length and sexM do not pass the T-test. Also, the standard errors of estimated coefficients are not small and this might be caused by the high correlations among variables.

3.1.2. Residual plots

Figure 6 And Figure 7 show the residual plots, which indicate the residuals are heavy-tailed. The model assumption does not hold well. For remedy, box-cox procedure is applied to identify proper transformation.

Additionally, there seem to be non-linear relationship between residuals and fitted values, so second-order models should be considered.

3.1.3. Box-cox procedure

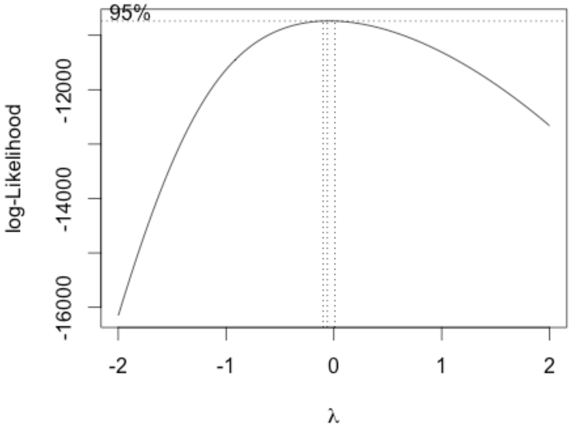


Figure 8 Plot of Box-cox of Model 1

From Figure 8, it suggests a logarithm transformation on rings.

3.2. Transformation

$$Y^* = log(Y)$$

3.3. First-order Models fitted on transformed Y: Y^*

3.3.1. First-order model with all variables

Fitted Model 2:

$$\begin{aligned} Y^* = 1.386637 &- 0.096859\text{sexI} + 0.009086\text{sexM} + 0.601379\text{length} \\ &+ 1.279768\text{diameter} + 0.851934\text{height} + 0.546883\text{whole} \\ &- 1.559631\text{sucked} - 0.727490\text{viscera} + 0.712404\text{shell} \end{aligned}$$

Model Assumptions: $\varepsilon_i \sim_{i.i.d} N(0, \sigma^2)$

The Adjusted R-squared is 0.5979, not very large which means this model can only describe 59.79% of the whole data set, but it is larger than model 1.
This model passes the F-test for regression relation while variable sexM does not pass the T-test.

From Figure 9, it is clear that the residuals are approximately linear.
From Figure 10, there seems to be a quadric relationship between residuals and fitted values.

3.3.2. Best subsets selection

Given large number of potential variables in the model, stepwise procedure is applied.
Fitted Model 3:

$$
\begin{aligned}
Y^* = 1.386637 - 0.096859\text{sexI} + 0.009086\text{sexM} + 0.601379\text{length} \\
+ 1.279768\text{diameter} + 0.851934\text{height} + 0.546883\text{whole} \\
- 1.559631\text{sucked} - 0.727490\text{viscera} + 0.712404\text{shell}
\end{aligned}
$$

Model Assumptions: $\varepsilon_i \sim_{i.i.d} N(0, \sigma^2)$

Model 3 is the same as Model 2.

3.4. Second-order models

3.4.1. Second-order model with all variables

Fitted Model 4:

Table 2 Coefficients of Model 4

Variable	Estimate	Variable	Estimate
(Intercept)	0.28173	sexM:shell	0.01277
sexI	0.06176	length:diameter	-10.94476
sexM	0.06142	length:height	10.34798
length	2.89706	length:whole	-4.68226
diameter	6.02761	length:sucked	6.17182
height	3.22897	length:viscera	-0.19648
whole	2.21967	length:shell	7.51997
sucked	-6.22558	diameter:height	-19.51632
viscera	-1.27580	diameter:whole	0.89219
shell	-0.62012	diameter:sucked	3.01683
sexI:length	-0.73399	diameter:viscera	6.94875
sexM:length	0.19077	diameter:shell	-0.47355
sexI:diameter	-0.52523	height:whole	7.68118
sexM:diameter	-0.40149	height:sucked	-6.35741
sexI:height	1.12584	height:viscera	-13.33099
sexM:height	-0.49264	height:shell	-5.91291
sexI:whole	-0.35319	whole:sucked	0.55346
sexM:whole	0.20523	whole:viscera	-1.08830
sexI:sucked	1.59160	whole:shell	-0.96853
sexM:sucked	-0.03072	sucked:viscera	1.38317
sexI:viscera	-0.09139	sucked:shell	-2.01544
sexM:viscera	-0.49408	viscera:shell	1.35131
sexI:shell	0.44627		

Model Assumptions: $\varepsilon_i \sim_{i.i.d} N(0, \sigma^2)$

The Adjusted R-squared is 0.6555, which means this model can describe 65.55% of the whole data set.

This model passes the F-test for regression relation but several factors cannot pass the T-test. And there are too many variables (45) in the model, which may lead to overfitting.

3.4.2. Best subsets selection

Applied the stepwise procedure, the selected model is :

Fitted Model 5:

$$
\begin{aligned}
Y = 0.60663 &+ 4.33238 diameter - 4.95865\, sucked\; + 2.10391\, shell \\
&+ 1.23509 whole\; + 0.06692 sexI + 0.06023\mathrm{sexM} \\
&- 0.66775 viscera\; + 10.76521 height \\
&+ 0.38326 sucked{:}\, whole\; + 1.02538 sucked{:}\, sexI \\
&+ 0.15501 sucked{:}\, sexM + 5.85535 whole{:}\, height \\
&- 0.93058 shell{:}\, whole - 29.55180 diameter{:}\, height \\
&+ 7.95805 diameter{:}\, sucked - 1.10198\, diameter{:}\, sexI \\
&- 0.28275 diameter{:}\, sexM - 2.82440 diameter{:}\, whole \\
&- 6.42414 sucked{:}\, height
\end{aligned}
$$

The Adjusted R-squared is 0.6515, which means this model can describe 65.15% of the whole data set.
This model passes the F-test for regression relation and only sex factor can not pass the T-test. It would not affect the significance of the whole model.

From Figure 11 and 12, the relationship between residuals and fitted values is approximately independent. The distribution of residuals is nearly normal. This indicates model assumption holds.

3.5. Remedial Measure for Multicollineartiy

The departures from model assumptions are fixed through Model 5, however multicollinearity should also be considered.
Based on Model 5, variance inflation factors (VIF) of each coefficient are shown in Table 3.

VIF's are much bigger than 10, which indicate high inter-correlation among X variables. Thus, the LS estimators are unduly influenced and will affect the predictions based on the model. Ridge regression is applied to fix this through introducing biased estimators.

Figure 13 shows the ridge trace plot based on Model 5.
$\lambda = 0.019$ is suggested by the generalized cross-validation (GCV) criterion.

