Fitting a Model for Abalone Age

Pamela Patterson- ppatterson@ucdavis.edu
Tongke Wu- tkwu@ucdavis.edu
Wenqian Li- wqvli@ucdavis.edu

Abstract

In order to determine age of an abalone, scientists have to cut through the shell, dye the flesh, and examine the flesh under a microscope to count the rings. This is time consuming and it would be faster if scientists could predict age based on other, easier to measure, characteristics. Data was collected on several size and weight variables, as well as gender, from approximately 4000 abalone to determine if there is a relationship between these variables and the age of the abalone. We explored several first order and polynomial models, including first order pairwise interaction terms. It appears that the residuals do not follow a normal distribution, as they are right-skewed. The best model we found was a quadratic model with p = 12 and 3 interaction terms. We also discovered that age appears to be a factor, but we did not treat it as a factor in our analysis.

Introduction:

We used data collected about abalone for the purposes of finding relationships between various abalone measurements and their age.

There are 4177 cases and 9 variables. The variables as we coded them in R are as follows:

Sex: M, F, and I (infant)
Length: longest shell measurement (mm)
Diameter: perpendicular to length (mm)
Height: with meat in shell (mm)
Whole: whole abalone weight (grams)
Shucked: weight of meat (grams)
Viscera: gut weight after bleeding (grams)
Shell: weight after being dried (grams)
age: number of rings +1.5

Some questions of interest we had beginning the analysis are:

1. How is the age determined by other characteristics of the abalone?
2. Are all of the weight variables necessary for an adequate model?
3. How do the weight variables and the other variables interact?
4. How will the gender classifications affect age predictions?
5. At what age do abalone develop their gender?
6. What is the distribution of the data (male vs. female, etc)?

Methods and Results:

Data exploration and processing:
We first examined the data and looked at the variable types. The variable types were appropriate for the data, so no changes were needed here. The only categorical variable is the sex of the abalone, with 3 factor levels. We found no missing values, and we changed the last column (the response variable) from “rings” to “age” by adding 1.5 to the values in the column.

By examining the histograms of each variable (Figure 1), we can see that age is right skewed, so a transformation should be explored via box cox. There is nothing very concerning about the distribution of the rest of the variables.

In the pairwise scatterplot matrix for quantitative variables (Figure 2), we can see that there appears to be a nonlinear relationship between each of the weight variables and length and diameter. There doesn’t appear to be any other non-linearity among the variables.

Examining the categorical variable (Figure 3), we can see that the data is evenly distributed between male, female, and infant. We can also see that infants tend to have a lower age, but there is no clear difference in age between male and female.

We then split the data into a training set and validation set using a random 50:50 split. Side-by-side boxplots (Figure 4) show that the distribution of each variable in each set is roughly the same.

After fitting the full first order model with no interaction terms, we performed box cox (Figure 5) to determine what transformation is needed. It appears that a log transformation would suffice since the lambda value is close to 0. The distribution of the transformed response variable (Figure 6) appears more normal like. We performed the log transformation and replaced the age column with log(age). For ease of notation, we still refer to this variable as "age" from this point forward.

Model selection:
Best subsets (Figure 7) of all first order terms, without interactions shows that including all quantitative variables, and only the infant level of the sex variable yields the smallest Cp, AIC, and BIC for a first order model. However, we need to explore other types of models, including interaction terms and higher order terms. Figure 8 shows some comparisons between forward selection of all first order terms, and all first order terms including interaction. It is clearly better to include interaction terms.

Before we explored different orders of models, we fitted a linear regression model, fit6, including all the first-order, interaction, 2nd order polynomial, and 3rd order polynomial, to estimate the error variance σ^2. We obtained the MSE of fit6 as MSE.123 being 0.0247978.

We added interaction terms into fit4 model, with total of 45 estimated regression coefficients. We performed forward stepwise procedure to model fit4, using AIC as the selection criterion, and obtained the model fs4, with 26 estimated regression coefficients left. The model formula follows:
age ~ Shell + Shucked + Diameter + Whole + Sex + Viscera + Height + Length + Shucked:Whole + Shucked:Sex + Shell:Whole + Diameter:Viscera + Whole:Viscera + Diameter:Sex + Diameter:Height + Shell:Height + Shell:Viscera + Diameter:Length + Shucked:Height + Shucked:Length + Viscera:Height + Shucked:Viscera

The residual plot and Q-Q plot of fs4 model (Figure 9) show great improvement compared with previous model. The residuals do not exhibit obvious nonlinearity but still shows a little curvilinear. The Q-Q plot shows obvious right-skewed, indicating a moderate deviation from normality.

Therefore, we added the squared terms into a new model, fit5, with total of 52 estimated regression coefficients. Again, we performed the forward stepwise procedure with AIC being the standard and obtained model fs5, with 26 estimated regression coefficients left.
age ~ Shell + I(Shucked^2) + I(Shell^2) + Shucked + Diameter + I(Diameter^2) + Whole + Sex + Viscera + I(Whole^2) + I(Length^2) + Length + Height + I(Height^2) + Shucked:Whole + Shucked:Sex + Diameter:Viscera + Sex:Length + Diameter:Length + Length:Height + Viscera:Length + Whole:Length

The residual plot and Q-Q plot of fs5 (Figure 10) does not show obvious improvement compared with fs4.

Being curious of whether 3rd order of predictor variables would help improve the model, we added the third-order polynomial terms into a new model, fit6, with total of 59 estimated regression coefficients. The forward selection procedure with AIC being the standard gave the model fs6, with 31 estimated regression coefficients. Again, the residual plot and Q-Q plot of fs6 (Figure 11) look good but still does not show obvious improvement compared to fs5.

Considering fs6 as a full model, we performed the best subset procedure to find out whether some terms can be removed from the model to decrease model complexity. Using regsubsets function, we got the top 1 best subset of all subset sizes up to 31. Then we got SSE, R^2, Ra^2, Cp, AIC, and BIC for each model (Figure 12, Figure 13). If we take BIC as the standard, p=12 model has the smallest BIC. For other parameters, even though they increase not much in p>12 models compared with p=12 model, it seems p=20 model is better if we take SSE, R^2, Ra^2, Cp, and AIC as standards. Therefore, we further analyzed these two models.

We noticed that in p=12 and p=20 models, only SexI is included in the model. But if we put Sex, a 3-level factor, and other terms into a model, we would get p > 12 or p>20. Therefore, we transformed Sex into a two-level factor variable SexI with infant being 1 and F/M being 0. P=12 model is age ~ Shell + Whole + Shucked + Viscera + Diameter +

I(Shucked^2) + I(Whole^2) + I(Diameter^2) + Viscera:Diameter + Shucked:SexI + Diameter:SexI. P=20 model is log(age) ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) + I(Shell^2) + I(Diameter^2) + I(Height^2) + I(Shucked^3) + I(Diameter^3) + I(Height^3) + I (Shell^3) + Whole:Shucked + Viscera:Diameter + Shucked:SexI + Diameter:SexI + Shucked:Height + Shell:Shucked. The residual plot and Q-Q plots of these two models are shown in (Figure 14) and (Figure 15) respectively. They are comparable with previous models, but p=12 and p=20 models have less variables in the model, hence smaller overall in-sample variance. Therefore, we will use p=12 and p=20 models for further validation.

Diagnostic for Candidate Models

Figure 16 and 17 are diagnostics for Model 12 and Model 20 respectively. As they show, there is no nonlinearity in residuals and they spread out evenly in both models. Yet the distribution of residuals are both right skewed though we have already transformed the response variable.

Validation

Validation for Model with 12 parameters

After fitting the model on training data, we obtained the fitted regression coefficients and displayed them in Table 1. Compared with the model regressed on training data, there are same signs and similar magnitude between the two sets of estimated coefficients and standard errors. Then we calculated SSE, MSE, R_a^2, Press and MSPE of both models and compared them in Table 2. We found MSPE of validation model is not much larger than MSE and Press/n based on the training data, indicating the Model 12 is stable in the fitted regression coefficients, plausible in model function and well able to predict.

Validation for Model with 20 parameters

As Table 3 shows, validation Model 20 has three fitted regression coefficients changing sign from training data to validation data. So it is eliminated from further consideration. Therefore, Model 12 is chosen as the final model. We refitted the model with the full dataset and obtained the following model equation:

age = 1.150 + 0.793 x Shell + 1.725 x Whole - 3.419 x Shucked - 3.305 x Viscera + 6.097 x Diameter + 1.678 x $Shucked^2$- 0.387 x $Whole^2$ - 8.025 x $Diameter^2$ + 5.640 x Viscera x Diameter + 0.816 x Shucked x SexI - 0.744 x Diameter x SexI.

Diagnostic for Final Model

Check Assumption

In Figure 18, there is no nonlinear pattern in residuals, indicating the linear function is appropriate. Since the residuals spread out around 0 evenly, the assumption that error terms have constant variance is appropriate. From the Q-Q plot, the distribution of residuals is still skewed to the right after log transformation of the response variable. The assumption that error terms are normally distributed cannot hold, which may lead to imprecise inference in the future.

Identify Outlying Cases

We calculated all the studentized deleted residuals and conducted Bonferroni's procedure.

The Bonferroni's threshold at level 0.05 is 4.383 then case 2184 and 481 are identified as outlying Y cases. The result is displayed in Figure 19.
As for X value outliers, we obtained the leverage values of all the cases and compared them with the threshold 2p/n = 0.0057, then 333 cases are identified as outlying with regard to its X values.

In order to find influential cases, we performed Cook's distance and detected 107 outlying cases showed in Figure 20 that may have substantial influence on the fitted values.

After removing all the influential cases, we refitted the model and got its equation
age = 1.200 + 0.665 x Shell + 2.311 x Whole - 4.372 x Shucked - 3.501 x Viscera + 5.744 x Diameter + 2.384 x $Shucked^2$ - 0.548 x $Whole^2$ - 7.726 x $Diameter^2$ + 5.468 x Viscera x Diameter + 0.810 x Shucked x SexI - 0.722 x Diameter x SexI
Comparing critical values in model with outliers and model without outliers, we find the model is better to fit the data after dropping outliers. The diagnostic can be found in Figure 21.

