Prediction for the age of abalone using multiple linear regression analysis

Sin Wing Leong

Abstract: Multiple linear regression is applied for predicting the age of abalone. It was found that multicollinearity between predictors, which makes the prediction more difficulty. Initially, ridge regression is used.Model selection using LASSO is applied in the model .

I. Introduction

Abalone is one of marine snails distributed aggregately. For different place they survive, they most likely show different appearance and recover at different rate when the place is favourable. For example, Blacklip abalone generally lives where complex rocky substrate supporting their inhabitation in Australia. Hence, different species of abalone with difference appearance is in favour of different environment, impacting how long they lives for. The quota of catching the abalone varied from time to time in the past and decreased substantially from 1980s (Warwick, 1994). Investigation of the age of abalone is essential as it related to how many quotas to catch them. In this article, the multiple linear regression with analysis are used to predict the age of abalone.

II. Data description

In this research, the data with no missing value is obtained from Tasmania from the past research project by Warwick Nash in 1994. The feature contains abalone's sex, length(mm), diameter(mm), height(mm), whole weight(grams), shucked weight(grams), viscera weight(grams) and shell weight(grams), which are potential predictors. The response is how many rings they have as the rings are generally determined for the age of abalone by counting the number of rings ranging from 1 to 29. In this dataset, sex is the only categorical variable and the others are continuous variable. Rings are the only integer.

Scatter Plot

Scatter plot showed the variable matrix which is graphically convenient to proceed data analysis.

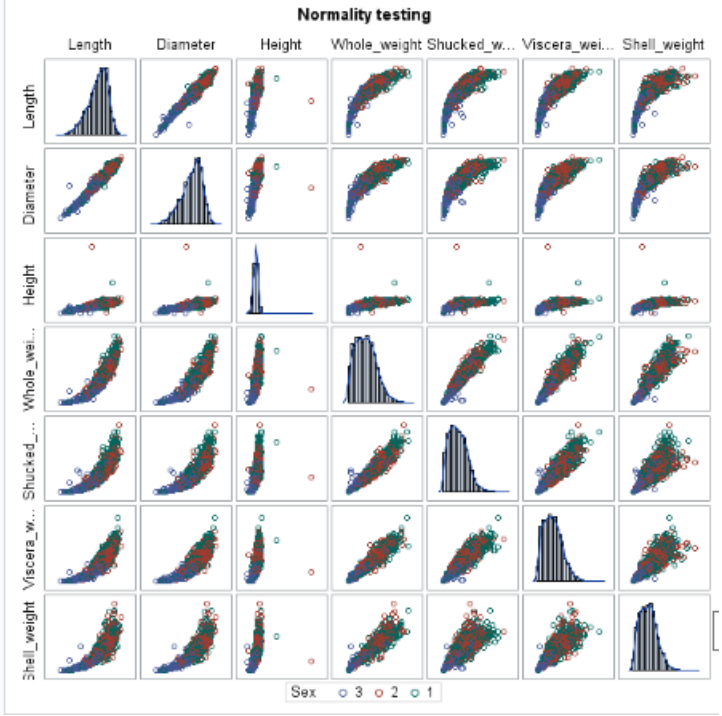


Figure 1 Scatter plot for all variable with different sex level

For the diagonal graph, there is kernel density estimation and histogram for each variable showing the distribution of each predictor. For Whole weight, Shucked weight, Viscera weight and shell weight, there is right-skewed distribution but length and diameter may exist different distribution from the graph.

Additionally, the relationship between diameter and length is almost linear. It is noticeable that the graphs in the first and second row are almost the same, which it may exist special property between length and diameter. The graph for 4 weight variables looks similar.

In sense of extreme value, some potential extreme points which are far away from the trend of scatters are obviously in the scatter matrix.

Analysis of the response

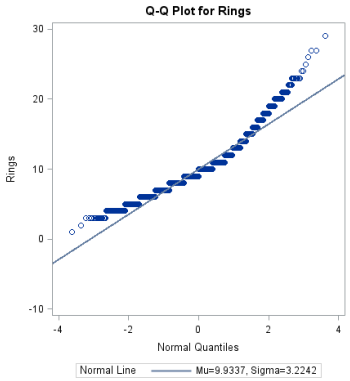


Figure 2QQplot for response(Rings)

Obviously in Figure 2,Y does not follow the normal distribution as it showed skew-right distribution. By nonparametric test, the result was showed below.