Model 6:

Table 4 Coefficients of Model 6

Variable	**Estimate**	**Variable**	**Estimate**
(Intercept)	0.62019487	sucked:sexI	1.01171714
diameter	4.26791205	sucked:sexM	0.15245902
sucked	-4.79843379	whole:height	5.46074020
shell	2.12148048	shell:whole	-0.94217526
whole	1.17677829	diameter:height	-29.05478957
sexI	0.05964185	diameter:sucked	7.36272574
sexM	0.05761861	diameter:sexI	-1.07417139
viscera	-0.66774376	diameter:sexM	-0.27449216
height	10.67764580	diameter:whole	-2.58281165

sucked:whole	0.40327701	sucked:height	-5.80564748

4. Model diagnostic and validation

4.1. Criterions of Models

Table 5 Criterions of Models

	P	MSE	C_P	R^2	R_a^2	Press
Model 1	10	4.650231	10.000	0.5383	0.5368	13032.20
Model 2	10	0.040804	466.947	0.5992	0.5979	112.96
Model3/4	10	0.040804	466.947	0.5992	0.5979	112.96
Model 4	45	0.034621	45.000	0.6609	0.6554	101.02
Model 5	20	0.035017	51.640	0.6539	0.6515	99.65
Model 6	20	0.034768				

4.2. Model validation

4.2.1. Internal Validation

According to Table 5,

Model 1 has the largest MSE since it only includes the linear effects. And $C_P = p$ indicates it has no bias.

Model 2 and Model 3 are the same and they all have model bias.

Model 4 has the largest number of parameters and it has no bias.

Model 5 has smallest Press value, which indicates good prediction. C_P is not far away from p so it can be considered with little bias. Additionally, only the model assumptions of Model 5 holds.

Based on Model 5, Model 6 is constructed by ridge regression to fix the muliticollinearity. Because it is a biased model so C_P is not considered.

4.2.2. Cross-validation

Applied Fitted Model 5 and Fitted Model 6 on the test data, and compare the estimated coefficients.

From Table 6, estimated coefficients are similar, thus model 5 has consistency on parameters.

Table 7 Predictions of Model 5 and Model 6

	Data Set	**SSE**	**MSE**	**R2_2**	**Press**	**Press/n**	**MSPR**
Model 5	Training	96.789	0.035	0.652	99.652	0.036	--
	Validation	50.096	0.036	0.654	--	--	0.03866423
Model 6	Training	96.793	0.035				
	Validation	53.981	0.039				0.03875174

From Table 7, the predictions on test data show small differences between Model 5 and Model 6. Though MSPR of Model 5 is unexpectedly a little bigger than that of Model 6, Model 6 has better predictive ability when extrapolation occurs. However, when extrapolation does not occur, Model 5 is better since it is not biased and has smaller MSPR.

4.3. Outlying and Influential

4.3.1. Outlying Y

Based on Model 5, studentized residuals against fitted values are drawn in Figure 13. According to Bonferroni's Procedure at level 0.05, case 237, 481, 2052, 2184, 2628 are identified as outlying Y observations (Figure 14).

4.3.2. Outlying X

According to leverage, $h_{ii} > \frac{2p}{n}$, there are 329 cases are identified as outlying with regard to its X value.

4.3.3. Influential

There are three measure values are calculated to identify whether some outlying cases may leads to major changes of the fitted regression function with their exclusion. Results are shown in Table 8.

Table 8 Test of outlying cases

Ith case	**DFFITS**	**cov.r**	**Cook's Distance**	**hat**
237	-1.2708668	0.9208513	0.08020785	0.052135275
481	0.4947060	0.9324790	0.01218592	0.013179800
2052	-8.1253051	3.9289875	3.28602344	0.767648615
2184	-0.4746297	0.8809143	0.01118816	0.007693966
2628	3.4615821	1.2288649	0.59506603	0.289746356

As DFFITS shows, all of those five cases are influential with regard their fitted values.
DFBETAS considering each estimated regression coefficient is calculated And Table 9.1 and Table 9.2shows whether each case has influence on each estimated coefficient.

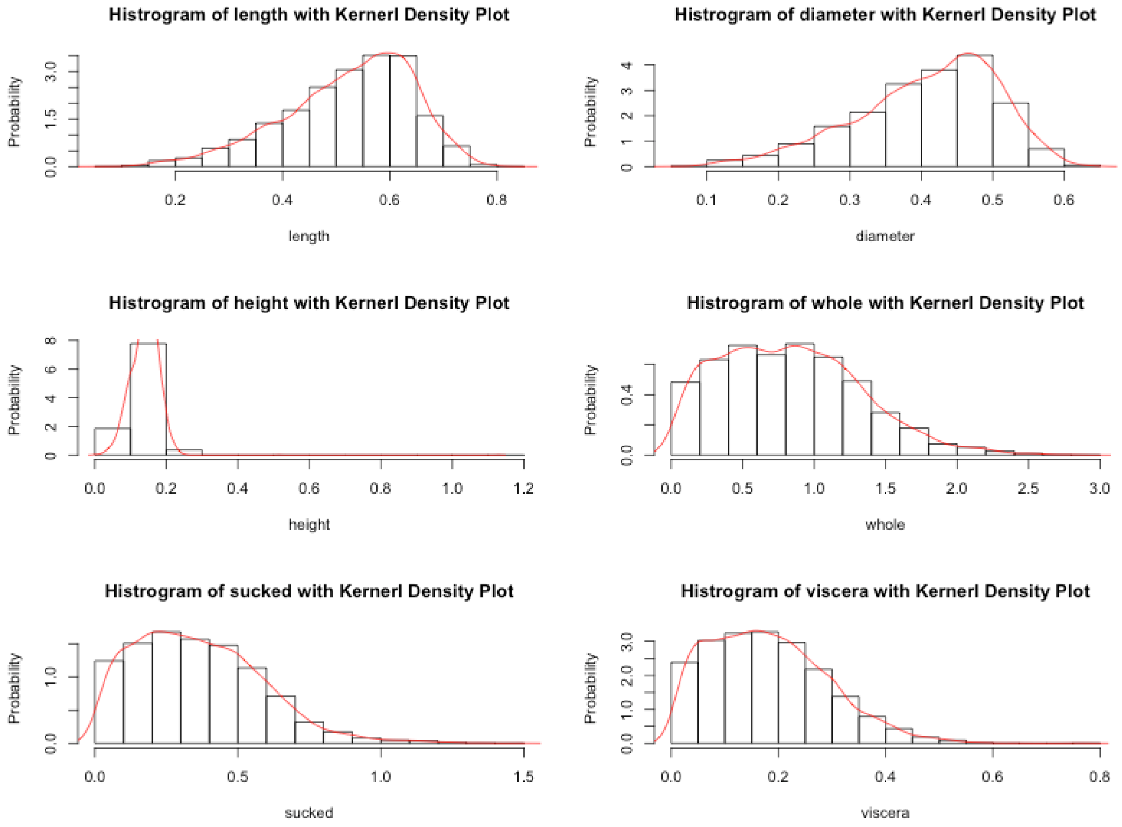
Conclusions and Discussion:

- Original data set is spited into Test and Training data since the data is large enough and it is good to do cross validation for searching a relatively good prediction model. The results of cross validation suggest a good prediction of Model 5 and Model 6. Additionally, the number of influential cases is quite small comparing with data size. It can be considered the results of Model 5 and 6 will not be influenced.
- Multicollinearity is quite serious in this data. Ridge regression reduces the variances of estimators by introducing bias. While comparing the ridge regression and Model 5 in cross validation, the results are similar because no extrapolation occurs.
- If extrapolation does not occur, Model 5 is a better choice than Model 6. If extrapolation occurs, Model 6 is better than Model 5 although it is biased estimation

References:

1. Kutner M H. Applied linear statistical models[M]. Chicago: Irwin, 1996.
2. Sam Waugh (1995) "Extending and benchmarking Cascade-Correlation", PhD thesis, Computer Science Department, University of Tasmania.
3. Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (_Haliotis_ species) in Tasmani a. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait ", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)
4. David Clark, Zoltan Schreter, Anthony Adams "A Quantitative Comparison of Dystal and Backpropagation", submitted to the Australian Conference on Neural Networks (ACNN'96).

Appendix 1



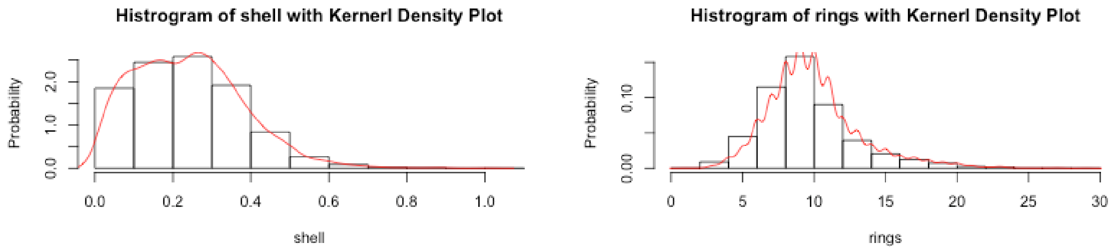


Figure 1 Histrogram Plot

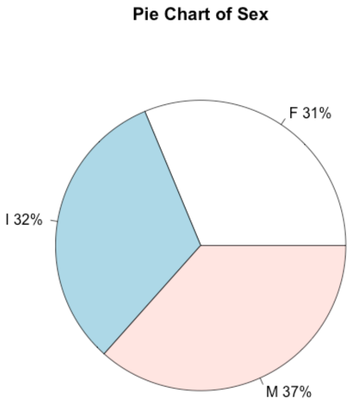
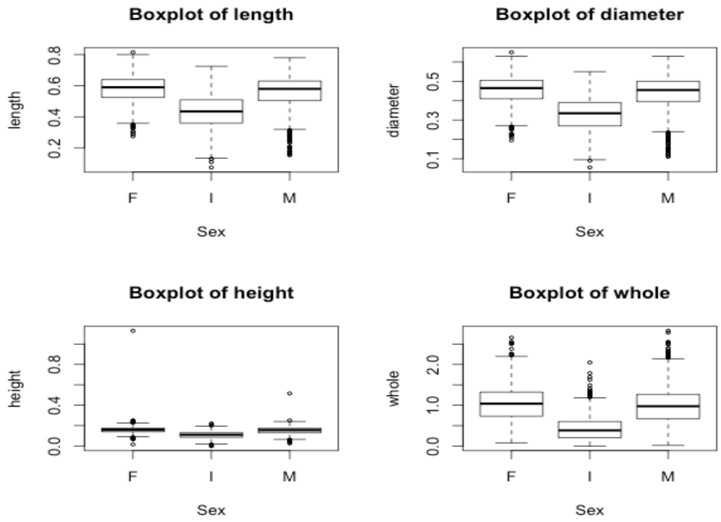


Figure 2 Pie Chart of Sex



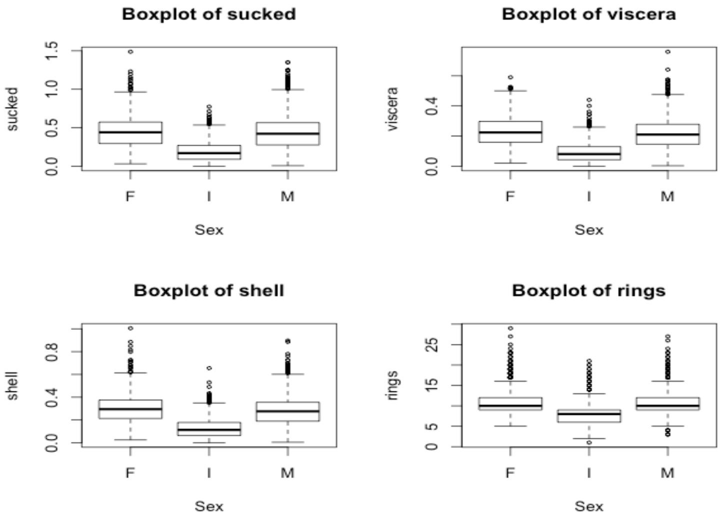
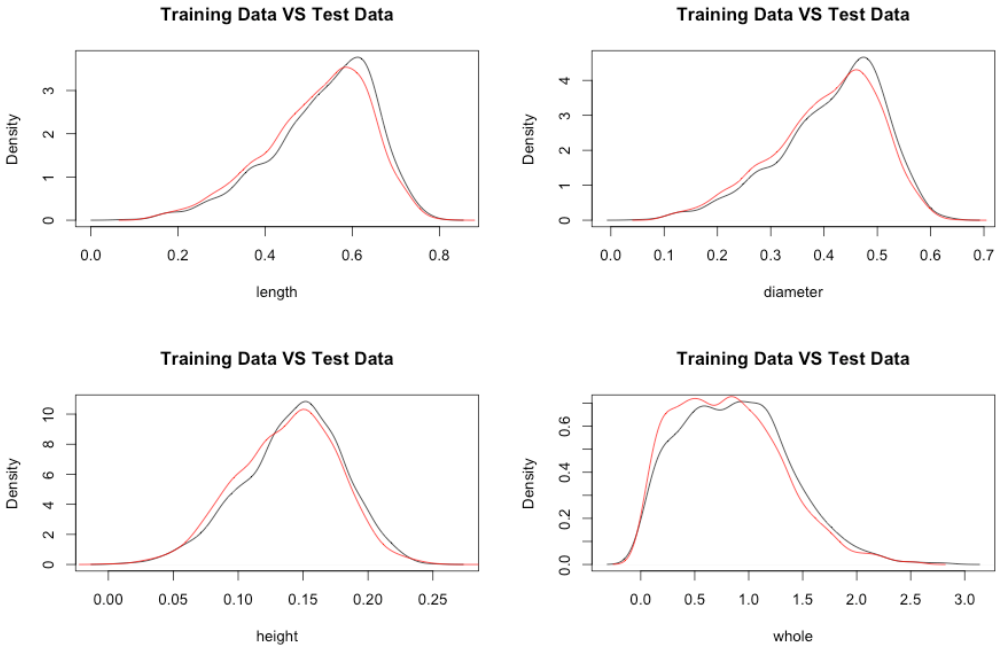