Conclusion

Our fitted model shows that the age of abalone is determined by Shell, Whole, Shucked, Viscera, Diameter and maturity of abalone, in linear terms of Shell, Whole, Shucked, Viscera and Diameter, quadratic terms of Shucked, Whole and Diameter and interactions between Viscera and Diameter, Shucked and Infant, Diameter and Infant. All the weight variables are necessary in this model. Yet gender classifications of Male and Female do not affect the age predictions, whether abalone is mature, saying infant or adult, does determine the age.
Since the dataset has over 5000 cases and only 107 of them are identified as outliers and removed, our model can be applied in a general way. However, since the distribution of residuals is right skewed, indicating the assumption of normally distributed errors cannot hold, further statistical inference, such as confidence intervals, will be less precise.

Further Discussion

During our work, we found interesting pattern of lines in all the residuals vs fitted values plots. We believe the fact that age is actually a categorical variable, as Figure 22 showing, accounted for the pattern.

In addition, although we have already transformed the response variable by log, the residuals are still right skewed. We need further research to figure out better remedial measures.

Appendix 1:

Figure 1: Histograms of quantitative variables

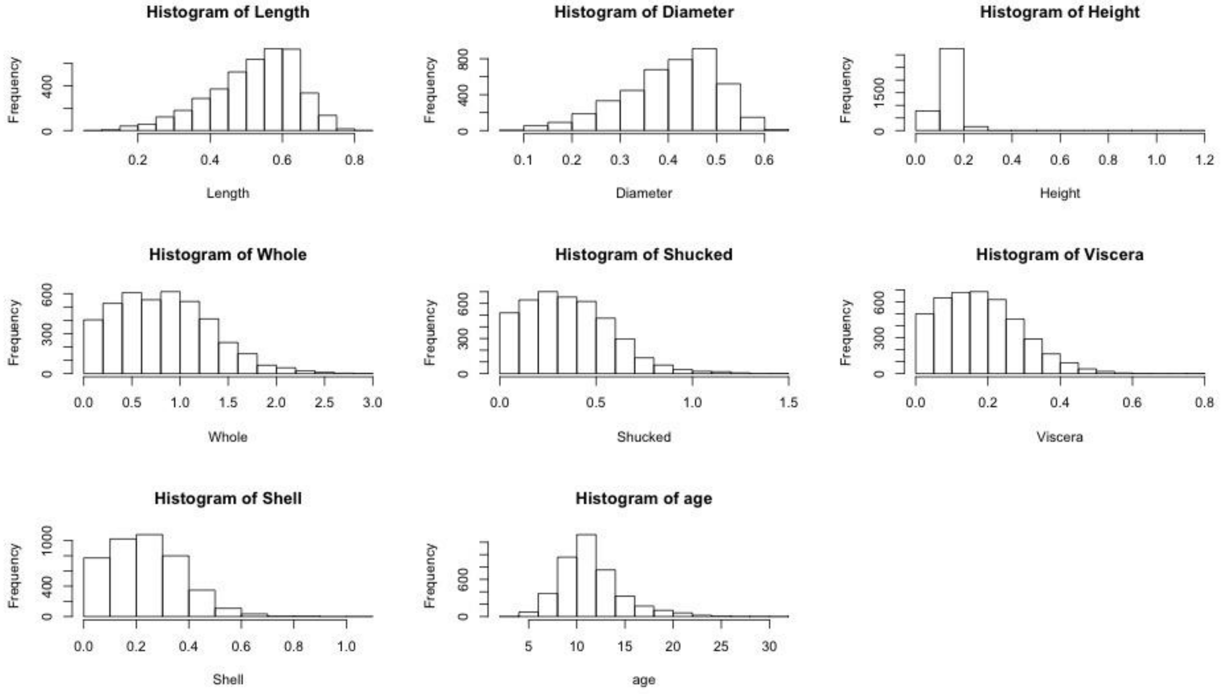


Figure 2: Pairwise scatterplot matrix of quantitative variables

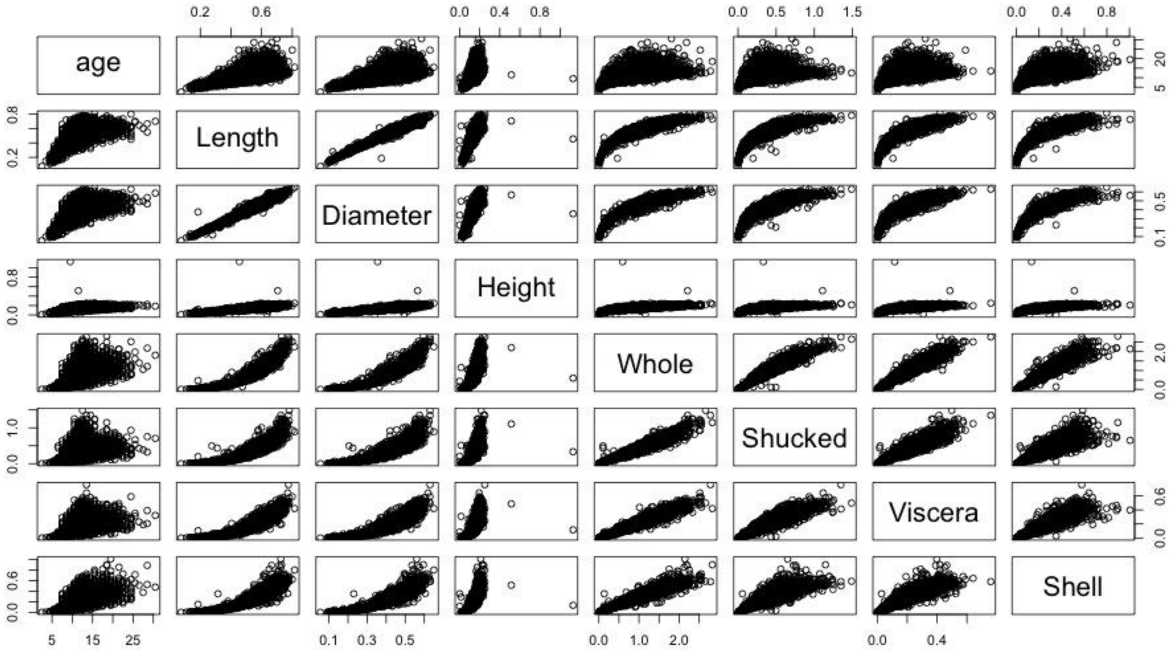


Figure 3: Pie chart and side-by-side box plot of categorical variable

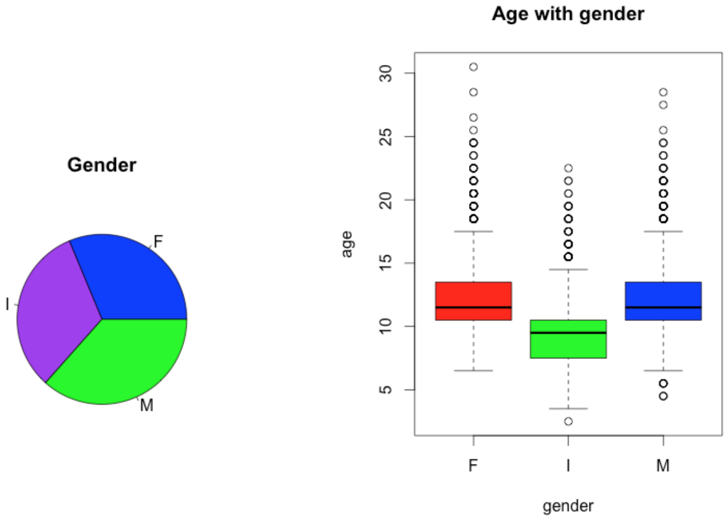


Figure 4: Side-by-side box plot of quantitative variables: training vs. validation