Test	Statistics	p-value
Kolmogorov-Smirnov	D = 0.145374	< 0.0100
Cramer-von Mises	W-Sq = 13.99257	< 0.0050
Anderson-Darling	A-Sq = 78.60015	<0.0050

Set significance level α = 0.05. The null hypothesis (H_0) is that y follows normal distribution. In contrast, the alternative hypothesis (H_1) is that y does not follow normal distribution. For three tests above, p-value < 0.01< $\alpha = 0.05$. Therefore, we reject H_0 *so the rings data does not follow normal distribution at significance level* α.

By the hypothesis above, Rings do not follow normal distribution and showed skewed distribution, which implies transformation on the response maybe one of methods for prediction but it is not necessary.

III. MULTIPLE LINEAR REGRESSION

Initial Regression Model (Ordinary Linear Regression)

$$y_i = \beta_0 + \sum_{k=1}^{2} \alpha_k * g_{i,k} + \sum_{j=1}^{7} \beta_j * x_{i,j} + \sum_{k=1}^{2}\sum_{j=1}^{7} \gamma_{k,j} * g_{i,k} * x_{i,j} + e_i \quad and\ assume\ \ e_i \sim N\ (0, {\sigma_i}^2)$$

, where α_k and β_j are coefficient of sex with k[th] group and j[th] continuous variable respectively. $\gamma_{k,j}$ is the interaction term between sex with k[th] group and j[th] continuous variable and β_0 is the intercept. To obtain inversible design matrix, we treat 3[rd] sex as a reference group so k is either 1 or 2 without 3.

For other form of regression model for k[th] sex abalone

$$y_{ik} = \beta_0 + \alpha_k + \sum_{j=1}^{7} \beta_j * x_{i,j} + \sum_{j=1}^{7} \gamma_{k,j} * x_{i,k} + e_{ik}$$

By performing $min \sum_{i=1}^{4177} (\hat{y_i} - y_i)^2$ respect to those coefficients, we obtain

Coefficient table ($R^2 = 0.55283$)

Coefficient	α_1	α_2	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Estimate	2.2544	5.1209	-2.7461	5.9345	28.825	8.2272	-14.662	-11.279	10.661
$\gamma_{1,j}$			2.3165	-0.87997	-13.925	0.44225	-4.1306	1.1095	0.018891
$\gamma_{2,j}$			-5.8861	6.4544	-25.111	2.4390	-6.5638	2.4779	-3.4173

Residual Analysis

In figure 3a, Q-Q, it demonstrated that residual does not follow $N(0, \sigma^2)$ as the scatter point does not lie on the diagonal line.

In figure 3b, standardized residual plot showed there is dispersion along the predicted value, which indicated that σ^2 is suspected not to be a constant which is one of the assumptions.

Therefore, the original regression model is invalid as it is under invalid assumption.

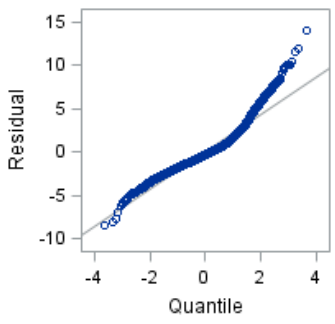


Figure 3a Q-Q plot for the Residual.

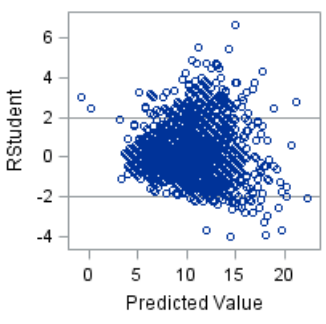


Figure 3b R-student standardized Residual plot for the regression model

Transformation on y_i

To solve the problem of invalid assumption, the transformation on y_i is initially investigated. By Box-Cox Transformation,

$$y_i^* = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & if\ \lambda \neq 0 \\ log(y_i)\ if\ \lambda = 0 \end{cases}$$