Figure 3 Boxplot of Variables



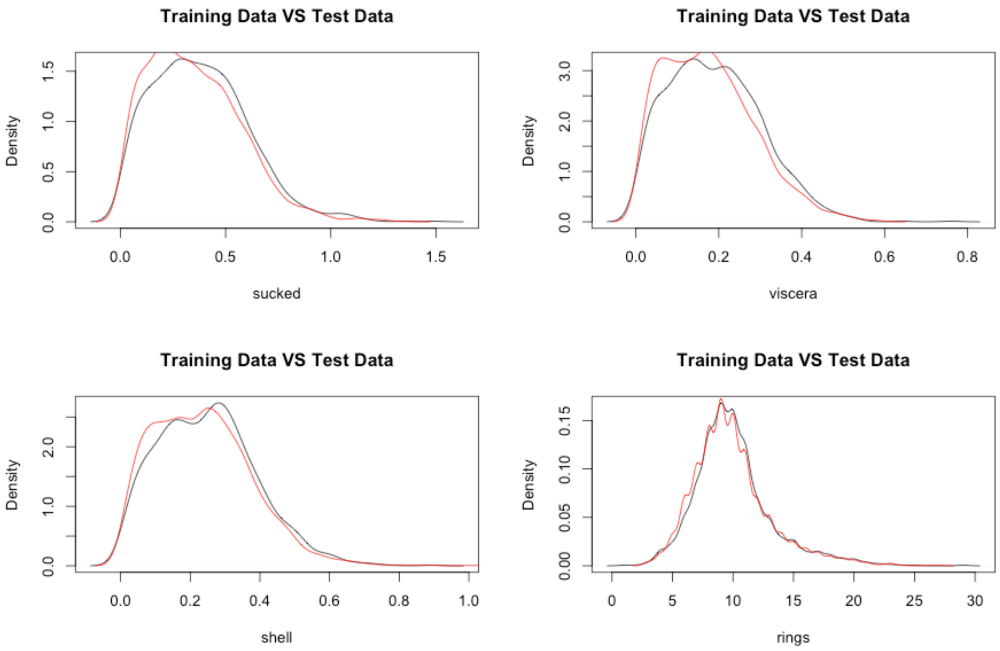


Figure 4 Line Chart of Training Data VS Test Data

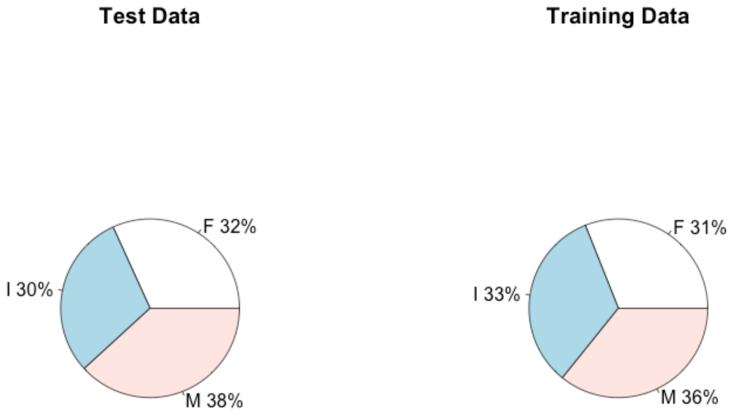


Figure 5 Pie Chart of Training Data VS Test Data

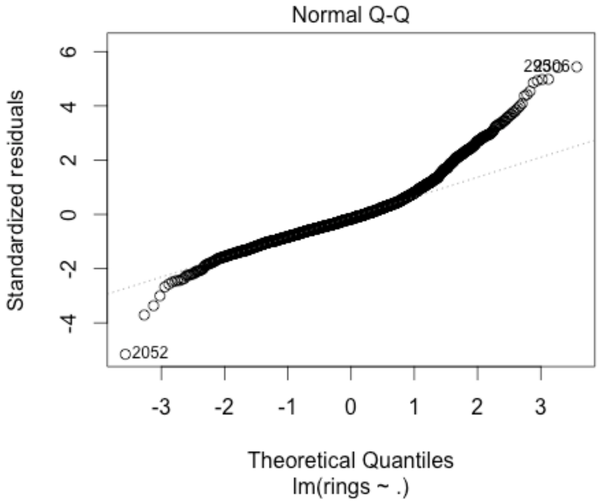


Figure 6 Plot of Normal Q-Q of Model 1

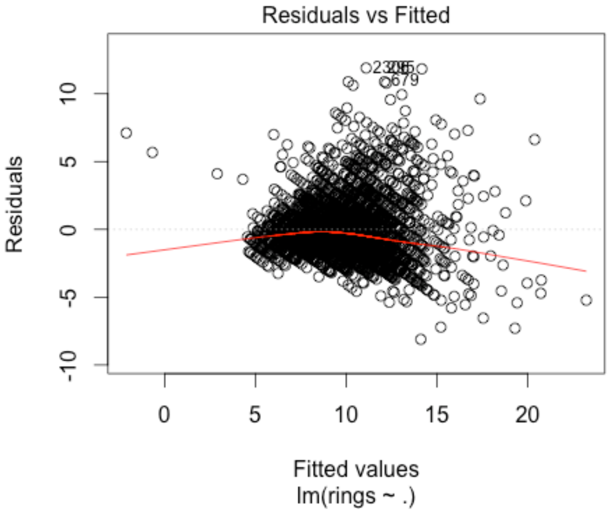


Figure 7 Plot of Residuals vs. Fitted of Model 1

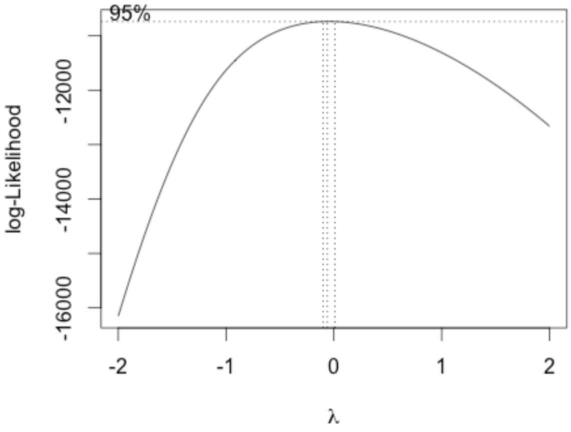


Figure 8 Plot of Box-cox of Model 1

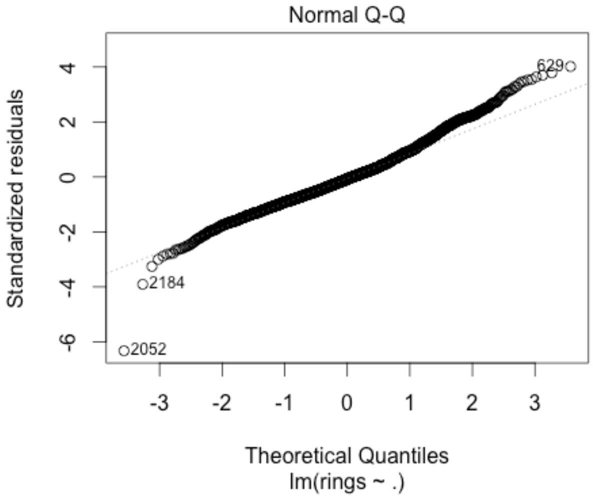


Figure 9 Plot of Normal Q-Q of Model 2

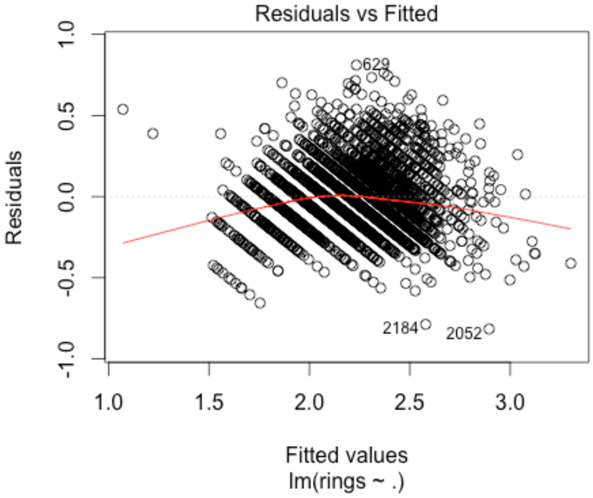


Figure 10 Plot of Residuals vs. Fitted of Model 2

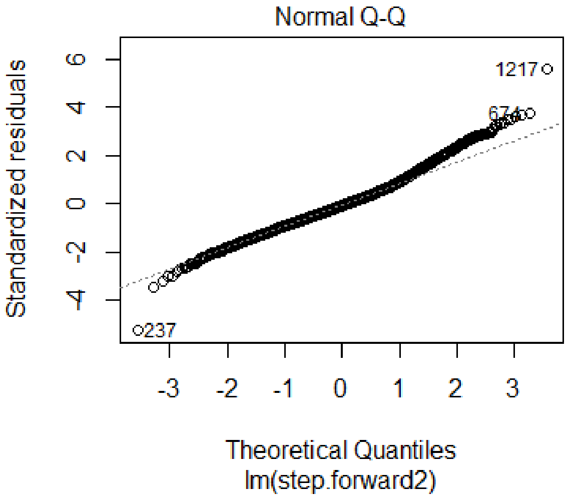


Figure 11 Plot of Normal Q-Q of Model 5

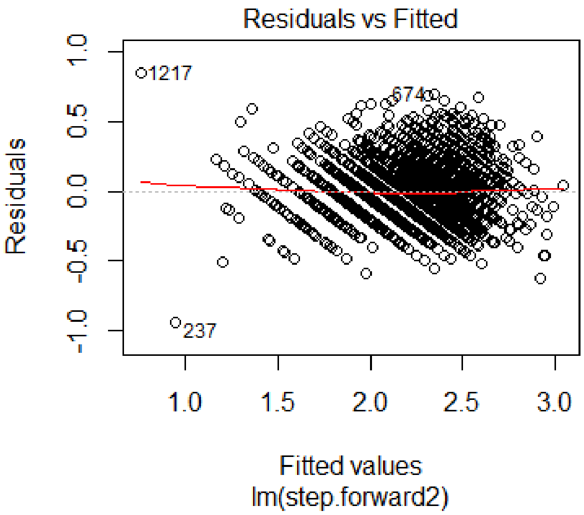


Figure 12 Plot of Residuals vs. Fitted of Model 5

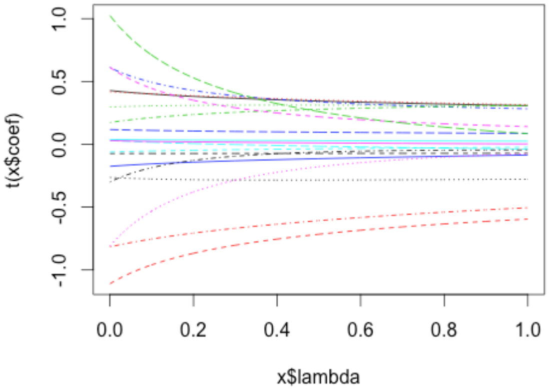


Figure 13 The ridge trace plot of Model 5

