A MULTIPLE LINEAR REGRESSION MODEL ANALYSIS ON ABALONE AGE PREDICTION

Siyao Wang, Jieyun Wang, Zhi Zhang
{syywang, jynwang, wwzzhang}@ucdavis.edu
id{917856208,917781483,917834518}

ABSTRACT

This project focused on analyzing the relationship between abalone age and physical measurements. By employing least square principle and stepwise procedures several reasonable models were built. Through internal validation and external validation, the best model was chosen. After detecting and deleteing influential outliers from whole dataset, the final model was fitted on the remaining dataset which will be used to accurately predict the age of abalone. All technical details for building the model are presented. The whole analysis is rigorous and integral. In addition, other novel model selection methods, including Principal Component Regression(PCR) and shrinkage methods, were implemented and their results were compared.

1 Introduction

The economic value of abalone is positively correlated with its age. Farmers usually cut the shells and count the rings through microscopes to determine age of abalones, which is demanding and time consuming. The complex method also increases the price of abalone and limits its popularity. Therefore, a better approach based on the current data driven methods should be developed to predicate the abalone age effectively.

Multiple linear regression (MLR) is commonly used to model the relationship between a scalar response and more than one explanatory variable (Freedman, 2009). Our problem might be properly solved by a multiple linear regression model. We are interested in building a MLR model that gave reliable regression estimates and generalized well. The first challenge is to build a model that can meet all of the above requirements at all. The second challenge stems from the robustness of the model building method, the effect of different predictors with the response variable, and how different linear regression methods behaved, how well the model fit the abalone's data set.

Motivated by these challenges, our paper focuses on the multiple linear regression model building with the following specific contributions:

Comprehensive and robust analysis for model building Our paper contained a series of processes including data processing, data exploratory, preliminary model investigation, model selection, model diagnostic and validation. Our full stack analysis tackled the model building from every angle to create a robust model.

Answered analytical questions related to model building Through out our analysis, following questions were answered: the associations between different predictor variables and the response variable, the robustness of models' performance on the training data and the predicting ability on the validation data.

Compare different linear regression methods for model building. Besides subset selection, we also compared our method with the shrinkage method and methods using derived input directions. The subset selection is a discrete process which often exhibits high variance(Friedman *et al.*, 2001). Shrinkage methods is more continuous, and don't suffer as much from high variability. PCR is commonly used in the situations that predictor variables are highly correlated. We compared the predictive abilities of different models built by these methods.

2 Methods and Results

2.1 Data collection and processing

The data set is provided with the UCI machinery repository. The original data is sourced from the (Nash *et al.*, 1994), which recorded about the abalone population age and physical measurements. Over 30 studies have referenced these data. The data has no missing value and the categorical variable is coded as factor, so no correction is needed and the variable sex should be treated as a categorical variable with three levels M, F, and I. We updated the rings by adding 1.5 to be as the age variable.

2.2 Exploratory data analysis

The scatter plot matrix Figure 1 were drawn. The sex variable looks like dots on lines. It is because sex only takes several values. There are no weird patterns for other variables. There appears to be a light non-linear relationship between the response variable and other predictors.

Since obviation from normal will deteriorate model assumptions, following with the histogram of the response variable, we checked if any transformation was needed. The Figure 2 box-cox procedure returns a $\lambda = -0.3$, so logarithm transform would be more preferable. Besides, the distribution of transformed response variable under different transformations were drawn, as shown in Figure 3. It confirms that the logarithm transformation is highly approximate to the normal distribution, so we chose it. For clear notation, denote Y as the logarithm transformation of age.

Scatter plot matrix among the quantitative variables in the interest of new transformed response Figure 4 shows that there appears to be a light non-linear, in fact quadratic relationship in between Y and other predictors.

The pie chart and side-by-side box plots Figure 5 for the response variable with respect to sex indicates that female and male abalone tend to have higher age than infant.

Since there is enough data, we randomly split our data as 80% training and 20% validation.

To examine whether the training data and validation data look alike, side-by-side box plots for each quantitative variable with respect to training data and validation data were drawn respectively. Figure 6 shows that all quantitative variables have similar distributions between training data and validation data.

The correlation matrix and VIF_k, show X variables are highly correlated.

2.3 Model Building and Diagnostic

2.3.1 First-order Model

Table 1 lists the response variable and predictor variables. By regressing the response variable independently on each predictor variable and calculating the correlation matrix, there is a significant association between the response variable and each X variable, so a first-order model with all predictor variables was chosen.

First-Order Model with all predictors

Figure 7 shows that the first-order model with all predictor variables is $\hat{Y} = 1.627913062 - 0.080374265X_1 + 0.006340914X_2 + 0.350780731X_3 + 1.018792870X_4 + 2.127506842X_5 + 0.573877043X_6 - 1.509800019X_7 - 0.767203991X_8 + 0.436939034X_9$.

Best Subset Selection Since there were only night X variables, exhaustive method and different criteria such as R_a^2, C_p, AIC_p, BIC_p could be used to select the best model. Because X_1 and X_2 are corresponding to a qualitative variable as a group, if the best model selected according to one criterion just included one, we still included both of them. Then the followings are the best models:
Under SSE, R_2, R_a^2, C_p, AIC_p criterion, first-order full model is chosen.
Under BIC_p criterion: model with $X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$ variables is chosen, X_3 is dropped.

Stepwise Regression

Four commonly used stepwise procedures: forward stepwise procedure, forward selection procedure, backward stepwise procedure and backward elimination procedure were also employed. For each procedure, both AIC

and BIC are used as pre-specified criteria. According to forward stepwise procedure with AIC, forward selection procedure with AIC, backward stepwise procedure with AIC and backward elimination procedure with AIC, first-order full model was chosen. According to forward stepwise procedure with BIC, forward selection procedure with BIC, backward stepwise procedure with BIC and backward procedure elimination with BIC, the model with $X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$ variables was chosen. Then, **Model 1** was defined as the model regressing Y on $X_1, X_2, X_4, X_5, X_6, X_7, X_8, X_9$ and **Model 2** was defined as the first-order full model.

After building these models, the residual plots of these two models were plotted.Figure 8 shows that these two models seem to be reasonable. However, the residual vs. fitted plot shows non linearity and the Q-Q plot indicates heavy-tailed errors. Thus, second-order models were considered.

2.3.2 Second-order Model

In second-order model, to reduce the correlation between the linear terms and the quadratic terms, the predictor variables were centered.