To determine the value of λ, it is regarding to maximization of *Loglikelihood* (y_i^*). In the following, it showed loglikelihood by λ

Box-Cox Transformation Information for Rings

λ	−2.0	−1.5	−1.0	−0.5	−0.0	−0.5	−1.0	−1.5	−2.0
Loglike (y_i^*)	-7992	-5112	-3387	-2696	-2584	-2792	-3220	-3830	-4601

By Box-Cox Transformation, the optimal $\lambda^* = 0.0$, which mean $y^* = log(y_i)$. The R^2 is 0.6237 which higher than 0.55283. Root Mean square error (MSE)= 0.19655

Comparing to figure 2a and figure 2b, Q-Q plot in figure 2a is more nearby the diagonal line, which showed it possibly follow normal distribution. For the R-studentized residual plot in figure 4b, it demonstrated no dispersion along the line $Rtudent = 0$. It may support the normality assumption.

However, it should be formally tested for the assumption whether it is valid or not.

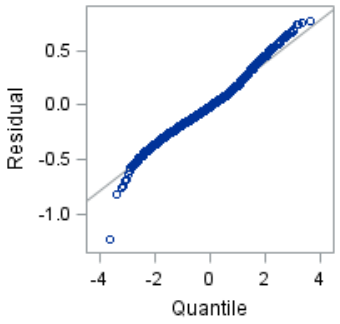


Figure 4a Q-Q plot for the Residual.

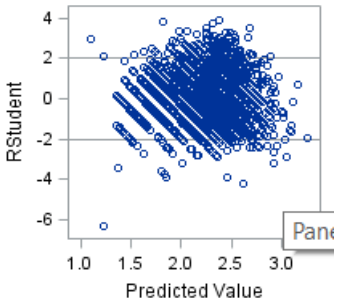


Figure 4b R-student standardized Residual plot for the regression model

Test	Chi-Square	p-value
Test of First and Second Moment Specification	305.73	< 0.001

The null hypothesis is jointly statement which consisted of homoscedastic error, independent error each other and the alternative hypothesis is that it does not satisfy at least one of statements. From Test of First and Second Moment specification, p-value is $< 0.001 < \alpha = 0.05$. Therefore, we reject this assumption. Then we should test whether R-studentized residual follows normal distribution. To investigate this result, we consider normality test.

Test	Statistics	p-value
Kolmogorov-Smirnov	D = 0.052358	< 0.0100
Cramer-von Mises	W-Sq = 2.897685	< 0.0050
Anderson-Darling	A-Sq = 17.93305	<0.0050

The table on the left showed three tests for the normality assumption. p-value $< 0.01 < \alpha = 0.05$. Therefore, we reject H_0 so the R-studentized residual does not follow normal distribution.

By the tests above, even though using Box-Cox transformation on y_i, the normality assumption is not satisfied.
Instead of another transformation on y_i or predictors, detecting the autocorrelated error first.

Test	Statistic
Durbin-Watson D	D = 2.013

By the Durbin-Watson statistic, the value near 2.0 showed we cannot reject the assumption of independent errors. Therefore, it is believed that the independence of error should not be rejected so in the following, detecting outliers and influential points is essential to fit the model correctly.

If outliers exist, they significantly impact the value of the intercept by only one or few observations so detecting them is valuable to fit the model correctly instead of including them. It is found that the maximum R-studentized residual is 6.3670 at 564th obs. Therefore, by the criteria of $|t_i| > 3$ to identify the outliers, it is currently suspected that the outliers exist. After searching for observations with $|t_i| > 3$, there are 31 observations with $|t_i| > 3$. However, Even though detecting them, the normality assumption is not satisfied and obtain the similar p-value result mentioned above. Moreover, as there are multiple outliers, it is hard to delete a set of outliers or one outlier without normality assumption by Bonferroni correction. Hence, keeping them into the consideration of fitting model is possible for further investigation.

Therefore, finding another problem are more efficient to find the optimal model, such as multicollinearity problem. Even though we can choose transformation on x_i but it makes more complicated interpretation so it is used only if detecting multicollinearity cannot solve the invalid assumption mentioned above.