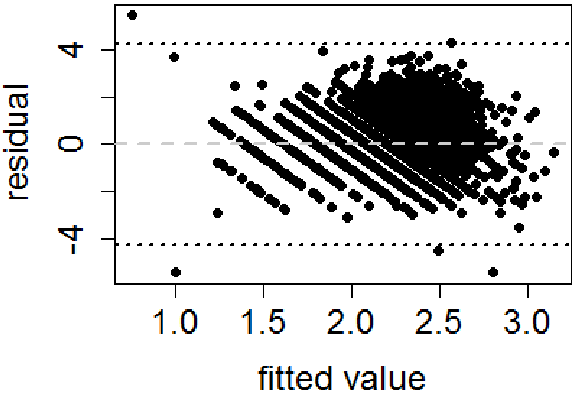


Figure 14 Studentized Residuals vs. Fitted of Model 5

Table 1 Correlation matrix among quantitative variables

	length	diameter	height	whole	sucked	viscera	shell	rings
length	1.0000000	0.9868116	0.8275536	0.9252612	0.8979137	0.9030177	0.8977056	0.5567196
diameter	0.9868116	1.0000000	0.8336837	0.9254521	0.8931625	0.8997244	0.9053298	0.5746599
height	0.8275536	0.8336837	1.0000000	0.8192208	0.7749723	0.7983193	0.8173380	0.5574673
whole	0.9252612	0.9254521	0.8192208	1.0000000	0.9694055	0.9663751	0.9553554	0.5403897
sucked	0.8979137	0.8931625	0.7749723	0.9694055	1.0000000	0.9319613	0.8826171	0.4208837
viscera	0.9030177	0.8997244	0.7983193	0.9663751	0.9319613	1.0000000	0.9076563	0.5038192
shell	0.8977056	0.9053298	0.8173380	0.9553554	0.8826171	0.9076563	1.0000000	0.6275740
rings	0.5567196	0.5746599	0.5574673	0.5403897	0.4208837	0.5038192	0.6275740	1.0000000