Second-Order Model with all predictors

In this model, we started with adding all possible first-order and second-order terms. The fitted model is presented in Figure 9. Since there are in total 51 predictor variables in the model, the number of possible models would roughly be 2^{51}. Evaluating each and every one of those models is quite time consuming and might be even infeasible computationally. Thus, only stepwise regression procedures were emploied.

Stepwise Regression

As the procedure in first-order model, four commonly used stepwise procedures were used with both AIC and BIC as pre-specified criteria. Theoretically, these procedures will lead to 8 models. After carefully comparison, we noticed that backward elimination(BIC) and backward stepwise(BIC) ended up with a same model, backward elimination(AIC) and backward stepwise(AIC) ended up with a same model. Thus, after stepwise regression procedure, additional 6 models were selected. A detailed summary of these models could be found in Appendix B Rcodes and Outputs:
Model 3 with 18 predictor variables, chosen by forward stepwise with BIC.
Model 4 with 18 predictor variables, chosen by forward selection with BIC.
Model 5 with 21 predictor variables, chosen by backward stepwise and backward selection with BIC.
Model 6 with 23 predictor variables, chosen by forward stepwise with AIC.
Model 7 with 26 predictor variables, chosen by forward selection with AIC.
Model 8 with 25 predictor variables, chosen by backward stepwise and backward selection with AIC.

2.4 Model Validation

In this part both internal validation and external validation were adopted to choose the best model between the models selected by different approaches. Through checking whether $C_p \approx p$, $Press_p$ is not much larger than SSE_p, the parameter estimation is consistent, and $MSPE_v$ based on the validation data is not much larger than $\frac{SSE_p}{n}$ and $\frac{Press_p}{n}$ based on the training data, we found only Model 5 has the property that same sign between the two sets of estimated coefficients is same. Furthermore, Model 5 also satisfies other criteria. Thus, Model 5 was chose as the final model. Denote $MSPE_{LS}$ to be the mean squared prediction error of Model 5, $MSPE_{LS} = 0.02541876$.

2.5 Model Finalization

2.5.1 Detecting Outliers

After fitting the final model with all data, studentized deleted residuals and Bonferroni's procedure at $\alpha = 0.1$ were implemented to identify the outlying Y observations. The Bonferroni's threshold is $t(1, 1 - \frac{\alpha}{2n}, n - p - 1) = 4.383448$. The Y observations corresponding to studentized deleted residuals which are greater than the Bonferroni's threshold could be deemed as significant outlying observations. They are cases 481 and 2184.
Then, the leverage values were compared with the value of $\frac{2p}{n} = 0.01053388$. The cases with $h_{ii} > \frac{2p}{n}$ were defined as outlying X observations. There are 291 cases.

2.5.2 Detecting Influential points

Finally, influential cases were identified via Cook's distance. There are 7 cases: 892, 1210, 1217, 1418, 2052, 2628 and 3997. Then, these influential cases were dropped, and the final model was refitted.

2.5.3 Final Model

After refitting, the following model is given:

$$\begin{aligned}\hat{Y} =& 2.459 - 0.013 SexI - 0.004 SexM - 0.519 Length^* + 0.585 Diameter^* + 0.991 Height^* \\ &+ 0.948 Whole\_weight^* - 2.124 Shucked\_weight^* - 0.584 Viscera\_weight^* \\ &+ 0.933 Shell\_weight^* - 5.66 Length^{*2} - 2.988 Diameter^{*2} + 5.232 Height^{*2} \\ &+ 2.924 Shucked\_weight^* + 0.457 SexI * Shucked\_weight^* + 0.052 SexM * Shucked\_weight^* \\ &+ 4.873 Length^* * Shucked\_weight^* + 0.74 Diameter^* * Whole\_weight^* \\ &- 7.969 Height^* * Viscera\_weight^* - 2.151 Whole\_weight^* * Shucked\_weight^* \\ &+ 1.393 Whole\_weight^* * Viscera\_weight^* - 2.74 Viscera\_weight^* * Shell\_weight^*\end{aligned}$$

where all starred variables are centered ones.
The residual plots are given in Figure 12 shows no obvious non-linearity, Q-Q plot indicates heavy-tailed distribution but only at a reasonable amount, and residual vs leverage plot gives no highly influential cases. All combined, this model could be considered as a sufficient model.

2.6 Additional Methods

We also implemented other state-of-art methods for building models, and compared them with the best model selected above.

2.6.1 Principle Component Regression

In PCR, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors.

The singular value decomposition (SVD) of the centered input $N \times p$ matrix $\mathbf{x}$ has the form $\mathbf{x} = \mathbf{udv}^T$, and the the sample covariance matrix is given by $\mathbf{s} = \mathbf{x}^T\mathbf{x}/N$. Therefore, we have $\mathbf{x}^T\mathbf{x} = \mathbf{vd}^2\mathbf{v}^T$. The first principle component $z_1 = \mathbf{x}v_1$ has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{x}$. The subsequent principal components z_j is subject to being orthogonal to the earlier ones. Since the $z_m := \{z_1, z_j\}$ are orthogonal, this regression is just a sum of univariate regressions, which are $\hat{y}_M^{pcr} = \bar{y}\mathbf{i} + \Sigma_{m=1}^M \theta_m \hat{}\mathbf{z}_m$, where $\hat{\theta}{}_m = \frac{<\mathbf{z}_m, \mathbf{y}>}{<\mathbf{z}_m, \mathbf{z}_m>}$, then the coefficients is expressed as:

$$\hat{\beta}^{pcr}(M) := \Sigma_{m=1}^M \hat{\theta}_m v_m \tag{1}$$

For $M < p$, the PCR discards the $p - M$ smallest eigenvalue components. Our R output shows that the $MSPE_{PCR} = 0.03059$, which is slightly larger than $MSPE_{LS}$

2.6.2 Shrinkage Methods

By retaining a subset of the predictors and discarding the rest, stepwise procedure produces models that have possibly lower prediction error than the full model. However, because it is a discrete process—variables are either retained or discarded—it often exhibits high variance. Shrinkage methods are more continuous, and don't suffer as much from high variability, so we investigated shrinkage methods and compared it with stepwise procedure.

Ridge Regression

The ridge estimators are:

$$\hat{\beta}_{ridge} := argmin_{\beta}\{\Sigma_{i=1}^n (y_i - \beta_0 - \Sigma_{j=1}^p x_{ij}\beta_j)^2 + \lambda\Sigma_{j=1}^p \beta_j^2\} \tag{2}$$

Since the predictor variables are highly correlated, regression coefficients can become poorly determined and exhibit high variance. We tried to use ridge regression to shrinkage the regression coefficients of the full first-order model. Ridge regression shrinks the regression coefficients by imposing a penalty on their size, and the parameter λ controls the amount of shrinkage.