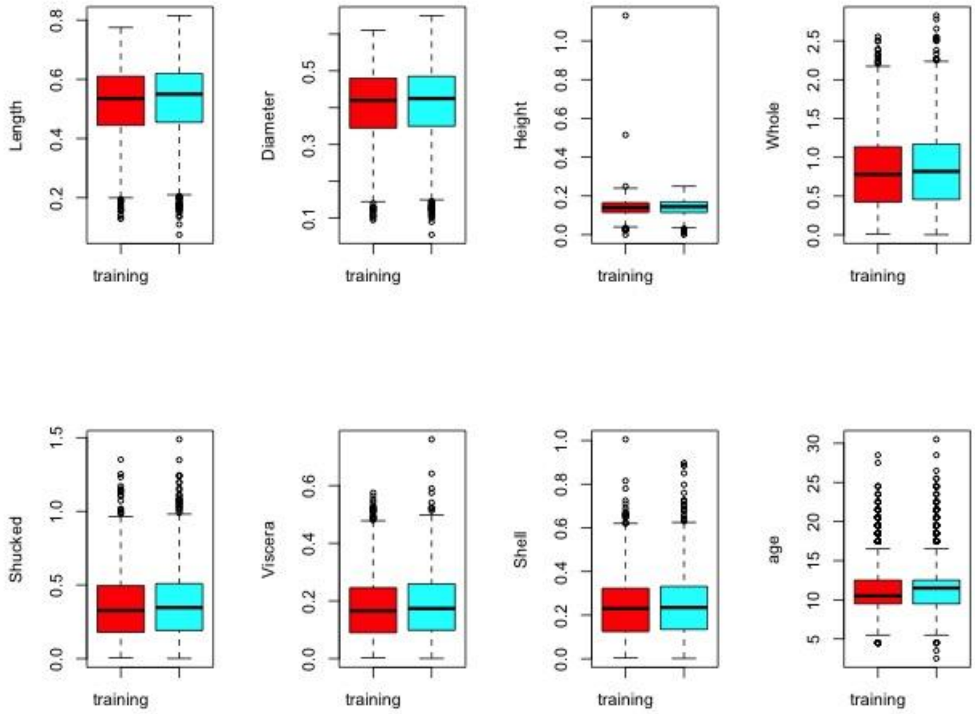


Figure 5: Box cox transformation

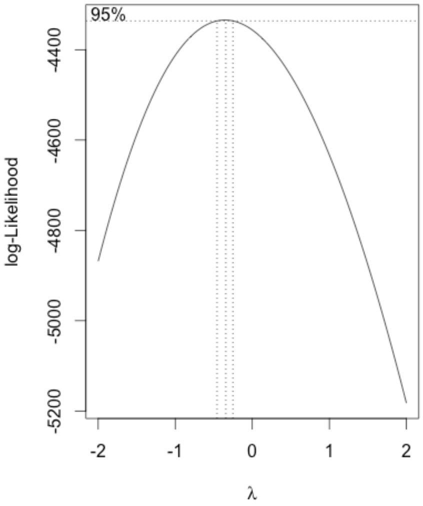


Figure 6: Histogram of age after log transformation

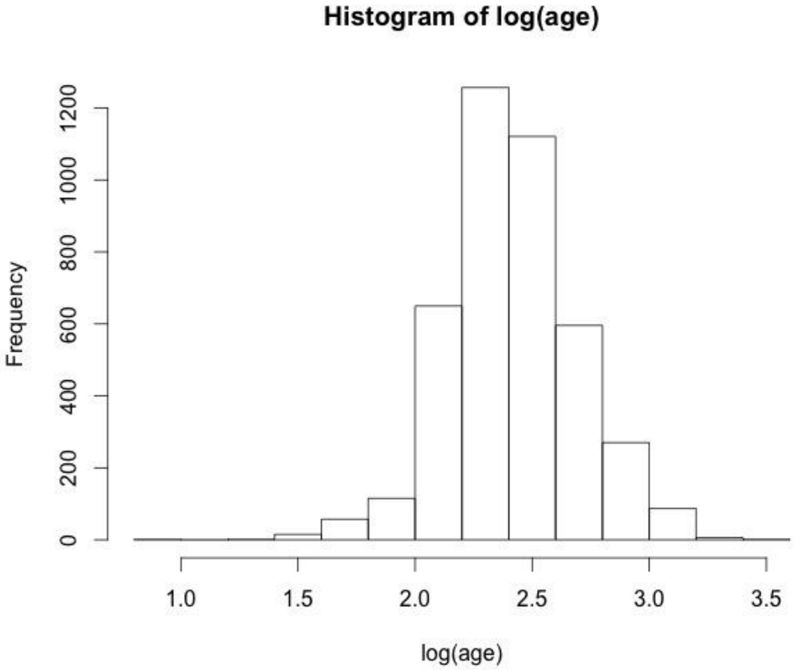


Figure 7: Output for best subsets on all first order variables, no interaction terms

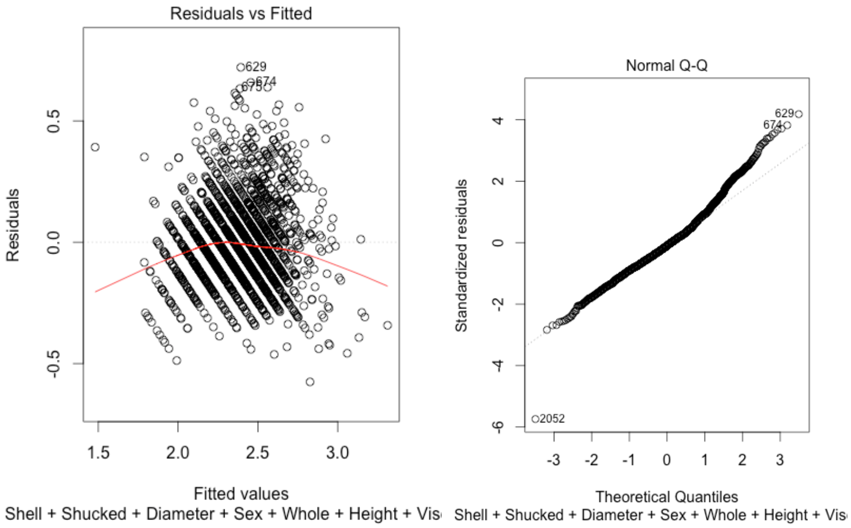
Best subsets:

	SexI	SexM	Len.	Diam.	Height	Whole	Shuck.	Visc.	Shell
1	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	1	0	1
3	0	0	0	1	0	1	1	0	0
4	1	0	0	1	0	0	1	0	1
5	1	0	0	1	0	1	1	0	1
6	1	0	0	1	1	1	1	0	1
7	1	0	0	1	1	1	1	1	1
8	1	0	1	1	1	1	1	1	1

	SSE	R2	Ra2	Cp	BIC	AIC
1	83.25	0.46	0.46	710.57	-16338.2	-16359.87
2	76.66	0.50	0.50	491.24	-16674.52	-16693.53
3	66.72	0.56	0.56	159.70	-17246.06	-17271.41
4	64.20	0.58	0.58	76.99	-17398.76	-17430.44
5	63.21	0.59	0.59	45.62	-17455.62	-17493.64
6	62.65	0.59	0.59	29.08	-17483.93	-17561.02
7	62.13	0.59	0.59	13.67	-17510.32	-17561.02
8	61.96	0.60	0.59	9.89	-17513.58	-17570.61

Figure 8

Forward stepwise selection of first order terms:
Step: AIC=-7326.31
age ~ Shell + Shucked + Diameter + Sex + Whole + Height + Viscera + Length



Forward stepwise for Model 2 (all variables and interaction terms-first order)
Step: AIC=-7620.31
age ~ Shell + Shucked + Diameter + Whole + Sex + Viscera + Height + Length + Shucked:Whole + Shucked:Sex + Shell:Whole + Diameter:Viscera +Whole:Viscera + Diameter:Sex + Diameter:Height + Shell:Height + Shell:Viscera + Diameter:Length + Shucked:Height + Shucked:Length + Viscera:Height + Shucked:Viscera

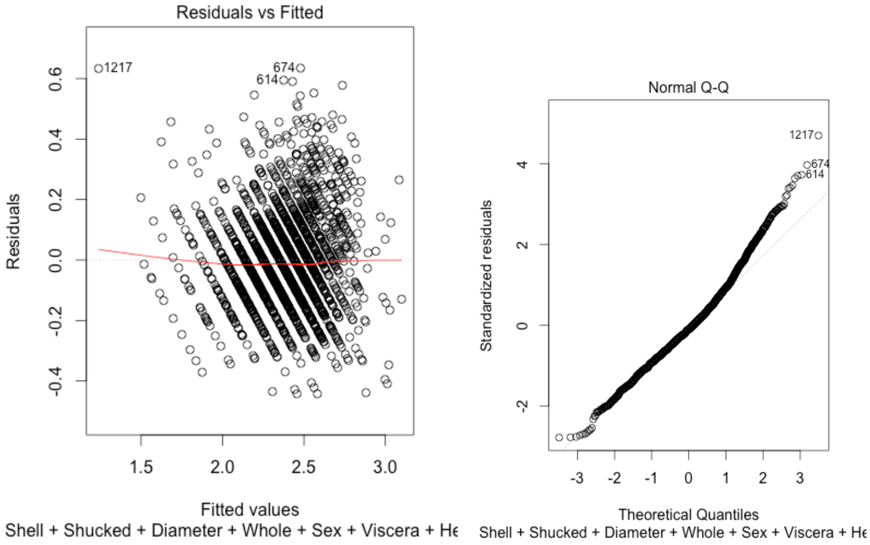


Figure 9. Residual plot and Q-Q plot of model fs4.

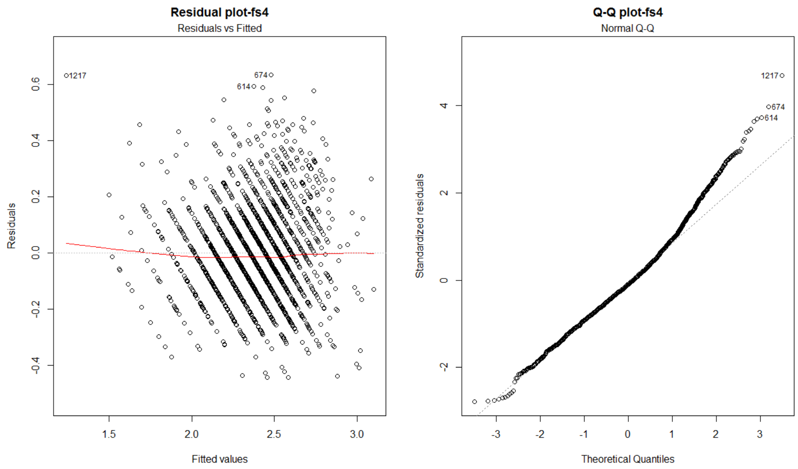


Figure 10. Residual plot and Q-Q plot of model fs5.

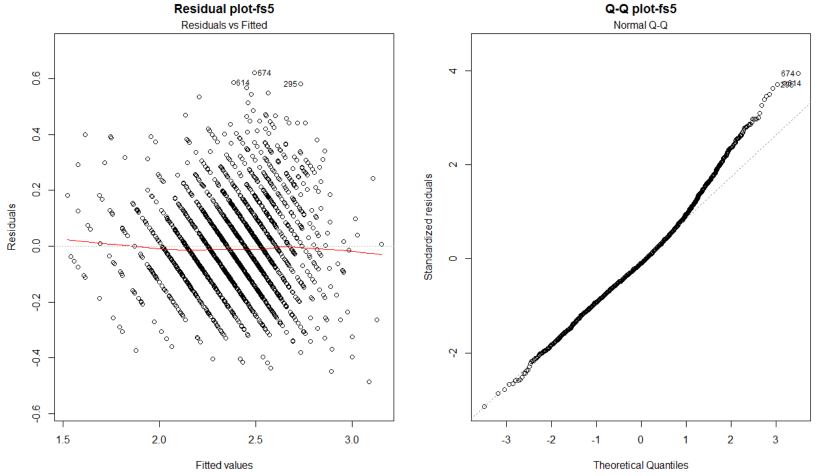


Figure 11. Residual plot and Q-Q plot of model fs6.