Table 3 VIF of each coefficient of Model 5

Variables	VIF
diameter	122.64659
sucked	1216.82980
shell	98.77648
whole	1796.37725
sex	73.41045
viscera	17.36716
height	172.83262
sucked:whole	248.32281
sucked:sex	56.10378
whole:height	2523.92985
shell:whole	171.86901
diameter:height	950.28949
diameter:sucked	3868.61398

diameter:sex	188.50517
diameter:whole	3916.00165
sucked:height	1604.51235

Table 6 Estimated coefficients of Model 5 and Model 6

	Model 5				Model 6
Variables	Coef. of Test Data	Standard error of Test Data	Coef. of Training Data	Standard error of Training Data	Coef. of Training Data
(Intercept)	0.96105	0.17723	0.60663	0.12251	0.62019
diameter	3.43015	0.71270	4.33238	0.44462	4.26791
sucked	-1.12632	0.73401	-4.95865	0.57503	-4.79843
shell	3.13452	0.45845	2.10391	0.25104	2.12148
whole	-0.63330	0.45814	1.23509	0.30859	1.17678
sexI	-0.18862	0.12840	0.06692	0.09168	0.05964
sexM	-0.09556	0.11471	0.06023	0.07848	0.05762
viscera	-0.59108	0.20412	-0.66775	0.13447	-0.66774
height	8.16823	1.73283	10.76521	1.23983	10.67765
sucked:whole	0.98643	0.20534	0.38326	0.12444	0.40328
sucked:sexI	0.55829	0.23424	1.02538	0.15361	1.01172
sucked:sexM	0.02871	0.13113	0.15501	0.08649	0.15246
whole:height	12.67629	2.05666	5.85535	1.71001	5.46074
shell:whole	-1.69657	0.33760	-0.93058	0.16443	-0.94218
diameter:height	-22.95112	6.65139	-29.55180	4.07992	-29.05479
diameter:sucked	3.94624	2.11245	7.95805	1.74066	7.36273
diameter:sexI	-0.10708	0.45956	-1.10198	0.32081	-1.07417
diameter:sexM	0.19229	0.35825	-0.28275	0.24289	-0.27449
diameter:whole	-1.45469	0.93477	-2.82440	0.78667	-2.58281
sucked:height	-22.09863	3.87939	-6.42414	3.04006	-5.80565

Table 9.1 Test of Influential

Ith case	dfb.1_	dfb.dmtr	dfb.sckd	dfb.shll	dfb.whol	dfb.sexI	dfb.sexM	dfb.vscr	dfb.hght	dfb.sckd:w
237	F	T	F	T	F	T	T	F	T	F
481	F	T	T	F	F	F	T	F	T	F
2052	T	T	F	T	T	F	T	T	F	T
2184	T	F	T	F	F	F	F	T	F	F
2628	F	T	T	T	F	T	F	T	T	T

Table 9.2 Test of Influential continued

Ith case	dfb.sc:I	dfb.sc:M	dfb.whl:	dfb.shl:	dfb.dmtr:h	dfb.dmtr:s	dfb.dm:I	dfb.dm:M	dfb.dmtr:w	dfb.sckd:h
237	F	T	T	F	F	T	F	F	T	F
481	F	T	F	F	F	F	F	F	T	T
2052	F	T	T	F	T	T	F	F	F	F
2184	F	T	F	T	T	F	T	F	T	T