We implemented the ridge regression cv.glmnet function in R to choose the best λ by minimization of expected generalization error in cross validation. Then we us the lambda.min to build a first-order model. The model is $\hat{Y} = 1.737950670 - 0.078639485X_1 + 0.002594522X_2 + 0.436885169X_3 + 0.778867901X_4 + 1.084187380X_5 + 0.036742765X_6 - 0.596676032X_7 - 0.171006721X_8 + 0.851200383X_9$. $MSPE_{ridge} = 0.031041$ is slightly larger than $MSPE_{LS}$.

The ridge regression was only applied on the first-order model since ridge is a proportional shrinkage which doesn't eliminate coefficients. Under a higher order model, if the ridge regression is adopted, no variable will be eliminated due to this property of proportional shrinkage, and a model which has too many predictors will be produced. The model will be hard to interpret.

Lasso

$$\hat{\beta}_{lasso} := argmin_{\beta}\{\frac{1}{2}\Sigma_{i=1}^{n}(y_i - \beta_0 - \Sigma_{j=1}^{p}x_{ij}\beta_j)^2 + \lambda\Sigma_{j=1}^{p}|\beta_j|\} \tag{3}$$

Lasso is another shrinkage method using regularization term but using the first order penalty term which reflected in Equation (3). This change will actually cause some of the coefficients to be exactly zero. We built the Lasso on the full second-order model, in order to choose significant variables.

The model built by Lasso is presented in Figure 10. The Lasso's $MSPE_{lasso} = 0.02599769$, which is slightly higher than the $MSPE_{LS}$, but it's lower than the $MSPE_{ridge}$. This is because the residual sum of squares in Lasso has elliptical contours, centered at the full least squares estimate. The constraint part region for ridge regression is the disk, while that for lasso is the diamond, so that if the solution occurs at a corner, then the parameter β will equal to zero.

Elastic net regularization

If there is a group of highly correlated variables, then Lasso tends to select one variable from a group and ignore the others, and thus the model built by Lasso only have few variables. To overcome these limitations, elastic net could be implemented which adds a quadratic part to the penalty. $(\|\beta\|^2\|\beta\|^2)$.

$$\hat{\beta} := argmin_{\beta}\{\Sigma_{i=1}^{n}(y_i - \beta_0 - \Sigma_{j=1}^{p}x_{ij}\beta_j)^2 + \lambda_1\Sigma_{j=1}^{p}\beta_j^2 + \lambda_2\Sigma_{j=1}^{p}|\beta_j|\} \tag{4}$$

The elastic net is a regularized regression method that linearly combines the L_1 and L_2 penalties of the Lasso and ridge methods. The model built by elastic net is presented in Figure 11. $MSPE_{elasticnet} = 0.02590794$ is similar with $MSPE_{LS}$.

3 Conclusions and Discussion

Using different model selection approaches, different models were built. We compared the $MSPE$ of different models and found there was no big difference among them. All of these models have relatively good predict abilities.

In a nutshell, when there are only few X variables, exhaustive methods or use ridge regression could be employ to build models. When there is a large number of predictor variables stepwise procedures, PCR, Lasso or elastic net might preferred to build models. In this project, each one of these methods provides us with a good model to predict the age of abalone.

Appendix A Figures and tables

Table 1: variables in the model

name	variable in model	type	description
Sex (M)	X_1	factor	Male
Sex (I)	X_2	factor	Infant
Length	X_3	numeric (mm)	Longest shell measurement
Diameter	X_4	numeric (mm)	perpendicular to length
Height	X_5	numeric (mm)	with meat in shell
Whole weight	X_6	numeric (grams)	whole abalone
Shucked weight	X_7	numeric (grams)	weight of meat
Viscera weight	X_8	numeric (grams)	gut weight (after bleeding)
Shell weight	X_9	numeric (grams)	after being dried
log(Age)	Y	numeric	the age in years log transformed