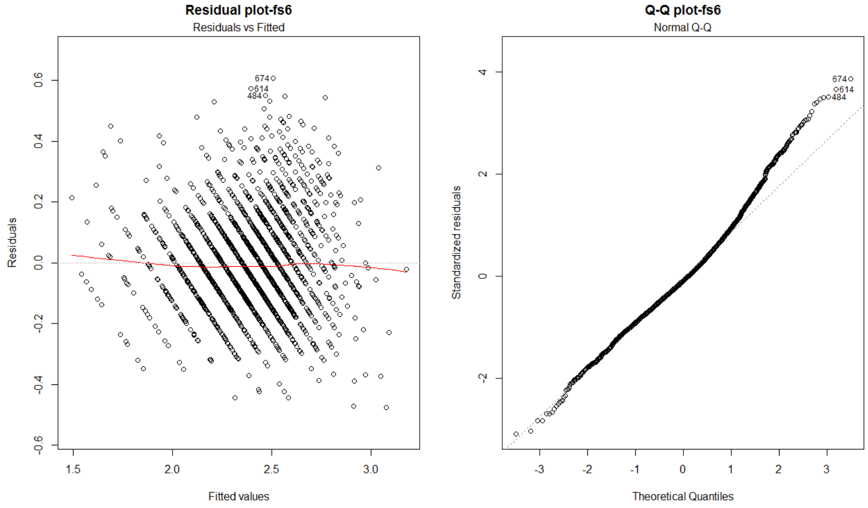


Figure 12. Parameters of top 1 best subset of all subset sizes from p=2 to p=31.

	SSE	R2	Ra2	Cp	AIC	BIC
1	83.2535	0.4570	0.4567	1267.6430	-6723.6863	-6712.3983
2	68.1809	0.5553	0.5549	662.8459	-7138.7136	-7121.7818
3	64.0604	0.5822	0.5816	498.9609	-7266.8761	-7244.3003
4	61.5167	0.5988	0.5980	398.5565	-7349.4766	-7321.2568
5	59.9122	0.6092	0.6083	335.9619	-7402.6596	-7368.7958
6	57.8759	0.6225	0.6214	255.9841	-7472.8606	-7433.3528
7	56.6144	0.6307	0.6295	207.1971	-7516.8764	-7471.7247
8	55.3317	0.6391	0.6377	157.5606	-7562.7253	-7511.9297
9	54.3148	0.6457	0.6442	118.6216	-7599.4564	-7543.0168
10	53.6260	0.6502	0.6485	92.8904	-7624.1062	-7562.0226
11	53.0104	0.6542	0.6524	70.1084	-7646.2131	-7578.4855
12	52.7066	0.6562	0.6542	59.8769	-7656.2147	-7582.8432
13	52.5481	0.6573	0.6551	55.4987	-7660.5006	-7581.4851
14	52.3485	0.6586	0.6563	49.4614	-7666.4485	-7581.7891
15	52.0616	0.6604	0.6580	39.9101	-7675.9246	-7585.6212
16	51.9386	0.6612	0.6586	36.9609	-7678.8609	-7582.9136
17	51.8210	0.6620	0.6592	34.2247	-7681.5958	-7580.0045
18	51.7147	0.6627	0.6598	31.9463	-7683.8822	-7576.6469
19	51.5336	0.6639	0.6608	26.6553	-7689.2073	-7576.3281
20	51.3804	0.6649	0.6616	22.4871	-7693.4244	-7574.9012
21	51.3292	0.6652	0.6618	22.4266	-7693.5054	-7569.3382
22	51.2448	0.6658	0.6622	21.0266	-7694.9437	-7565.1326
23	51.2003	0.6661	0.6623	21.2357	-7694.7571	-7559.3020
24	51.1749	0.6662	0.6623	22.2148	-7693.7915	-7552.6925
25	51.1585	0.6663	0.6623	23.5556	-7692.4596	-7545.7166
26	51.1299	0.6665	0.6623	24.4012	-7691.6303	-7539.2434
27	51.1144	0.6666	0.6622	25.7791	-7690.2615	-7532.2306
28	51.1013	0.6667	0.6622	27.2514	-7688.7970	-7525.1221
29	51.0968	0.6667	0.6620	29.0691	-7686.9821	-7517.6632
30	51.0951	0.6667	0.6619	31.0000	-7685.0522	-7510.0894

Figure 13. Plots of parameters of best subsets models of different sizes.

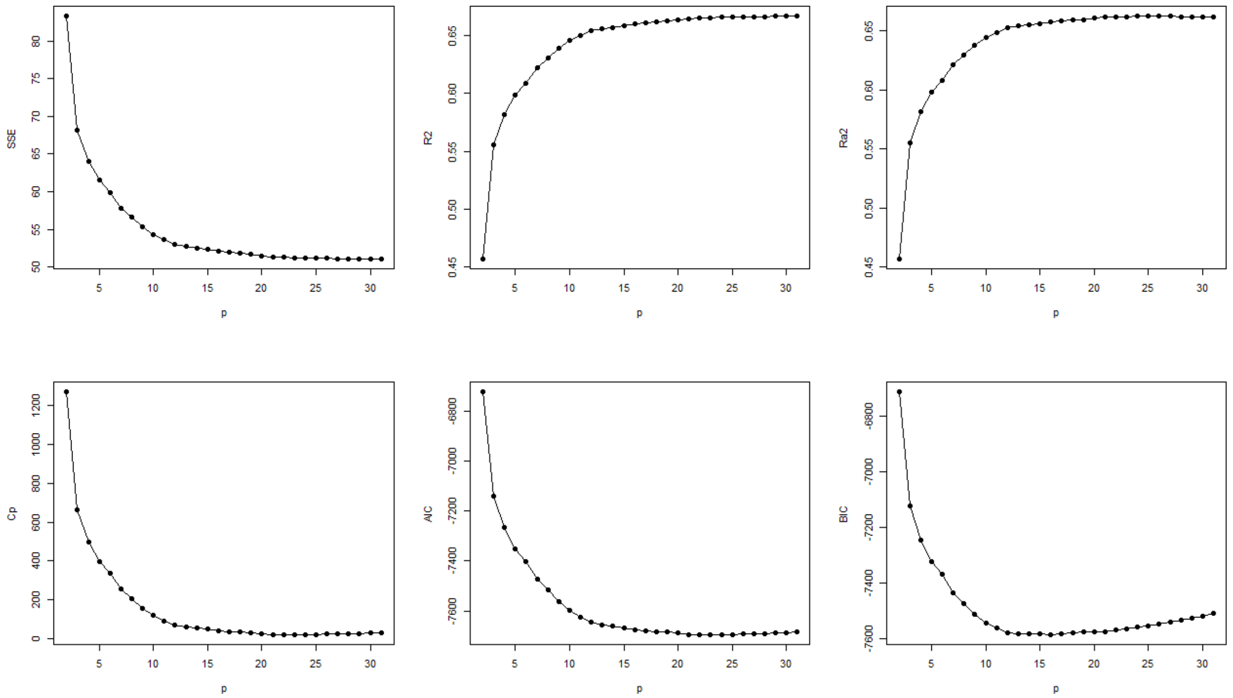


Figure 14. Residual plot and Q-Q plot of model p=12.

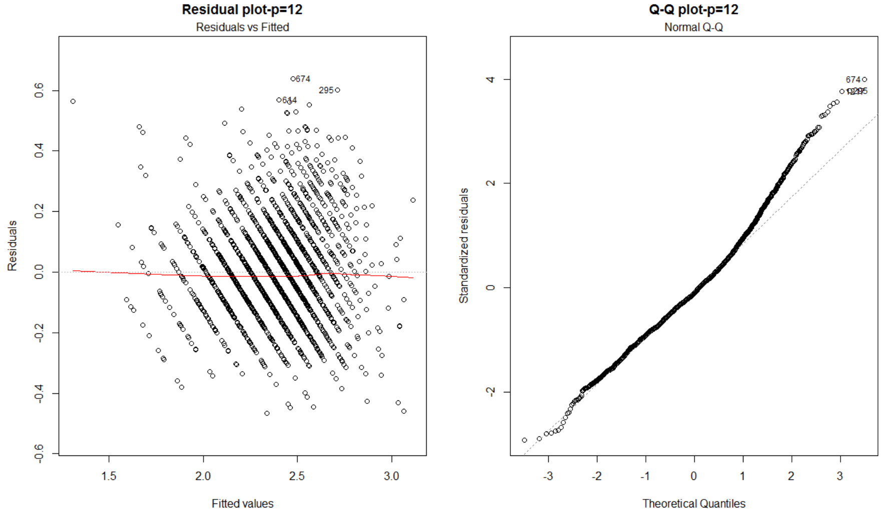


Figure 15. Residual plot and Q-Q plot of model p=20.