2628	T	F	T	F	F	F	F	T	T	F

Appendix 2

```
setwd('/Users/Shawn/Dropbox/lecture/stat206/project')
abalone = read.table("abalone.txt" , sep =',')
colnames(abalone) = c("sex","length","diameter","height","whole","sucked",
               "viscera","shell","rings")

a = abalone
a$sex = factor(a$sex)
#
#split data
#
set.seed(10)
n = nrow(a)
index.s = sample(1:n , size = n*2/3 , replace = FALSE)

a.train = a[index.s,]
a.test = a[-index.s,]

sapply(2:9, function(i) boxplot(a.train[,i],a.test[,i]))
##################################################################

boxplot(rings~sex,data = a.train,
     main='side-by-side boxplots',xlab='factor levels',
     ylab='observation',col=rainbow(6))


################################################################################
################################################################################
#transform response variables
a.train$rings = log(a.train$rings)
a.test$rings = log(a.test$rings)
#
#model selection
#
#
#selection of first-order effects
#
#
model.1st = lm(rings~. , data = a.train)
mse = summary(model.1st)$sigma^2
summary(model.1st)
plot(model.1st, which=1)
plot(model.1st, which=2)
plot(model.1st$residuals)
#
#best subsets selection
#
library(leaps)
sub.set = regsubsets(rings~. , data = a.train , nbest = 1,
                     nvmax = 16 , method = "exhaustive")
sum.sub = summary(sub.set)

#number of parameters in each model
num.p = as.numeric(rownames(sum.sub$which)) + 1L

#parameters in model
n.train = nrow(a.train)
sse = sum.sub$rss

#aic ,  pic
aic = n.train*log(sse/n) + 2*num.p
bic = n.train*log(sse/n) + log(n)*num.p

sub.table = cbind(sum.sub$which, sse, sum.sub$rsq, sum.sub$adjr2,
                  sum.sub$cp, aic ,bic)

#null model
fit0 = lm(rings~1, data = a.train)
sse0 = sum(fit0$residuals^2)
p0 = 1
c0 = sse0/mse - (n.train-2*p0)
aic0 = n.train*log(sse0/n.train) + 2*p0
bic0 = n.train*log(sse0/n.train) +log(n.train)*p0
none = c(p0, rep(0,9), sse0, 0, 0, c0, aic0, bic0)

sub.table = rbind(none, sub.table)
colnames(sub.table) = c(colnames(sum.sub$which), "sse", "R^2", "R^2_a", "cp",
                        "aic", "bic")

#
#forward stepwise procedure
#
library(MASS)

step.forward = stepAIC(fit0, scope = list(upper = model.1st, lower = ~1),
                       direction = "both", k=2)
#
#
#selection of first-order and second-order effects
#
#
```

```
model.2nd = lm(rings~.^2, data = a.train)
mse2 = summary(model.2nd)$sigma^2
#
#forward stepwise procedure
#
step.forward2 = stepAIC(fit0, scope = list(upper = model.2nd, lower = ~1),
                        direction = "both", k=2)
################################################
################################################
#
#model validation
#
#
#
#internal validation
#
model1 = lm(step.forward , data = a.train)
plot(model1, which = 1)
plot(model1, which = 2)

model2 = lm(step.forward2 , data = a.train)
plot(model, which = 1)
plot(model, which = 2)

sse.1st = anova(step.forward)["Residual" , 2]
p.1st = length(step.forward$coefficients)
cp.1st = sse.1st/mse2 - (n.train-2*p.2nd)
press.1st = sum(step.forward$residuals^2/(1-influence(step.forward)$hat)^2)
mse.1st = anova(step.forward)["Residuals",3]
#cp̮??

sse.2nd = anova(step.forward2)["Residual" , 2]
p.2nd = length(step.forward2$coefficients)
cp.2nd = sse.2nd/mse2 - (n.train-2*p.2nd)
press.2nd = sum(step.forward2$residuals^2/(1-influence(step.forward2)$hat)^2)
mse.2nd = anova(step.forward2)["Residuals",3]
#(cp??51????p 24????Щ?? ????????һ??’▯?????▯?ʰ????????Щ??Ç??�� ??????model bias)
#press.2nd/n = 0.00733 , mse.2nd = 0.00706. Little difference between these two variables
#supports the validity of the model. And the mse is small which shows a good ablity of
#the model

#
#external validation
#
#caculation
model2.v = lm(step.forward2 , data = a.test)

mspr2 =round (mean((predict.lm(model2, a.test)-a.test$rings)^2),3)
```

```
press.2nd/n.train

sse_model2 = round(anova(model2)["Residuals",2],3)

sse_model2.v = round(anova(model2.v)["Residuals",2],3)

mse_model2 = round(anova(model2)["Residuals",3],3)

mse_model2.v = round(anova(model2.v)["Residuals",3],3)

model2_R2_a = round(summary(model2)$adj.r.squared,3)

model2_R2_a.v = round(summary(model2.v)$adj.r.squared,3)
#model2
mod_sum_2 = cbind(coef(summary(model2.v))[,1], coef(summary(model2.v))[,2],
            coef(summary(model2))[,1],coef(summary(model2))[,2])
colnames(mod_sum_2) = c('coef validation','coef std.err validation',
                'coef ','coef std.err')

Training_2 = cbind(sse_model2,mse_model2,model2_R2_a,round(press.2nd,3),
             round(press.2nd/n.train,3),"--")
Validation_2 = cbind(sse_model2.v,mse_model2.v,model2_R2_a.v,"--","--",
              mspr2)
con_2 = rbind(Training_2,Validation_2)
rownames(con_2) = c('Training','Validation')
colnames(con_2) = c('sse','mse','R2_2','press','press/n','mspr')

mod_sum_2
con_2

#############################################################################
###
#
#outlying
#
#outlying y
model.final = lm(step.forward2, data = a)
hii = influence(model.final)$hat
mse = anova(model.final)["Residuals",3]
res = model.final$residuals
stu.res = res/sqrt(mse*(1-hii))   #studentized residuals

res.del = res / (1-hii)   # deleted residuals
library(MASS)
stu.res.del = studres(model.final)  #studentized deleted residuals
bon.thre = qt(1-0.1/(2*n),n-model.final$rank-1)
```

```
#residuals vs. fitted values plots
plot(model.final$fitted, stu.res.del , xlab="fitted value", ylab="residual",
    cex.lab=1.5, cex.axis = 1.5, pch = 19, cex = 1.5)
abline(h=0, col = grey(0.8), lwd = 2, lty = 2)
abline(h = bon.thre, lwd = 2, lty = 3)
abline(h = -bon.thre, lwd = 2, lty = 3)

#test for outlying Y
sse = sum((summary(model.final)$residuals)^2)
ti = res*sqrt((nrow(a)-fit$rank-1)/(sse*(1-hii)-res^2))
tt = qt(1-0.1/(2*nrow(a)) , nrow(a)-fit$rank-1 )
any(abs(ti)>tt)
index_outy = which(abs(ti)>tt)


#test for outlying X
any(hii>2*model.final$rank/nrow(a))
index_outx = which(hii>2*model.final$rank/nrow(a))

#cook's distance (outlying influence)
Di = stu.res^2*hii/(model.final$rank*(1-hii))
plot(Di,type="h",ylab = "Cook's distance")

Di = c(Di)
dd = pf(Di , model.final$rank, nrow(a)-model.final$rank)
any(dd>0.5)

#DFFITS DFBETAS
sta = influence.measures(model.final)

#DFFITS
2*sqrt(model.final$rank/n)

#DFBETAS
2/sqrt(n)


setwd("E:/206project")
data=read.table('E:/206project/abalone.txt', sep=",")
colnames(data)=c('sex','length','diameter','height','whole',
              'shucked','viscera','shell','rings')

boxplot(data$rings~data$sex)
pairs(data)
data1=data[,-1]
corr =cor(data1)
#x variables are highly correlated
data1=transform(data1, testr= shucked+viscera+shell-whole)
```

```
data1$testr
summary(data1$testr)
plot(data1$tesetr)

fitwhole=lm(rings ~ factor(sex)+length+diameter+height+whole+shucked
          +viscera+shell, data=data)
summary(fitwhole)
#mse 2.194, rsquare=0.5379. f_pvalue<2.2e-16