Multicollinearity

Multicollinearity means the variables correlated highly each other such that the value of one of them may leads to another variable. However, the regression model relies on the independence of variables. Otherwise, the determinant of the predictor is near 0 so the design matrix is singular or the regression coefficients varies much as it exists linearly dependent column vectors.
Intuitive situation is that for the abalone case, the total weight of abalone and shell weight are suspected highly correlated as the simple physical property between them exists.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

,where $p_{X,Y}$ is the Pearson correlation coefficients.

Pearson Correlation Coefficients

	Length	Diameter	Height	Whole weight	Shucked	Viscera	Shell
Length	1.0000	0.98681	0.82755	0.92526	0.89791	0.90302	0.89771
Diameter	0.98681	1.0000	0.83368	0.92545	0.89316	0.89972	0.90533
Height	0.82755	0.83368	1.0000	0.81922	0.77497	0.79832	0.81734
Whole weight	0.92526	0.92545	0.81922	1.0000	0.96941	0.96638	0.95536
Shucked	0.89791	0.89316	0.77497	0.96941	1.0000	0.93196	0.88262
Viscera	0.90302	0.89972	0.79832	0.96638	0.93196	1.0000	0.90766
Shell	0.89771	0.90533	0.81734	0.95536	0.88262	0.90766	1.0000

From the Pearson correlation coefficients, all the variable may contain positive linear correlation with each other as all the value is near 0.9 which implies strong linear correlation.
Therefore, Analysis of those linear-correlated variables is needed to conducted. Variance Inflation Factor (VIF) is one of tools to detect the correlated variables.

Factor	VIF
Sex = 1	96.236
Sex = 2	151.50
Length	142.35
Diameter	159.02
Height	36.705
Whole weight	854.40
Shucked	208.76
Viscera	197.47
Shell	167.86

The table on the left showed VIF for each predictors. By the criteria VIF > 10 (Hair et al., 1995), all the variables in the model maybe problematic. For instance, multilinearity leads to large variance for each regression coefficient such that there is significant large band-width confidence for certain distribution. The regression coefficients are not very confident in the model so multicollinearity is problematic. However, instead of VIF, investigating condition index should be done as analysis of multicollinearity is not yet finished.

Factor	Condition Index
Sex = 2 and Shucked weight	111.46
Sex = 2 and Viscera weight	214.46
Sex = 2 and Shell weight	219.38

There are 3 cases listed below and those cases contains large condition index, which condition index > 100 is one of criteria for finding highly correlated variables. When interaction term between sex = 2 and Viscera weight, eigenvalue proportion of length and diameter are 0.68904 and 0.69414 respectively. It indicated that highly correlated relationship between length and diameter. This result is consistent with the Pearson correlation coefficient and VIF. Therefore, it is highly suspected that there is highly correlated relationship between them.

When interaction term between Sex = 2 and shell weight, the eigenvalue proportion of whole weight, shucked weight, viscera weight and shell weight are 0.91403,0.57997,0.39740 and 0.42986, which showed that whole weight and viscera are suspected to be highly correlated variables.

By the test statement mention above, multicollinearity problem is detected. It existed highly-correlated variables. However, the difficulty next is how to deal with this problem as most of regression variables are highly correlated with some of them. If deleting them casually, it may lead underfitting probably. At least, tackling the multilinearity problem is the first mission.

Ridge regression

In order not to delete those variables, ridge regression is proposed to use for prediction. The ridge regression model is the same as y^* before but the regression coefficient by min $(\sum_{i=1}^{4177}(y-y*)^2 + \lambda \sum_{j=1}^{p} \beta_j^2)$ respect to β_i, where p is the amount of regression coefficient excluding the intercept. By lowering dimension of the regression model using the penalty term λ, the result is shown.