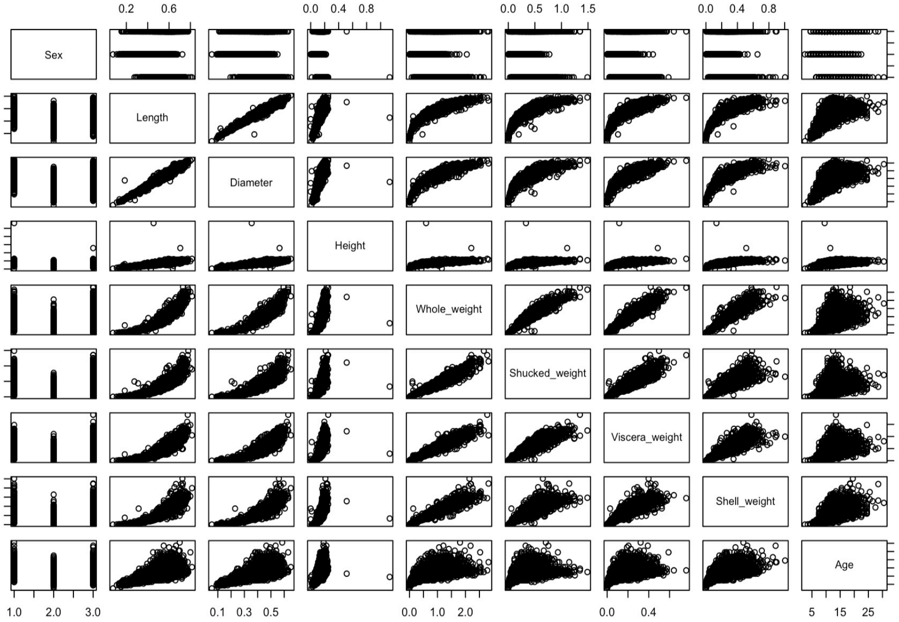


Figure 1: Scatter matrix before transformation

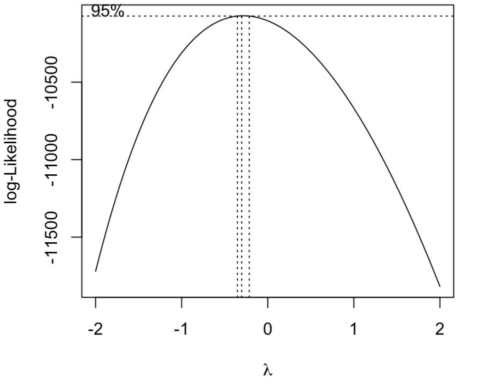


Figure 2: Box Cox procedure

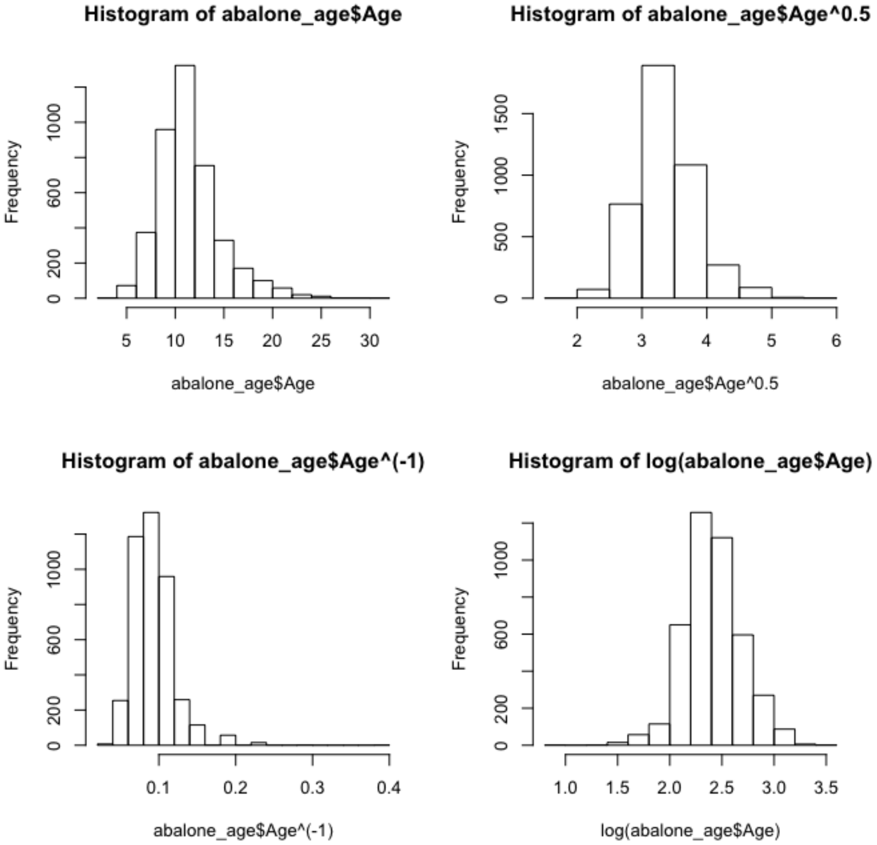


Figure 3: Y distribution

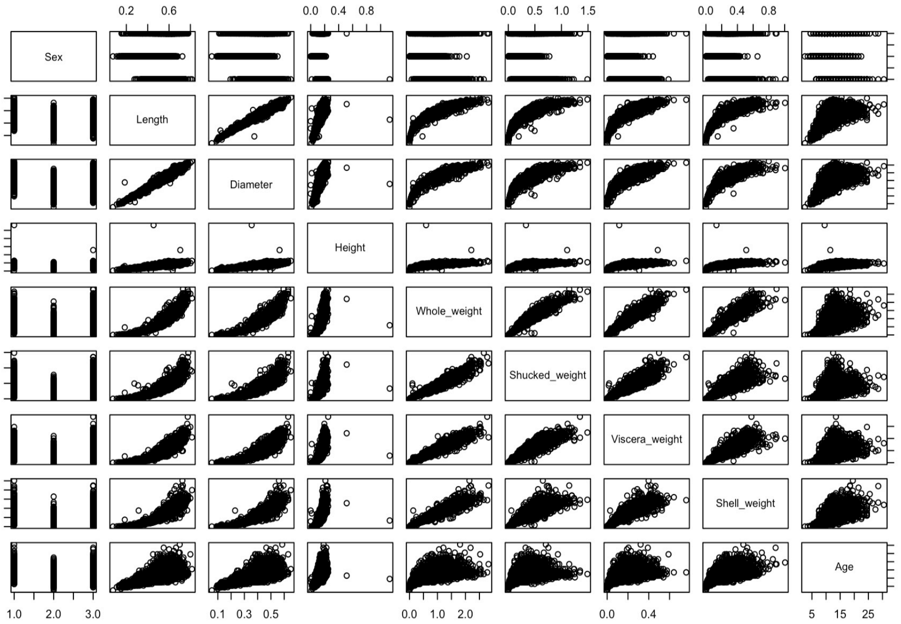


Figure 4: Scatter matrix after transformation

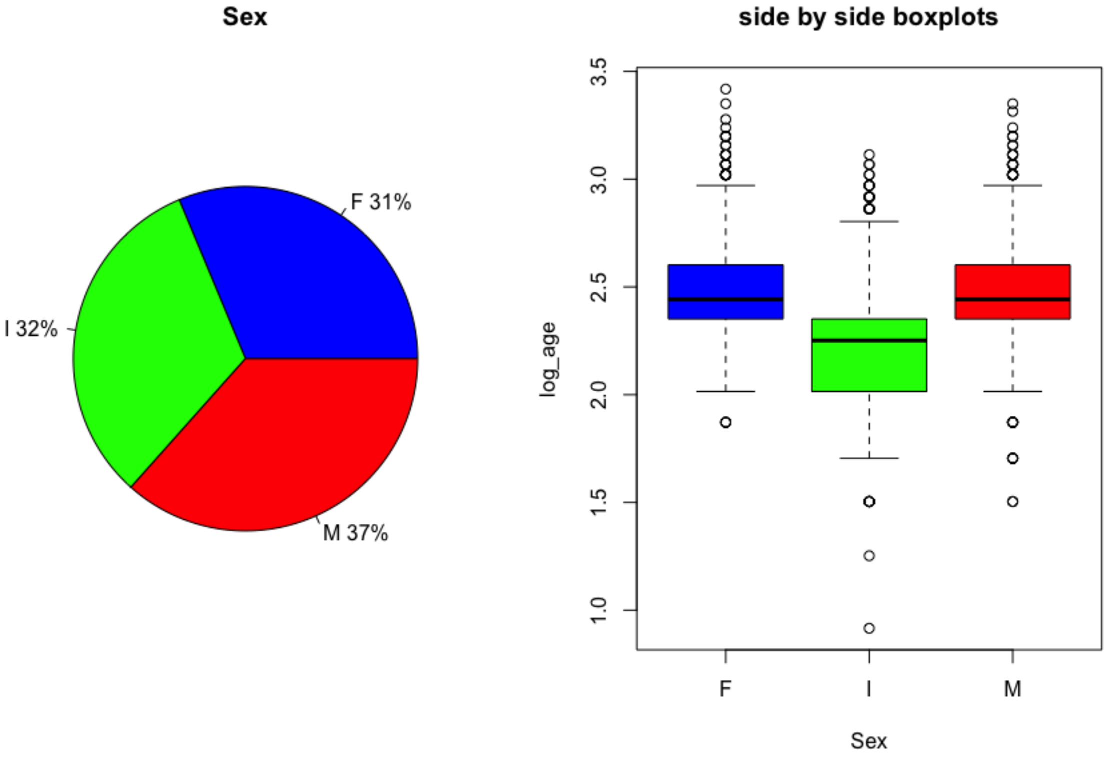


Figure 5: Pie and side-by-side box plot for Sex

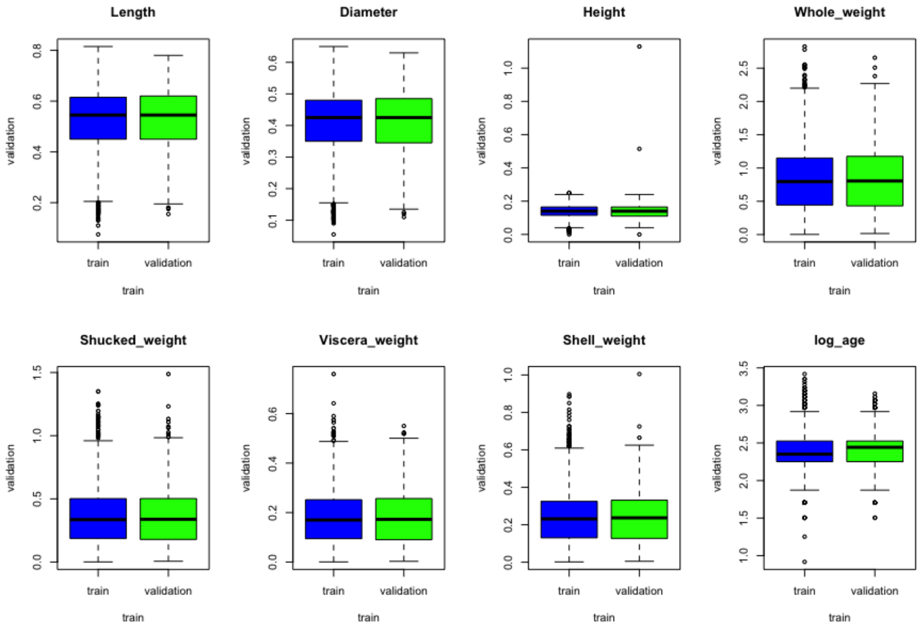
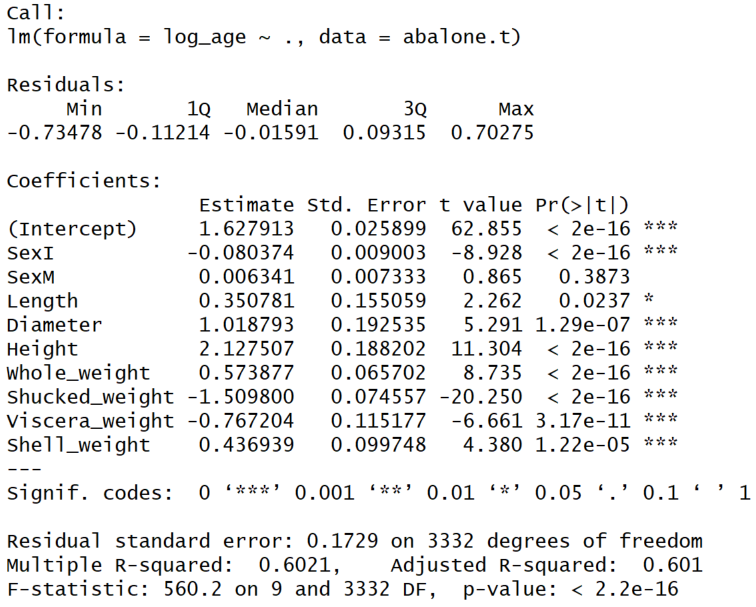


Figure 6: Side by Side train validation box plot



```
Call:
lm(formula = log_age ~ ., data = abalone.t)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73478 -0.11214 -0.01591  0.09315  0.70275 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.627913   0.025899  62.855  < 2e-16 ***
SexI           -0.080374   0.009003  -8.928  < 2e-16 ***
SexM            0.006341   0.007333   0.865   0.3873    
Length          0.350781   0.155059   2.262   0.0237 *  
Diameter        1.018793   0.192535   5.291 1.29e-07 ***
Height          2.127507   0.188202  11.304  < 2e-16 ***
Whole_weight    0.573877   0.065702   8.735  < 2e-16 ***
Shucked_weight -1.509800   0.074557 -20.250  < 2e-16 ***
Viscera_weight -0.767204   0.115177  -6.661 3.17e-11 ***
Shell_weight    0.436939   0.099748   4.380 1.22e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1729 on 3332 degrees of freedom
Multiple R-squared:  0.6021,    Adjusted R-squared:  0.601 
F-statistic: 560.2 on 9 and 3332 DF,  p-value: < 2.2e-16
```