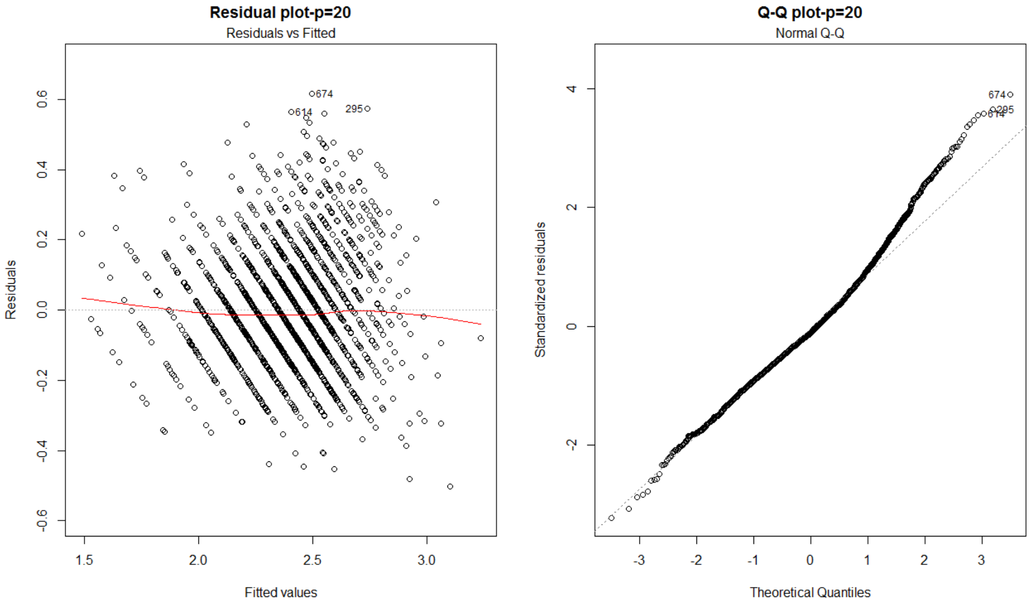


Figure 16: for model 12

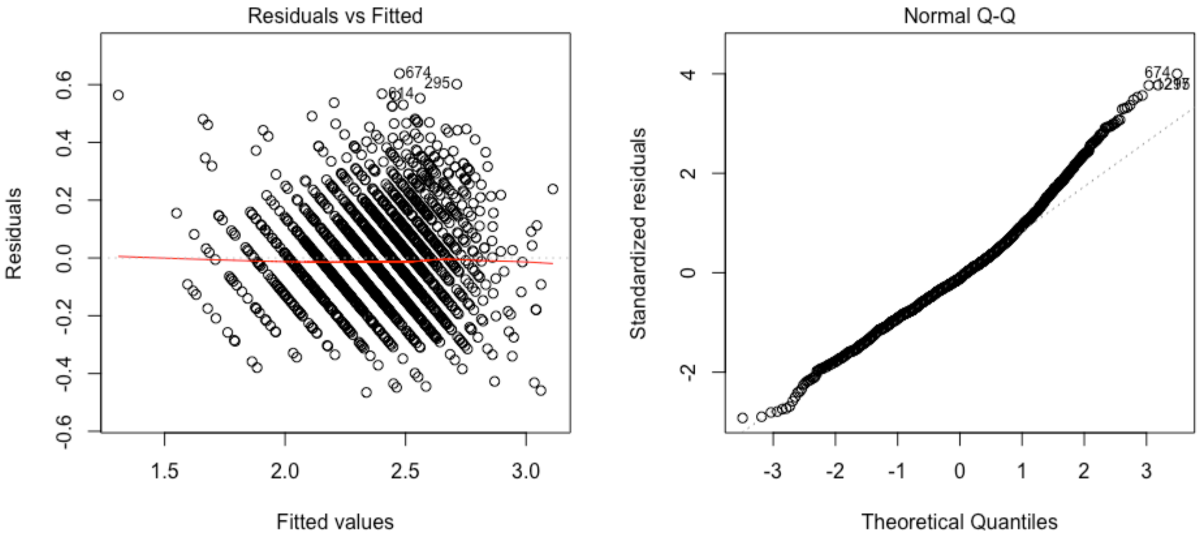


Figure 17: for model 20

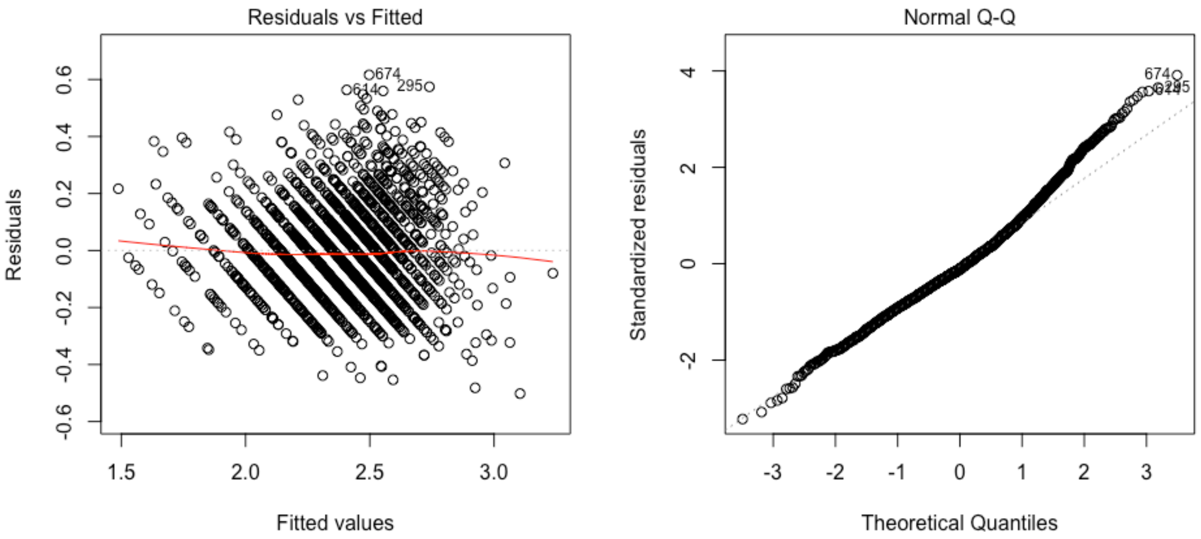


Table 1: Coefficients of Validation Model 12

	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	1.064	0.068	1.209	0.062
Shell	0.749	0.115	0.837	0.122
Whole	2.15	0.157	1.448	0.149
Shucked	-3.537	0.187	-3.384	0.189
Viscera	-5.996	0.846	-1.466	0.74
Diameter	7.081	0.514	5.401	0.471
I(Shucked^2)	1.859	0.145	1.563	0.138
I(Whole^2)	-0.565	0.053	-0.27	0.047
I(Diameter^2)	-10.68	0.984	-6.108	0.898

Viscera:Diameter	11.288	1.705	1.726	1.49
Shucked:SexI1	0.794	0.097	0.832	0.098
Diameter:SexI1	-0.733	0.074	-0.749	0.076

Table 2: Comparision of Critical Values

	SSE	MSE	R2_a	Press	Press/n	MSPE
Training	53.01	0.026	0.652	53.804	0.026	--
Validation	56.552	0.027	0.638	57.578	0.028	0.028

Table 3: Coefficients of Validation Model 20

	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	0.844	0.111	0.908	0.095
Shell	2.045	0.461	3.5	0.545
Whole	1.881	0.171	0.831	0.21
Shucked	-3.49	0.336	-3.763	0.36
Viscera	-4.308	0.901	0.431	0.86
Diameter	9.281	1.075	8.75	0.949
I(Shucked^2)	4.96	0.585	4.718	0.606
I(Shell^2)	-4.712	1.206	-3.603	1.3
I(Diameter^2)	-20.676	3.608	-20.085	3.257
I(Height^2)	11.074	2.375	9.98	4.556

I(Shucked^3)	-0.651	0.236	-1.205	0.198
I(Diameter^3)	12.092	3.845	16.417	3.506
I(Height^3)	-8.584	1.85	-10.968	18.976
I(Shell^3)	2.286	0.807	2.946	0.884
Whole:Shucked	-2.14	0.273	-0.25	0.339
Viscera:Diameter	7.69	1.825	-2.234	1.747
Shucked:SexI1	0.731	0.101	0.853	0.1
Diameter:SexI1	-0.67	0.077	-0.74	0.078
Shucked:Height	-4.606	1.087	-2.597	1.599
Shell:Shucked	1.973	0.662	-2.854	0.765