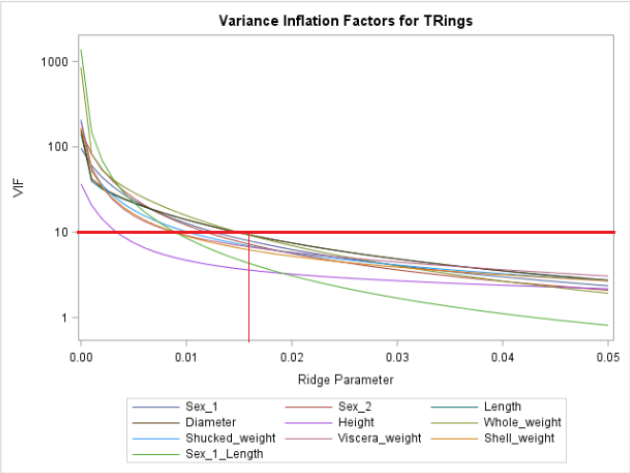


Figure 4a VIF graph with λ

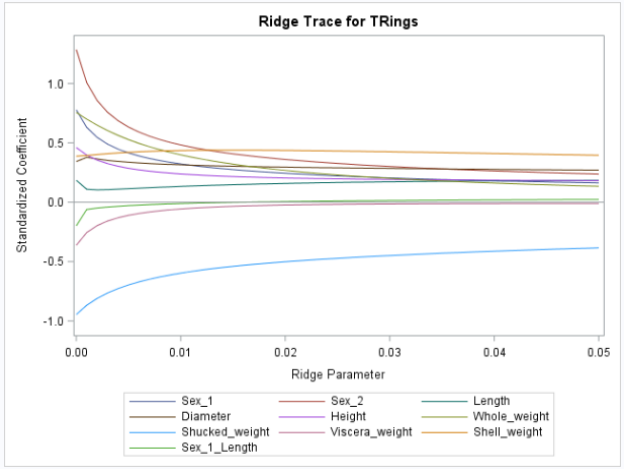
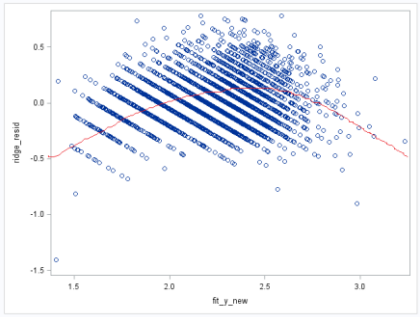


Figure 4b Standard deviation graph with λ

Choice of λ

In Figure 4a, it showed VIF for each variable respect to λ and standard deviation of coefficient respect to λ. As the criteria of VIF > 10, λ = 0.02 is conveniently chosen for ridge regression as the graph of ridge trace demonstrated that the standard deviation is stable comparing to λ < 0.02. In Figure 4a, it showed all VIFs of the coefficients are below 10 in the training dataset.