Figure 7: summary fit1

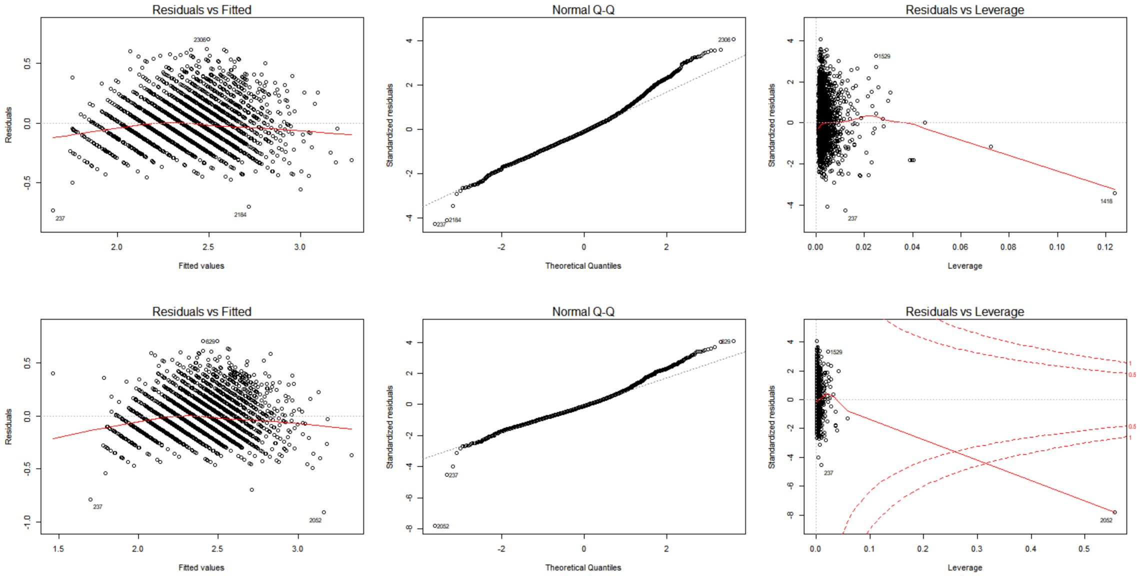
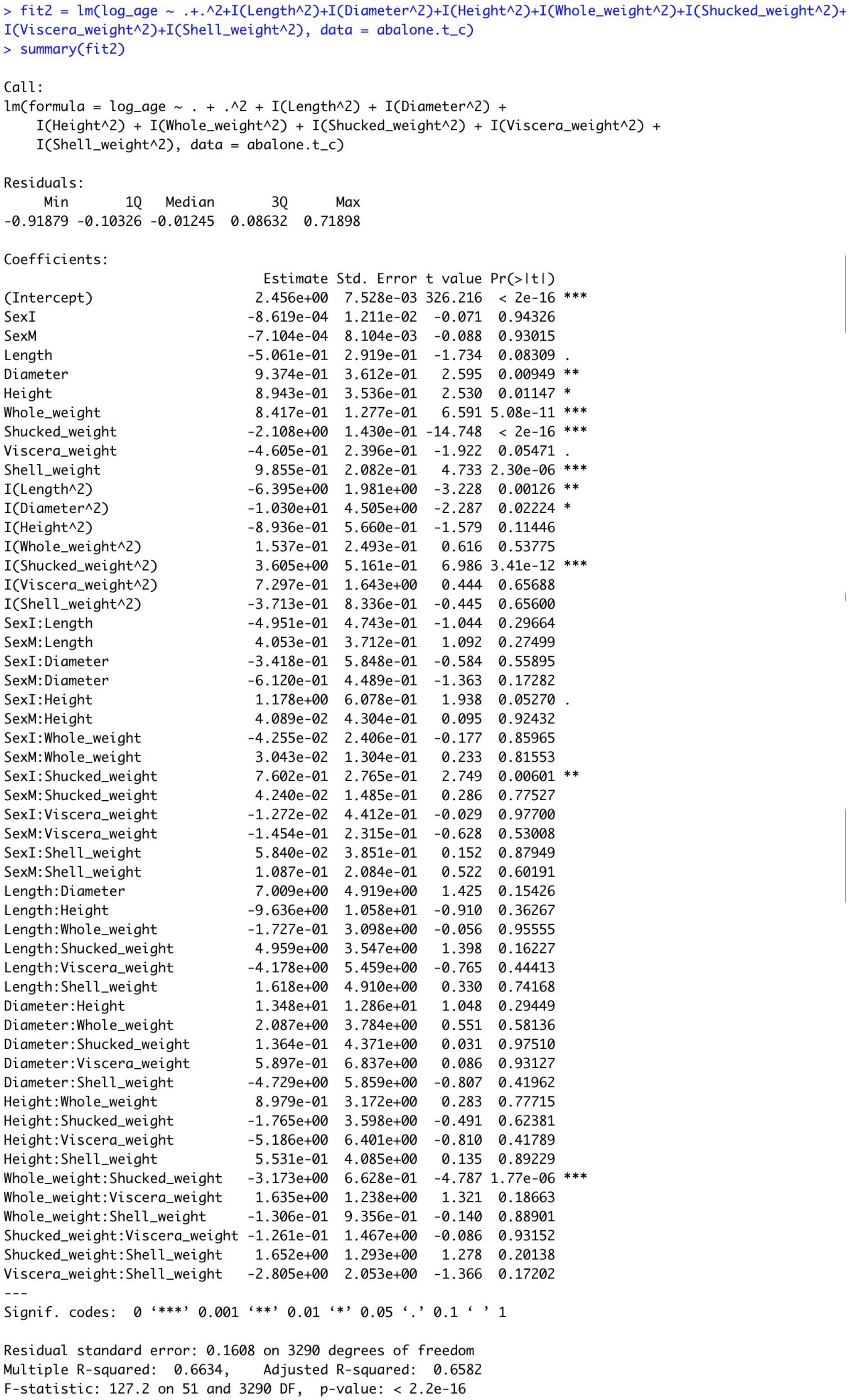
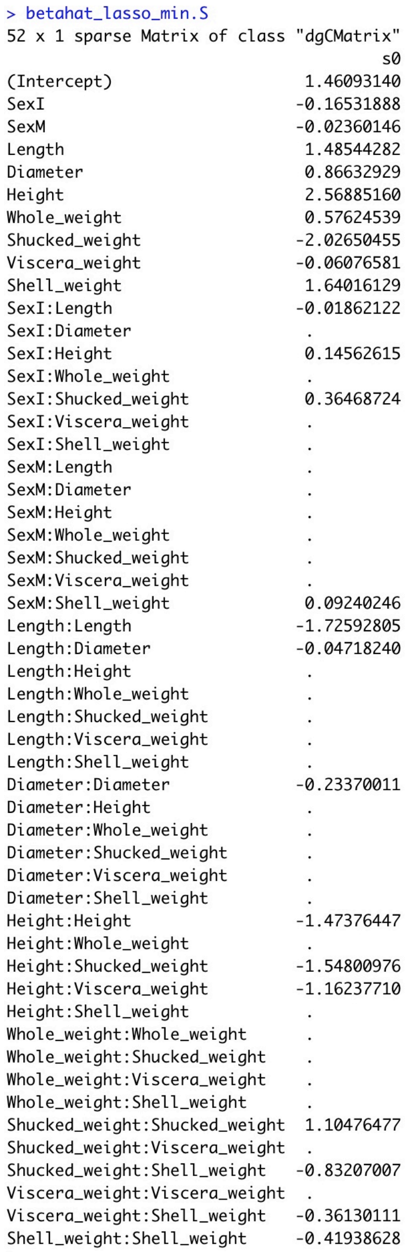