Figure 18: For model 12

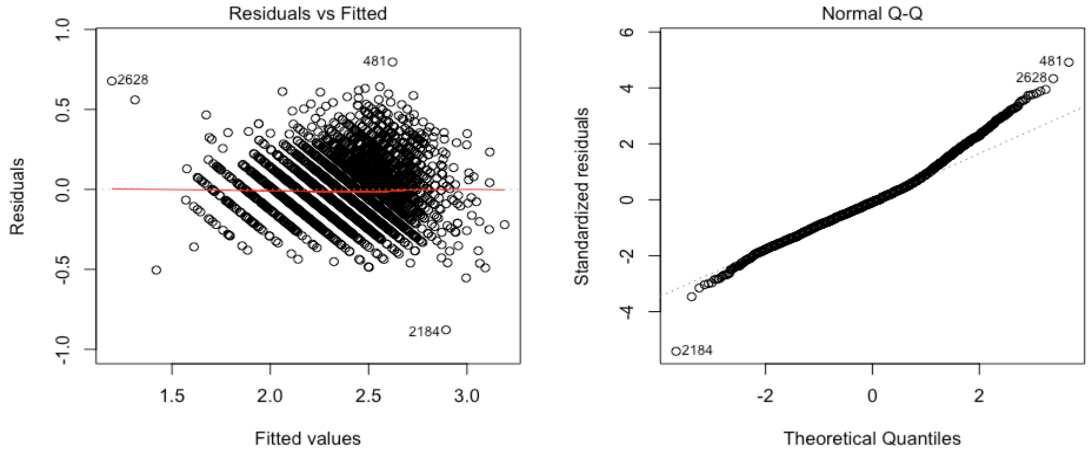


Figure 19: for model 12

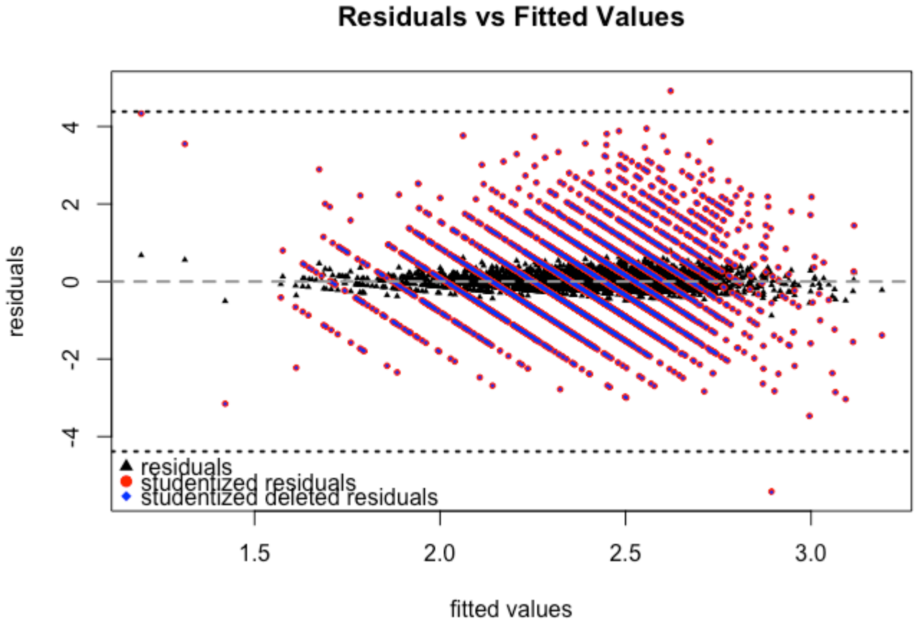


Figure 20: For model 12

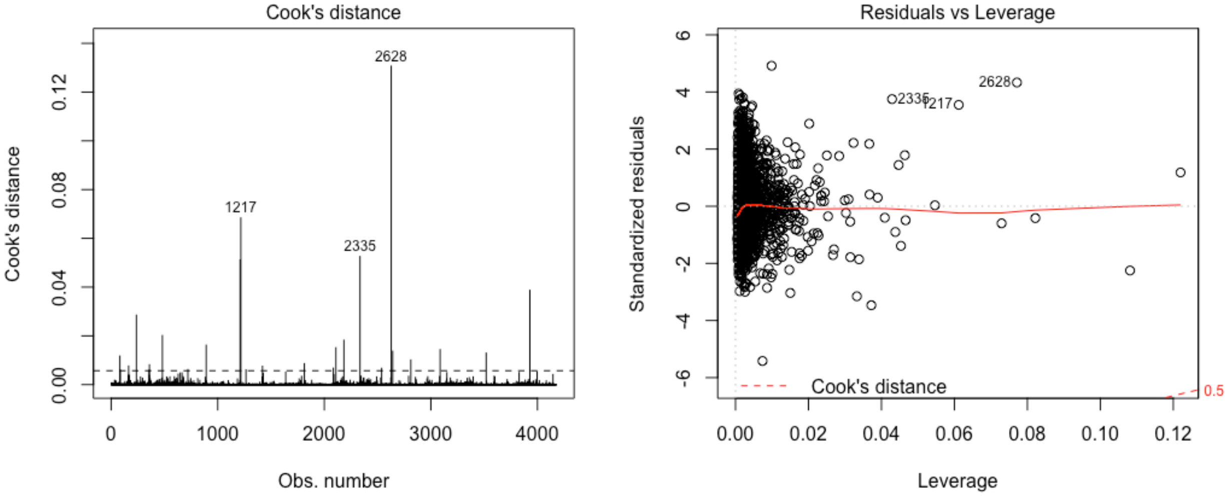


Table 4: Comparison of Critical Values

	SSE	MSE	R^2_a
Model with outliers	110.364	0.026	0.643
Model without outliers	98.83	0.024	0.651

Figure 21: For model 12

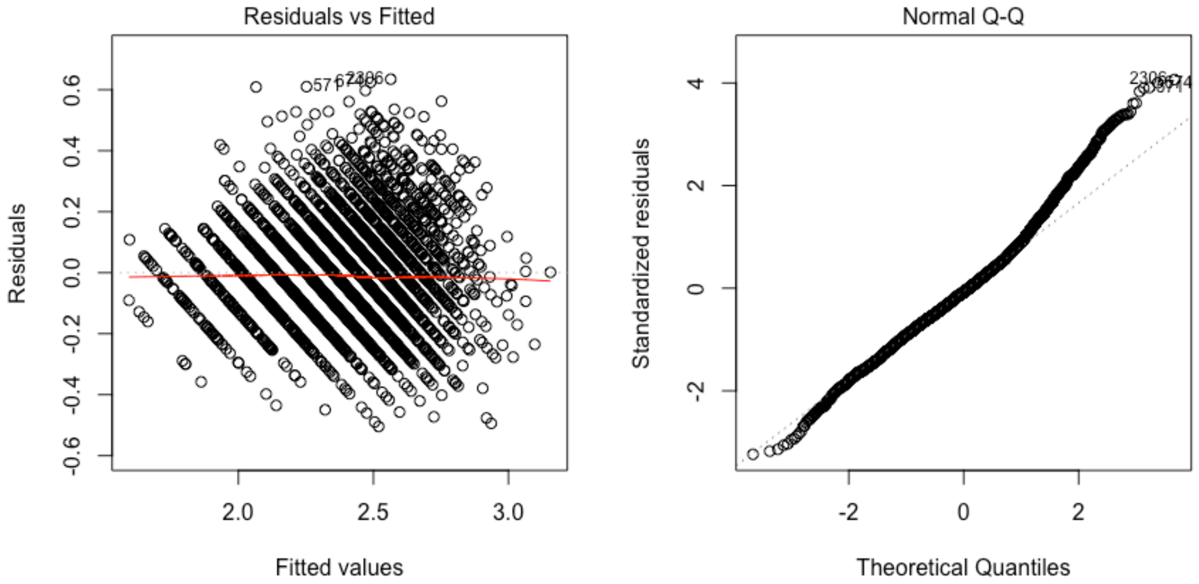
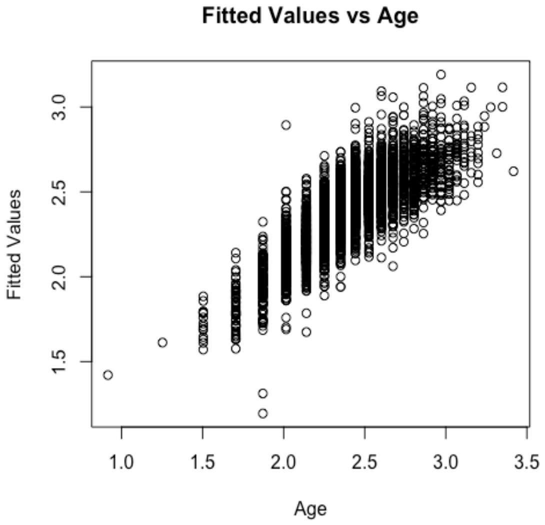


Figure 22:



Appendix 2: R Code

```
library(MASS)
library(leaps)
abalone = read.csv("abalone.csv", header=FALSE)
names(abalone) =
c("Sex","Length","Diameter","Height","Whole","Shucked","Viscera","Shell","Rings")
abalone$age = abalone$Rings + 1.5
n = nrow(abalone)
drops = c("Rings") #drop Rings from data
abalone = abalone[,!(names(abalone)%in%drops)]

sapply(abalone, class) #get variable type
sapply(abalone, function(x) sum(is.na(x))) #check missing values
summary(abalone) #summary statistics

par(mfrow = c(3,3))
for(i in 2:9) {
  hist(abalone[,i], xlab= names(abalone)[i], main = paste("Histogram of",
names(abalone)[i]))
} #histograms of variables

#pairwise scatter plot of quantitative variables
par(mfrow = c(1,1))
pairs(~age + Length + Diameter + Height + Whole + Shucked + Viscera + Shell, data =
abalone) #pairwise scatter plots
#pairwise correlation matrix of quantitative variables
round(cor(abalone[,-1]),2)
#pie chart of qualitative variable
par(mfrow = c(1,1))
pie(table(abalone$Sex), col = c('blue','purple','green'), main = 'Gender')

#boxplot of qualitative variable vs. Age
boxplot(abalone$age~abalone$Sex, main = 'Age with gender', xlab = 'gender', ylab =
'age', col = rainbow(3))

#transformation of age
par(mfrow = c(2,2))
hist(log(abalone$age), xlab = "log(age)", main = "Histogram of log(age)")
hist(sqrt(abalone$age),xlab = "square root age", main = "Histogram of square root age")
hist(1/abalone$age,xlab = "1/age", main = "Histogram of 1/age")

#split data
set.seed(10)
index=sample(1:n, size = 4176/2, replace = FALSE)
abalone.c = abalone[index,]
abalone.v = abalone[-index,]
```

```
#side-by-side box plots of training vs validation data
vars = c("age", "Length", "Diameter", "Height", "Whole", "Shucked", "Viscera", "Shell")
train.box = abalone.c[,(names(abalone.c)%in%vars)]
vali.box = abalone.v[,(names(abalone.v)%in%vars)]

par(mfrow = c(2,4))
for(i in 1:8){
  boxplot(cbind(train.box[,i],vali.box[,i]),names=c("training","validation"),ylab=names(trai
n.box[i]),col=rainbow(2))
}

#try different orders of the model:
fit1 = lm(age ~ ., data = abalone.c) #full model, no transformation
fitp = lm(age ~ Sex + Length + Diameter + Height+ poly(Whole, 2, raw=TRUE) +
poly(Shucked, 2, raw=TRUE) + poly(Viscera, 2, raw=TRUE) + poly(Shell, 2, raw=TRUE),
data = abalone.c) #quadratic model
summary(fit1)
summary(fitp)
par(mfrow = c(1,1))
boxcox(fit1)

#logtransformation of age:
abalone$age = log(abalone$age) #perform log transformation of age
abalone.c$age = log(abalone.c$age)
abalone.v$age = log(abalone.v$age)
#plots after transformation
par(mfrow = c(1,1))
hist(abalone$age, xlab="log(age)", main = "Histogram of log(age)")
pairs(~age + Length + Diameter + Height + Whole + Shucked + Viscera + Shell, data =
abalone) #pairwise scatter plots
boxplot(abalone$age~abalone$Sex, main = 'logage with gender', xlab = 'gender', ylab =
'age', col = rainbow(3))
#fit full model after transformation
fit_full = lm(age ~ ., data = abalone.c)
summary(fit_full)
hist(abalone.c$age)

#null model
none_mod = lm(age~1, data = abalone.c)
#best subsets
sub_set = regsubsets(age ~ ., data = abalone.c, nbest = 1, nvmax = 8 , method =
"exhaustive")
sum_sub = summary(sub_set)
p.m = as.integer(rownames(sum_sub$which))+1
ssto = sum((abalone.c$age - mean(abalone.c$age))^2)
sse = (1 - sum_sub$rsq)*ssto
```

```
aic = n*log(sse/n)+2*p.m
bic = n*log(sse/n)+log(n)*p.m
res_sub = cbind(sum_sub$which, sse, sum_sub$rsq, sum_sub$adjr2, sum_sub$cp, bic,
aic)
colnames(res_sub)[12]= "R2"
colnames(res_sub)[13]="Ra2"
colnames(res_sub)[14]="Cp"
round(res_sub,2)
#forward stepwise
fs1 = stepAIC(none_mod, scope = list(upper=fit_full), direction = "both", k = 2)
par(mfrow = c(1,1))
plot(fs1, which = 1)  #residuals vs. fitted values
plot(fs1, which = 2) #residuals Q-Q plot
#Model 2
fit2 = lm(age~.+.*., data = abalone.c)
summary(fit2)
#forward stepwise Model 2
fs2 = stepAIC(none_mod, scope = list(upper=fit2), direction = "both", k = 2)
par(mfrow = c(1,1))
plot(fs2, which = 1)  #residuals vs. fitted values
plot(fs2, which = 2) #residuals Q-Q plot
#forward selection
fs3 = stepAIC(none_mod, scope = list(upper=fit2), direction = "forward", k = 2)
#Model 3
fit3 = lm(age~Shell + Shucked + Diameter + Whole + Sex + Viscera + Height + Length +
Shucked:Whole + Shucked:Sex + Shell:Whole + Diameter:Viscera + Whole:Viscera +
Diameter:Sex + Diameter:Height + Shell:Height + Shell:Viscera + Diameter:Length +
Shucked:Height + Shucked:Length + Viscera:Height + Shucked:Viscera, data =
abalone.c)
summary(fit3)
sum_fs1 = summary(fs1)
sum_fs2 = summary(fs2)
n.c = nrow(abalone.c)
#info for fs1
sse_fs1 = (1 - 0.5962)*ssto
mse_full = 0.02979076
cp_fs1 = (sse_fs1/mse_full)-(n.c-2*9)
press_fs1 = sum((fs1$residuals/(1-influence(fs1)$hat))^2)
#info for fs2
sse_fs2 = (1 - 0.6546)*ssto
cp_fs2 = (sse_fs2/mse_full)-(n-2*23)
press_fs2 = sum((fs2$residuals/(1-influence(fs2)$hat))^2)

abalone = read.csv("Abalone.txt", header=FALSE)
```

```
names(abalone) =
c("Sex","Length","Diameter","Height","Whole","Shucked","Viscera","Shell","Rings")
abalone$age = abalone$Rings + 1.5
n = nrow(abalone)
drops = c("Rings") #drop Rings from data
abalone = abalone[,!(names(abalone)%in%drops)]

#split data
set.seed(10)
index=sample(1:n, size = 4176/2, replace = FALSE)
abalone.c = abalone[index,]
abalone.v = abalone[-index,]

#log transformation of age defined by previous results:
abalone$age = log(abalone$age) #perform log transformation of age
abalone.c$age = log(abalone.c$age)
abalone.v$age = log(abalone.v$age)

#null model
none_mod = lm(age~1, data = abalone.c)

fit6 = lm(age ~ . + .^2 + I(Length^2) + I(Diameter^2) + I(Height^2) + I(Whole^2)
+I(Shucked^2) + I(Viscera^2) + I(Shell^2) +I(Length^3) + I(Diameter^3) + I(Height^3) +
I(Whole^3) +I(Shucked^3) + I(Viscera^3) + I(Shell^3), data= abalone.c)
MSE.123 = summary(fit6)$sigma^2 #fit6 as the full model to calculate MSE
MSE.123

# All first order and interaction
fit4 = lm(age ~ . + .^2, data= abalone.c)
#the number of coefficients
nrow(summary(fit4)$coef)

# forward stepwise
fs4 = stepAIC(none_mod, scope = list(upper = fit4), direction = "both", k = 2 )

nrow(summary(fs4)$coef) #26 terms included
summary(fs4)

windows()
par(mfrow=c(1,2))
```

```
plot(fs4.both, which = 1, main = "Residual plot-fs4")
plot(fs4.both, which = 2, main = "Q-Q plot-fs4")

#-------------------------
#2nd polynomial with interaction
fit5 = lm(age ~ . + .^2 + I(Length^2) + I(Diameter^2) + I(Height^2) + I(Whole^2)
+I(Shucked^2) + I(Viscera^2) + I(Shell^2), data= abalone.c)
nrow(summary(fit5)$coef) #52 coefficients
#forward stepwise
fs5 = stepAIC(none_mod, scope = list(upper = fit5), direction = "both", k = 2 )
#AIC = -7686.67
summary(fs5)
nrow(summary(fs5)$coef) #26

windows()
par(mfrow=c(1,2))
plot(fs5, which = 1,main = "Residual plot-fs5")
plot(fs5, which = 2, main = "Q-Q plot-fs5")

#-----------------------------------
#first order + interaction + 2nd poly + 3rd poly
fit6 = lm(age ~ . + .^2 + I(Length^2) + I(Diameter^2) + I(Height^2) + I(Whole^2)
+I(Shucked^2) + I(Viscera^2) + I(Shell^2) +I(Length^3) + I(Diameter^3) + I(Height^3) +
I(Whole^3) +I(Shucked^3) + I(Viscera^3) + I(Shell^3), data= abalone.c)
nrow(summary(fit6)$coef)

fs6 = stepAIC(none_mod, scope = list(upper = fit6), direction = "forward", k = 2 )
nrow(summary(fs6)$coef) #31
#AIC = -7692.63


windows()
par(mfrow=c(1,2))
plot(fs6, which = 1, main = "Residual plot-fs6")
plot(fs6,which = 2, main = "Q-Q plot-fs6")


######################################
```

```
#best subsets based on fs6 as the full model
sub_fs6 = regsubsets(age ~ Shell + I(Shucked^2) + I(Shell^2) + I(Shucked^3) + Whole
+ factor(Sex) + Shucked + I(Whole^3) + I(Whole^2) + Viscera + Diameter +
I(Diameter^2) + I(Diameter^3) + I(Length^3) + Height + I(Height^2) + I(Height^3) +
I(Shell^3) + Whole:Sex + Whole:Shucked + Sex:Shucked + Whole:Viscera +
Sex:Diameter + Viscera:Diameter + Shucked:Height + Shell:Shucked, data = abalone.c,
nbest = 1, nvmax = 31, method = "exhaustive")
sum_sub_fs6 = summary(sub_fs6)
p.m_sub_fs6 = as.integer(rownames(sum_sub_fs6$which))+1
n.c = nrow(abalone.c)

#Get SSE, R^2, Ra^2, Cp, AIC, BIC
sse.sub_fs6 = sum_sub_fs6$rss
rsq.sub_fs6 = sum_sub_fs6$rsq
adjr2.sub_fs6 = sum_sub_fs6$adjr2
Cp.sub_fs6 = sum_sub_fs6$cp
aic.sub_fs6 = n.c*log(sse.sub_fs6/n.c)+2*p.m_sub_fs6
bic.sub_fs6 = n.c*log(sse.sub_fs6/n.c)+log(n.c)*p.m_sub_fs6
res_sub_fs6 = cbind(sum_sub_fs6$which, sse.sub_fs6,
rsq.sub_fs6,adjr2.sub_fs6,Cp.sub_fs6,aic.sub_fs6,bic.sub_fs6)

colnames(res_sub_fs6) =
c(colnames(sum_sub_fs6$which),"SSE","R2","Ra2","Cp","AIC","BIC")

colnames(res_sub_fs6)
res_sub_fs6.frame = as.data.frame(res_sub_fs6)
write.csv(res_sub_fs6.frame,"res_sub_fs6")
#Plot the SSE and rest.
par(mfrow = c(2,3))
for(i in 32:37){

plot(p.m_sub_fs6,res_sub_fs6.frame[,i],xlab="p",ylab=names(res_sub_fs6.frame[i]),pch
=19,type="o")
}

## based BIC, I found p=12 has best parameter combination.
p12.para = res_sub_fs6.frame[11,32:37] #look at the parameters of p=12
p12.para
#      SSE      R2      Ra2      Cp      AIC      BIC
# 11 53.01041 0.654245 0.6524129 70.10841 -7646.213 -7578.486
```

```
p12 = res_sub_fs6.frame[11,1:31] #get the variables 1 or 0 of p=12

p12pre = p12[which(p12[1,] == 1)] #get the variables included in p=12
abalone.trans = abalone.c
abalone.trans$SexI = NA
abalone.trans$SexI[which(abalone.trans$Sex == "I")] = 1