Residual

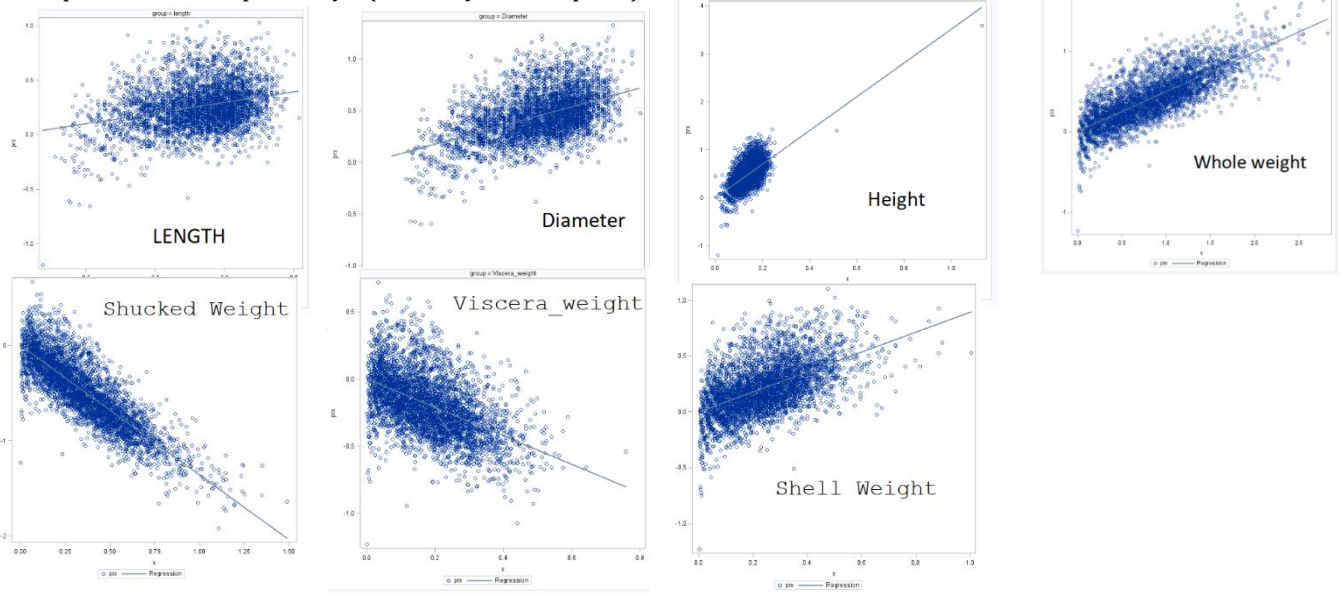


By the residual plot on the left, it showed there are nonlinearity for the plot of ridge regression residual on predicted value. Therefore, it is valuable for considering transformation on predictors.

Transformation on *predictors*

The transformation mentioned above showed transformation on y_i but it is noticeable that the relationship between predictors and the response is essential.

This is the partial residual plot for y^* (Ordinary Least Square)



By the partial residual plot, there are patterns in length , diameter and height with a tail in low predictor value. For the several weights, they behave linearly with partial residual. Hence, by partial residual plot, it is observed that transformation for length, diameter and height may be potential transformation for the linear regression.

Box-Tidwell Transformation

$$y_i{}^{*} = \beta_0 + \sum_{k=1}^{2} \alpha_k * g_{i,k} + \sum_{j=1}^{7} \beta_j * w_j(x_{i,j}) + \sum_{k=1}^{2}\sum_{j=1}^{7} \gamma_{k,j} * g_{i,k} * w_j(x_{i,j}) + e_i \qquad (1)$$

$$\text{,where } w_j(x_{i,j}) = \begin{cases} x_{i,j}{}^{\zeta_j} & if\ \zeta_j \neq 0 \\ log(x_{i,j}) & if\ \zeta_j = 0 \end{cases}$$

Motivation to choose the Box-Tidwell Transformation is that it is easy to choose the power index of $x_{i,j}$. If we choose naïve transformation, the power index of $x_{i,j}$ is very subjective and the choice to choose $log(x_{i,j})$ is not systematic.
As Box-Tidwell transformation is only valid for the continuous variable. For simplicity,

$$log(y_i) = \beta_0 + \sum_{j=1}^{7} \beta_j * w_j(x_{i,j}) + e_i$$

Even though our model is not the equation above, it may give us Box-Tidwell estimates for us to take reference. Therefore,

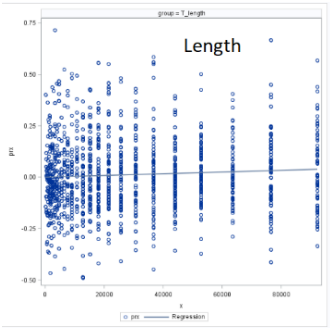
Box-Tidwell estimate Table

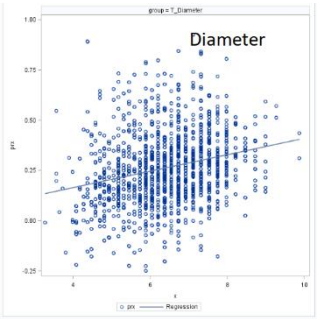
	Power	Standard error	Chosen power
Length	-22.750	0.98681	-22.75
Diameter	-2.7200	1.0000	-2.75
Height	0.1843	0.83368	0.20
Whole weight	0.5345	0.92545	0.50
Shucked	0.3958	0.89316	0.40
Viscera	1.0250	0.89972	1.00
Shell	11.3344	0.90533	11.0

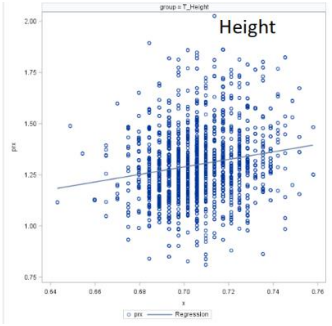
Then simplified expression is valid for Box-Tidwell transformation.
By the table on the left, it initially gives us power transformation for each predictor. However, for convenience, we use power of length, diameter, height, whole weight, shucked weight, viscera weight and shell weight to be chosen power shown and it make interpretation more easily.