Figure 8: residual plots of first-order models: Model 1(bottom) and Model 2



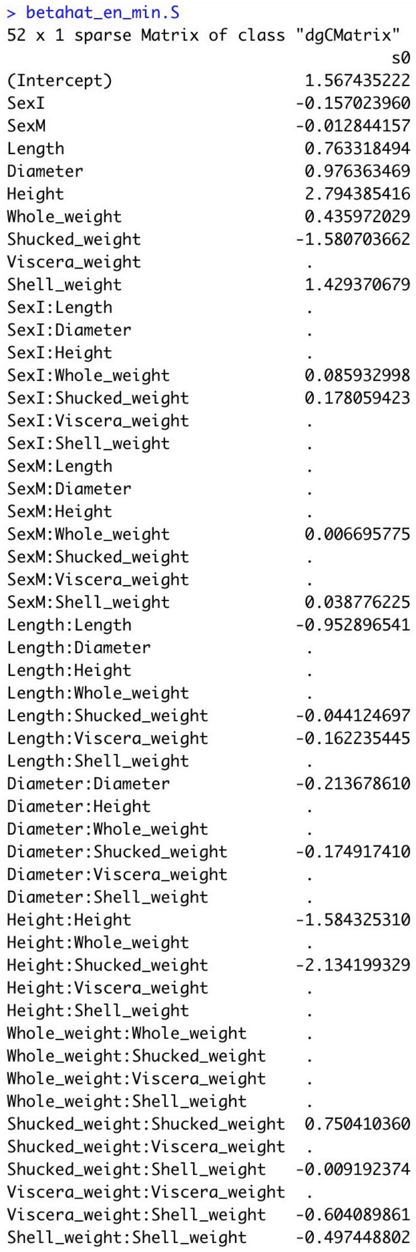
```
> fit2 = lm(log_age ~ .+.^2+I(Length^2)+I(Diameter^2)+I(Height^2)+I(Whole_weight^2)+I(Shucked_weight^2)+
I(Viscera_weight^2)+I(Shell_weight^2), data = abalone.t_c)
> summary(fit2)

Call:
lm(formula = log_age ~ . + .^2 + I(Length^2) + I(Diameter^2) + 
    I(Height^2) + I(Whole_weight^2) + I(Shucked_weight^2) + I(Viscera_weight^2) + 
    I(Shell_weight^2), data = abalone.t_c)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.91879 -0.10326 -0.01245  0.08632  0.71898 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    2.456e+00  7.528e-03 326.216  < 2e-16 ***
SexI                          -8.619e-04  1.211e-02  -0.071  0.94326    
SexM                          -7.104e-04  8.104e-03  -0.088  0.93015    
Length                        -5.061e-01  2.919e-01  -1.734  0.08309 .  
Diameter                       9.374e-01  3.612e-01   2.595  0.00949 ** 
Height                         8.943e-01  3.536e-01   2.530  0.01147 *  
Whole_weight                   8.417e-01  1.277e-01   6.591 5.08e-11 ***
Shucked_weight                -2.108e+00  1.430e-01 -14.748  < 2e-16 ***
Viscera_weight                -4.605e-01  2.396e-01  -1.922  0.05471 .  
Shell_weight                   9.855e-01  2.082e-01   4.733 2.30e-06 ***
I(Length^2)                   -6.395e+00  1.981e+00  -3.228  0.00126 ** 
I(Diameter^2)                 -1.030e+01  4.505e+00  -2.287  0.02224 *  
I(Height^2)                   -8.936e-01  5.660e-01  -1.579  0.11446    
I(Whole_weight^2)              1.537e-01  2.493e-01   0.616  0.53775    
I(Shucked_weight^2)            3.605e+00  5.161e-01   6.986 3.41e-12 ***
I(Viscera_weight^2)            7.297e-01  1.643e+00   0.444  0.65688    
I(Shell_weight^2)             -3.713e-01  8.336e-01  -0.445  0.65600    
SexI:Length                   -4.951e-01  4.743e-01  -1.044  0.29664    
SexM:Length                    4.053e-01  3.712e-01   1.092  0.27499    
SexI:Diameter                 -3.418e-01  5.848e-01  -0.584  0.55895    
SexM:Diameter                 -6.120e-01  4.489e-01  -1.363  0.17282    
SexI:Height                    1.178e+00  6.078e-01   1.938  0.05270 .  
SexM:Height                    4.089e-02  4.304e-01   0.095  0.92432    
SexI:Whole_weight             -4.255e-02  2.406e-01  -0.177  0.85965    
SexM:Whole_weight              3.043e-02  1.304e-01   0.233  0.81553    
SexI:Shucked_weight            7.602e-01  2.765e-01   2.749  0.00601 ** 
SexM:Shucked_weight            4.240e-02  1.485e-01   0.286  0.77527    
SexI:Viscera_weight           -1.272e-02  4.412e-01  -0.029  0.97700    
SexM:Viscera_weight           -1.454e-01  2.315e-01  -0.628  0.53008    
SexI:Shell_weight              5.840e-02  3.851e-01   0.152  0.87949    
SexM:Shell_weight              1.087e-01  2.084e-01   0.522  0.60191    
Length:Diameter                7.009e+00  4.919e+00   1.425  0.15426    
Length:Height                 -9.636e+00  1.058e+01  -0.910  0.36267    
Length:Whole_weight           -1.727e-01  3.098e+00  -0.056  0.95555    
Length:Shucked_weight          4.959e+00  3.547e+00   1.398  0.16227    
Length:Viscera_weight         -4.178e+00  5.459e+00  -0.765  0.44413    
Length:Shell_weight            1.618e+00  4.910e+00   0.330  0.74168    
Diameter:Height                1.348e+01  1.286e+01   1.048  0.29449    
Diameter:Whole_weight          2.087e+00  3.784e+00   0.551  0.58136    
Diameter:Shucked_weight        1.364e-01  4.371e+00   0.031  0.97510    
Diameter:Viscera_weight        5.897e-01  6.837e+00   0.086  0.93127    
Diameter:Shell_weight         -4.729e+00  5.859e+00  -0.807  0.41962    
Height:Whole_weight            8.979e-01  3.172e+00   0.283  0.77715    
Height:Shucked_weight         -1.765e+00  3.598e+00  -0.491  0.62381    
Height:Viscera_weight         -5.186e+00  6.401e+00  -0.810  0.41789    
Height:Shell_weight            5.531e-01  4.085e+00   0.135  0.89229    
Whole_weight:Shucked_weight   -3.173e+00  6.628e-01  -4.787 1.77e-06 ***
Whole_weight:Viscera_weight    1.635e+00  1.238e+00   1.321  0.18663    
Whole_weight:Shell_weight     -1.306e-01  9.356e-01  -0.140  0.88901    
Shucked_weight:Viscera_weight -1.261e-01  1.467e+00  -0.086  0.93152    
Shucked_weight:Shell_weight    1.652e+00  1.293e+00   1.278  0.20138    
Viscera_weight:Shell_weight   -2.805e+00  2.053e+00  -1.366  0.17202    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1608 on 3290 degrees of freedom
Multiple R-squared:  0.6634,	Adjusted R-squared:  0.6582 
F-statistic: 127.2 on 51 and 3290 DF,  p-value: < 2.2e-16
```