plot(fitwhole, which = 1)
plot(fitwhole, which = 2)
plot(fitwhole$residuals)
library(MASS)
boxcox(fitwhole)
#r=0 so log

fitlog=lm(log(rings) ~ factor(sex)+length+diameter+height+whole+shucked
          +viscera+shell, data=data)
summary(fitlog)
#Residual standard error: 0.2025 on 4167 degrees of freedom
#Multiple R-squared:  0.5991,  Adjusted R-squared:  0.5982
#F-statistic: 691.8 on 9 and 4167 DF,  p-value: < 2.2e-16
plot(fitlog, which = 1)
plot(fitlog, which = 2)
plot(fitlog$residuals)

vy=var(data$rings)
fitlogscale=lm(log(rings/vy) ~ factor(sex)+length+diameter+height+whole+shucked
         +viscera+shell, data=data)
summary(fitlogscale)
#Residual standard error: 0.2025 on 4167 degrees of freedom
#Multiple R-squared:  0.5991,  Adjusted R-squared:  0.5982
#F-statistic: 691.8 on 9 and 4167 DF,  p-value: < 2.2e-16
plot(fitlogscale, which = 1)
plot(fitlogscale, which = 2)
plot(fitlogscale$residuals)

rings ~ diameter + sucked + shell + whole + sex + viscera + height +
  sucked:whole + sucked:sex + whole:height + shell:whole +
  diameter:height + diameter:sucked + diameter:sex + diameter:whole +
  sucked:height

x=cbind(a.train$diameter,a.train$sucked,a.train$shell,a.train$whole,as.numeric(a.train$sex),

a.train$viscera,a.train$height,a.train$sucked*a.train$whole,a.train$sucked*as.numeric(a.train$
sex),
         a.train$whole*a.train$height,a.train$whole*a.train$shell,
         a.train$diameter*a.train$height,a.train$diameter*a.train$sucked,
         a.train$diameter*as.numeric(a.train$sex),a.train$diameter*a.train$whole,
      a.train$sucked*a.train$height)

rxx=cor(x)
VIF=as.data.frame(diag(solve(rxx)))
rownames(VIF)=c('diameter','sucked','shell', 'whole','sex','viscera','height',
          'sucked:whole','sucked:sex', 'whole:height','shell:whole',
          'diameter:height','diameter:sucked','diameter:sex','diameter:whole',
          'sucked:height')

library(MASS)

# Using R's automatic selection methods to select the biasing constant:
# R calls this constant "lambda"

select(lm.ridge(log(rings) ~.^2, data=a.train, lambda = seq(0,1,0.001)))
#modified HKB estimator is 0.128917
#modified L-W estimator is 21.90175
#smallest value of GCV  at 0.144

# The generalized cross-validation (GCV) criterion says
# the optimal biasing constant is .144

ridge.reg <- lm.ridge(log(rings) ~.^2, data=a.train, lambda = 0.144)


# Printing the ridge-regression coefficient estimates for this problem:

ridge.reg

rings ~ diameter + sucked + shell + whole + sex + viscera + height +
  sucked:whole + sucked:sex + whole:height + shell:whole +
  diameter:height + diameter:sucked + diameter:sex + diameter:whole +
  sucked:height

select(lm.ridge(log(rings) ~ diameter + sucked + shell + whole + sex + viscera + height +
          sucked:whole + sucked:sex + whole:height + shell:whole +
          diameter:height + diameter:sucked + diameter:sex + diameter:whole +
          sucked:height, data=a.train, lambda = seq(0,1,0.001)))
#modified HKB estimator is 0.1179347
#modified L-W estimator is 9.06262
#smallest value of GCV  at 0.019
plot(lm.ridge(log(rings) ~ diameter + sucked + shell + whole + sex + viscera + height +
          sucked:whole + sucked:sex + whole:height + shell:whole +
          diameter:height + diameter:sucked + diameter:sex + diameter:whole +
          sucked:height, data=a.train, lambda = seq(0,1,0.001)))

# The generalized cross-validation (GCV) criterion says
# the optimal biasing constant is .019
```

```
ridge.reg5 <- lm.ridge(log(rings) ~ diameter + sucked + shell + whole + sex + viscera + height +
              sucked:whole + sucked:sex + whole:height + shell:whole +
              diameter:height + diameter:sucked + diameter:sex + diameter:whole +
              sucked:height, data=a.train, lambda = 0.019)


# Printing the ridge-regression coefficient estimates for this problem:
SexI=as.numeric(a.train$sex=='I')
SexM=as.numeric(a.train$sex=='M')
trainX=cbind(rep(1,2784),a.train$diameter,a.train$sucked, a.train$shell, a.train$whole, SexI,
      SexM, a.train$viscera, a.train$height, a.train$sucked*a.train$whole,
      a.train$sucked*SexI,a.train$sucked*SexM, a.train$whole*a.train$height,
      a.train$shell*a.train$whole, a.train$diameter*a.train$height,
a.train$diameter*a.train$sucked,
      a.train$diameter*SexI, a.train$diameter*SexM, a.train$diameter*a.train$whole,
      a.train$sucked*a.train$height)
fitted.ridge.reg5=sapply(1:20, function(i) coef[i]*trainX[,i])
fitted.model6=apply(fitted.ridge.reg5, 1, sum)
#mse of model6
mse6=mean((fitted.model6-log(a.train[,9]))^2)
# 0.03476759
sse6=sum((fitted.model6-log(a.train[,9]))^2)
#96.79298
summary(ridge.reg5)
ridge.reg5
newSexI=as.numeric(a.test$sex=='I')
newSexM=as.numeric(a.test$sex=='M')
newX=cbind(rep(1,1393),a.test$diameter,a.test$sucked, a.test$shell, a.test$whole, newSexI,
      newSexM, a.test$viscera, a.test$height, a.test$sucked*a.test$whole,
      a.test$sucked*newSexI,a.test$sucked*newSexM, a.test$whole*a.test$height,
      a.test$shell*a.test$whole, a.test$diameter*a.test$height, a.test$diameter*a.test$sucked,
      a.test$diameter*newSexI, a.test$diameter*newSexM, a.test$diameter*a.test$whole,
      a.test$sucked*a.test$height)
coef=coef(ridge.reg5)
ridgecoef=as.data.frame(coef)
predict.ridge.reg5=sapply(1:20, function(i) coef[i]*newX[,i])
predict.model6=apply(predict.ridge.reg5, 1, sum)
predict.y6=exp(predict.model6)
predict.y5=exp(predict.model5)
model5=lm(log(rings) ~ diameter + sucked + shell + whole + sex + viscera + height +
       sucked:whole + sucked:sex + whole:height + shell:whole +
       diameter:height + diameter:sucked + diameter:sex + diameter:whole +
       sucked:height, data=a.train)
summary(model5)
predict.model5=predict(model5, a.test[,-9])
mspr5=mean((predict.y5-a.test[,9])^2)
mspr6=mean((predict.y6-a.test[,9])^2)
```

```
#mspr5 of log y
mean((predict.model5-log(a.test[,9]))^2)
#0.03866423
mean((predict.model6-log(a.test[,9]))^2)
#0.03875174
sum((predict.model6-log(a.test[,9]))^2)
#53.98118

#
model3=lm(log(rings) ~ shell + sucked + diameter + sex + height + whole + viscera +
        length, data=a.train)
summary(model3)

#
select(lm.ridge(log(rings) ~ diameter + sucked + shell + whole + sex + viscera + height +
           sucked:whole + sucked:sex + whole:height + shell:whole +
           diameter:height + diameter:sucked + diameter:sex + diameter:whole +
           sucked:height, data=a.test, lambda = seq(0,1,0.001)))
#modified HKB estimator is 0.1384103
#modified L-W estimator is 8.947052
#smallest value of GCV  at 0.036
ridge.reg5.test <- lm.ridge(log(rings) ~ diameter + sucked + shell + whole + sex + viscera + height
+
                sucked:whole + sucked:sex + whole:height + shell:whole +
                diameter:height + diameter:sucked + diameter:sex + diameter:whole +
                sucked:height, data=a.test, lambda = 0.036)
ridge.test=as.data.frame(coef(ridge.reg5.test))
```