abalone.trans$SexI[which(abalone.trans$Sex == "F")] = 0
abalone.trans$SexI[which(abalone.trans$Sex == "M")] = 0
abalone.trans$SexI = as.factor(abalone.trans$SexI)
#write out the linear regression as p12.fun. In p12, there are Shucked:SexI,
Diameter:SexI. But I have Shucked:factor(Sex), and Diameter:factor(Sex) instead in
p12.fun
p12.fun = lm(age ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) +
I(Whole^2) + I(Diameter^2) + Viscera:Diameter + Shucked:SexI + Diameter:SexI, data =
abalone.trans)

p12.sum = summary(p12.fun)
p12.sum
p12.SSE = anova(p12.fun)["Residuals","Sum Sq"] #53.01041
p12.SSE
p12.sqr = p12.sum$r.squared #0.654245
p12.sqr
p12.adjr = p12.sum$adj.r.squared #0.6524129
p12.adjr
p12.Cp = p12.SSE/MSE.123 - (n.c-2*12) #73.70579
p12.Cp
p12.AIC = n.c * log(p12.SSE/n.c) + 2*12 #-7646.213
p12.AIC
p12.BIC = n.c * log(p12.SSE/n.c) + (log(n.c))*12 #-7578.486
p12.BIC
windows()
par(mfrow=c(1,2))
plot(p12.fun, which = 1, main = "Residual plot-p=12")
plot(p12.fun,which = 2, main = "Q-Q plot-p=12")

#All these parameters are the same as in the regsubset matrix.
############################################################
## based on the plot, I found p=20 is also good.
p20.para = res_sub_fs6.frame[19,32:37] #look at the parameters of p=20
```

```
p20.para

p20 = res_sub_fs6.frame[19,1:31] #get the variables 1 or 0 of p=20

p20pre = p20[which(p20[1,] == 1)] #get the variables included in p=20
p20pre
#write out the linear regression as p20.fun. In p20, there are Shucked:SexI,
Diameter:SexI. But I have Shucked:factor(Sex), and Diameter:factor(Sex) instead in
p12.fun
p20.fun = lm(age ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) +
I(Shell^2) + I(Diameter^2) + I(Height^2) + I(Shucked^3) + I(Diameter^3) + I(Height^3) + I
(Shell^3) + Whole:Shucked + Viscera:Diameter + Shucked:SexI + Diameter:SexI +
Shucked:Height + Shell:Shucked, data = abalone.trans)

p20.sum = summary(p20.fun)
nrow(p20.sum$coef) #20

p20.SSE = anova(p20.fun)["Residuals","Sum Sq"] #51.53361
p20.SSE
p20.sqr = p20.sum$r.squared #0.6638772
p20.sqr
p20.adjr = p20.sum$adj.r.squared #0.660789
p20.adjr
p20.Cp = p20.SSE/MSE.123 - (n.c-2*20) #30.15243
p20.Cp
p20.AIC = n.c * log(p20.SSE/n.c) + 2*20 #-7689.207
p20.AIC
p20.BIC = n.c * log(p20.SSE/n.c) + (log(n.c))*20 #-7576.328
p20.BIC

windows()
par(mfrow=c(1,2))
plot(p20.fun, which = 1, main = "Residual plot-p=20")
plot(p20.fun,which = 2, main = "Q-Q plot-p=20")
```

```
abalone = readLines('abalone.txt')
library(stringr)
test = str_split(abalone, ',', simplify = F)
test1 = data.frame(test)
abalone = data.frame(unname(t(test1)), stringsAsFactors = F)
abalone[2:9] = sapply(abalone[2:9], as.numeric)
```

```
names(abalone) =
c("Sex","Length","Diameter","Height","Whole","Shucked","Viscera","Shell","Rings")
abalone$age = abalone$Rings + 1.5
n = nrow(abalone)
drops = c("Rings") #drop Rings from data
abalone = abalone[,!(names(abalone)%in%drops)]
# perform log transformation of age
abalone$age = log(abalone$age)
# add SexI variable
abalone$SexI = NA
abalone$SexI[which(abalone$Sex == "I")] = 1
abalone$SexI[which(abalone$Sex == "F")] = 0
abalone$SexI[which(abalone$Sex == "M")] = 0
abalone$SexI = as.factor(abalone$SexI)
# split data
set.seed(10)
index=sample(1:n, size = 4176/2, replace = FALSE)
abalone.c = abalone[index,]
abalone.v = abalone[-index,]

library(MASS)
library(leaps)
# Validation for p12.fun
# model based on trainning data
p12.fun = lm(age ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) +
I(Whole^2) + I(Diameter^2) + Viscera:Diameter + Shucked:SexI + Diameter:SexI, data =
abalone.c)
sum.p12 = summary(p12.fun)
# validation p12.fun
p12.fun.v = lm(age ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) +
I(Whole^2) + I(Diameter^2) + Viscera:Diameter + Shucked:SexI + Diameter:SexI, data =
abalone.v)
sum.p12.v = summary(p12.fun.v)
# get estimates and statistics
est1 = cbind(sum.p12$coefficients[,1:2],sum.p12.v$coefficients[,1:2])
est1 = round(est1,3)
write.csv(est1,'valid1.csv')
sse1 = c(anova(p12.fun)['Residuals',2], anova(p12.fun.v)['Residuals',2])
mse1 = c(anova(p12.fun)['Residuals',3], anova(p12.fun.v)['Residuals',3])
Ra1 = c(sum.p12$adj.r.squared, sum.p12.v$adj.r.squared)
press1 = c(sum(p12.fun$residuals^2/(1-influence(p12.fun)$hat)^2),
sum(p12.fun.v$residuals^2/(1-influence(p12.fun.v)$hat)^2))
# MSPE on validation set
mspe1 = c(NA, mean((predict.lm(p12.fun, abalone.v[,-9])-abalone.v$age)^2))
# display results
stats1 = cbind(sse1, mse1, Ra1, press1, press1/(n/2), mspe1)
```

```
rownames(stats1) = c('Training','Validation')
colnames(stats1) = c('SSE','MSE','R2_a','Press','Press/n', 'MSPE')
stats1 = round(stats1,3)
write.csv(stats1,'valid2.csv')

# Validation of p20
p20.fun = lm(age ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) +
I(Shell^2) + I(Diameter^2) + I(Height^2) + I(Shucked^3) + I(Diameter^3) + I(Height^3) + I
(Shell^3) + Whole:Shucked + Viscera:Diameter + Shucked:SexI + Diameter:SexI +
Shucked:Height + Shell:Shucked, data = abalone.c)
sum.p20 = summary(p20.fun)
p20.fun.v = lm(age ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) +
I(Shell^2) + I(Diameter^2) + I(Height^2) + I(Shucked^3) + I(Diameter^3) + I(Height^3) + I
(Shell^3) + Whole:Shucked + Viscera:Diameter + Shucked:SexI + Diameter:SexI +
Shucked:Height + Shell:Shucked, data = abalone.v)
sum.p20.v = summary(p20.fun.v)
# get estimates and statistics
est2 = cbind(sum.p20$coefficients[,1:2],sum.p20.v$coefficients[,1:2])
est2 = round(est2,3)
write.csv(est2,'valid3.csv')
sse2 = c(anova(p20.fun)['Residuals',2], anova(p20.fun.v)['Residuals',2])
mse2 = c(anova(p20.fun)['Residuals',3], anova(p20.fun.v)['Residuals',3])
Ra2 = c(sum.p20$adj.r.squared, sum.p20.v$adj.r.squared)
press2 = c(sum(p20.fun$residuals^2/(1-influence(p20.fun)$hat)^2),
sum(p20.fun.v$residuals^2/(1-influence(p20.fun.v)$hat)^2))
# MSPE on validation set
mspe2 = c(NA, mean((predict.lm(p20.fun, abalone.v[,-9])-abalone.v$age)^2))
# display results
stats2 = cbind(sse2, mse2, Ra2, press2, press2/(n/2), mspe2)
rownames(stats2) = c('Training','Validation')
colnames(stats2) = c('SSE','MSE','R2_a','Press','Press/n', 'MSPE')
stats2 = round(stats2,3)
write.csv(stats2,'valid4.csv')
########################################
####   Final model: p12.fun
# use entire data set to re-fit the final model
Model = lm(age ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) +
I(Whole^2) + I(Diameter^2) + Viscera:Diameter + Shucked:SexI + Diameter:SexI, data =
abalone)
sum.model = summary(Model)
round(Model$coefficients,3)
# diagnosis of assumption
par(mfrow = c(1,2))
plot(Model,1)
plot(Model,2)
#### Outlying Y cases
```

```
res = Model$residuals # residuals
hii = influence(Model)$hat # diagonal of the hat matrix: leverage values
stu.r = res/sqrt(anova(Model)['Residuals',3]*(1-hii)) # studentized residuals
del = res/(1-hii) # deleted residuals

stud.d = studres(Model) # studentized deleted residuals
head(sort(abs(stud.d), decreasing = T),10) # 10 largest studendized deleted residuals
qt(1-0.05/(2*n),n-12-1) # Bonferroni's thresholed at level 0.05
index.Y = which(abs(stud.d)>qt(1-0.05/(2*n),n-12-1))
index.Y # case 481 and 2184 identified as outliers
# display the results of 3 kinds of residuals
plot(Model$fitted.values, res, xlab = 'fitted values', ylab = 'residuals', pch = 17, cex = .5,
ylim = c(-5.5,5), main = 'Residuals vs Fitted Values')
points(Model$fitted.values, stu.r, col = 'red', pch = 19, cex = 0.5)
points(Model$fitted.values, stud.d, col = 'blue', pch = 18, cex = .4)
abline(h=0, col = grey(0.6), lwd = 2, lty = 2)
abline(h=qt(1-0.05/(2*n),n-12-1), lwd = 2, lty = 3)
abline(h = -qt(1-0.05/(2*n),n-12-1), lwd = 2, lty = 3)
legend('bottomleft', legend = c('residuals','studentized residuals','studentized deleted
residuals'), pch = c(17,19,18), col = c('black','red','blue'), bty = 'n')
#### Outlying X Cases
# use leverage values
head(sort(hii, decreasing = T)) # the leverage values which are diagonal elements of hat
matrix
index.X = which(hii>2*12/n)
index.X = unname(index.X)
length(index.X) # 333 cases detected
# Identify influential cases
# Cook's distance
cook = res^2*hii/(12*anova(Model)['Residuals',3]*(1-hii)^2)
outl = which(cook[index.X] > 4/(n-12))
outl = as.numeric(names(outl)) # outliers may have substantial influence on fitted values
length(outl)
# plot of cook's distance and leverage values
par(mfrow = c(1,2))
plot(Model,4)
abline(h = 0.0057, lty = 2)
plot(Model,5)
#############################################
# Remove outliers and refit the model
abalone.n = abalone[-outl,]
final.model = lm(age ~ Shell + Whole + Shucked + Viscera + Diameter + I(Shucked^2) +
I(Whole^2) + I(Diameter^2) + Viscera:Diameter + Shucked:SexI + Diameter:SexI, data =
abalone.n)
sum.fm = summary(final.model)
round(final.model$coefficients,3)
```

```
plot(Model$residuals[-outl], final.model$residuals, pch = 20, cex = .5, main =
'Comparision of Residuals after Dropping Outliers', xlab = 'Residuals in Model', ylab =
'Residuals in Final Model')
abline(a=0,b=1, col = 'red')
# compare SSE etc.
ra = c(sum.model$adj.r.squared, sum.fm$adj.r.squared)
ss1 = anova(final.model)['Residuals',2:3]
ss2 = anova(Model)['Residuals',2:3]
comp = rbind(ss2,ss1)
comp = cbind(comp, ra)
colnames(comp) = c('SSE','MSE','R^2_a')
rownames(comp) = c('Model','Final model')
comp = round(comp,3)
write.csv(comp, 'valid5.csv')
```