Figure 9: summary fit2



```
> betahat_lasso_min.S
52 x 1 sparse Matrix of class "dgCMatrix"
                                       s0
(Intercept)                    1.46093140
SexI                          -0.16531888
SexM                          -0.02360146
Length                         1.48544282
Diameter                       0.86632929
Height                         2.56885160
Whole_weight                   0.57624539
Shucked_weight                -2.02650455
Viscera_weight                -0.06076581
Shell_weight                   1.64016129
SexI:Length                   -0.01862122
SexI:Diameter                  .
SexI:Height                    0.14562615
SexI:Whole_weight              .
SexI:Shucked_weight            0.36468724
SexI:Viscera_weight            .
SexI:Shell_weight              .
SexM:Length                    .
SexM:Diameter                  .
SexM:Height                    .
SexM:Whole_weight              .
SexM:Shucked_weight            .
SexM:Viscera_weight            .
SexM:Shell_weight              0.09240246
Length:Length                 -1.72592805
Length:Diameter               -0.04718240
Length:Height                  .
Length:Whole_weight            .
Length:Shucked_weight          .
Length:Viscera_weight          .
Length:Shell_weight            .
Diameter:Diameter             -0.23370011
Diameter:Height                .
Diameter:Whole_weight          .
Diameter:Shucked_weight        .
Diameter:Viscera_weight        .
Diameter:Shell_weight          .
Height:Height                 -1.47376447
Height:Whole_weight            .
Height:Shucked_weight         -1.54800976
Height:Viscera_weight         -1.16237710
Height:Shell_weight            .
Whole_weight:Whole_weight      .
Whole_weight:Shucked_weight    .
Whole_weight:Viscera_weight    .
Whole_weight:Shell_weight      .
Shucked_weight:Shucked_weight  1.10476477
Shucked_weight:Viscera_weight  .
Shucked_weight:Shell_weight   -0.83207007
Viscera_weight:Viscera_weight  .
Viscera_weight:Shell_weight   -0.36130111
Shell_weight:Shell_weight     -0.41938628
```

Figure 10: model build by Lasso



```
> betahat_en_min.S
52 x 1 sparse Matrix of class "dgCMatrix"
                                        s0
(Intercept)                    1.567435222
SexI                          -0.157023960
SexM                          -0.012844157
Length                         0.763318494
Diameter                       0.976363469
Height                         2.794385416
Whole_weight                   0.435972029
Shucked_weight                -1.580703662
Viscera_weight                 .          
Shell_weight                   1.429370679
SexI:Length                    .          
SexI:Diameter                  .          
SexI:Height                    .          
SexI:Whole_weight              0.085932998
SexI:Shucked_weight            0.178059423
SexI:Viscera_weight            .          
SexI:Shell_weight              .          
SexM:Length                    .          
SexM:Diameter                  .          
SexM:Height                    .          
SexM:Whole_weight              0.006695775
SexM:Shucked_weight            .          
SexM:Viscera_weight            .          
SexM:Shell_weight              0.038776225
Length:Length                 -0.952896541
Length:Diameter                .          
Length:Height                  .          
Length:Whole_weight            .          
Length:Shucked_weight         -0.044124697
Length:Viscera_weight         -0.162235445
Length:Shell_weight            .          
Diameter:Diameter             -0.213678610
Diameter:Height                .          
Diameter:Whole_weight          .          
Diameter:Shucked_weight       -0.174917410
Diameter:Viscera_weight        .          
Diameter:Shell_weight          .          
Height:Height                 -1.584325310
Height:Whole_weight            .          
Height:Shucked_weight         -2.134199329
Height:Viscera_weight          .          
Height:Shell_weight            .          
Whole_weight:Whole_weight      .          
Whole_weight:Shucked_weight    .          
Whole_weight:Viscera_weight    .          
Whole_weight:Shell_weight      .          
Shucked_weight:Shucked_weight  0.750410360
Shucked_weight:Viscera_weight  .          
Shucked_weight:Shell_weight   -0.009192374
Viscera_weight:Viscera_weight  .          
Viscera_weight:Shell_weight   -0.604089861
Shell_weight:Shell_weight     -0.497448802
```

Figure 11: model build by elastic net

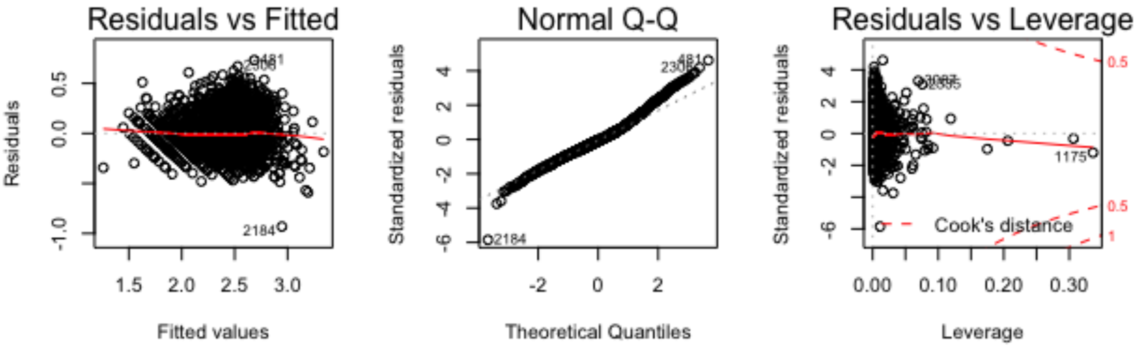


Figure 12: final model