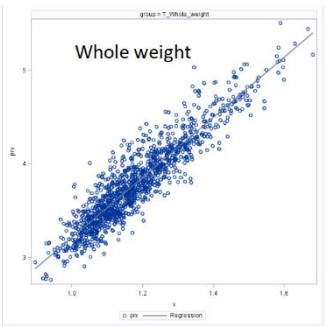
Addictive model

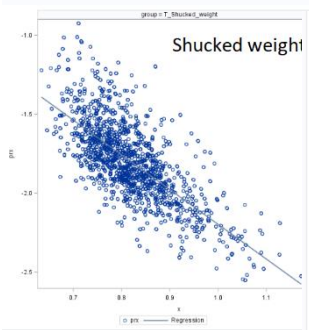
Here is the model,
$y^{*} = \beta_0 + \sum_{k=1}^{2} \alpha_k * g_{i,k} + \sum_{j=1}^{7} \beta_j * x_{i,j} + \sum_{k=1}^{2}\sum_{j=1}^{7} \gamma_{k,j} * g_{i,k} * x_{i,j} + \left(\sum_{j=1}^{7}\sum_{l=1}^{p} s_{l(x_{i,j})}\right) + \sum_{k=1}^{2}\sum_{j=1}^{7}\sum_{l=1}^{p} p_{k,j,p} * g_{i,k} * s_{l(x_{i,j})}$
, where $s(x_{i,j}) = \sum_{l=1}^{p} s_{l(x_{i,j})}$ is non-linear transformation function, p is the amount of how many components of non-linear transformation. Therefore, $\sum_{j=1}^{7}\sum_{l=1}^{p} s_{l(x_{i,j})}$ is the non-linear effect on $x_{i,j}$. $\sum_{k=1}^{2}\sum_{j=1}^{7}\sum_{l=1}^{p} p_{k,j,p} * g_{i,k} * s(x_{i,j})$ is the interaction term for sex and those non-linear terms.
However, the interaction term between group and non-linear transformation term make interpretation extremely difficultly.

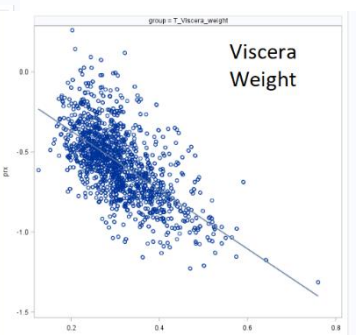


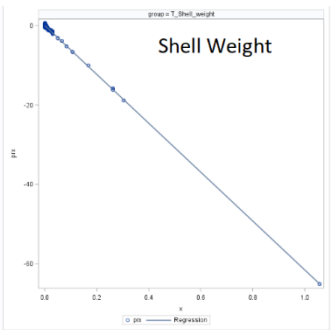












After Box-Tidwell_Transformation, the partial residual plot looks linear for each variable and the variance maybe a constant. MSE $= 0.18911 < 0.19655$ and $R^2 = 0.6516 > 0.6237$ so we do not reject this method in term of R^2 and MSE. The linearity and equal variance enable us to adopt ridge regression for the model with transformed x and y.

Model selection

Ridge regression is applied to Box-Tidwell transformed and Box-Cox Transformed model. Before further analysis, deleting influential point is essential. By DFFITS, COV RATIO, LEVERAGE, there are 33 observations simultaneously fulfilled the requirement for suspected influential point among 4177 observations, such as 110th, 121th, 239th etc. For example, the 110th observation fulfilled DFFITS $> 2 * (p'/n)^{1/2}$, |COV RATIO -1|> 3*p'/n and $h_{ii} > 2$*p'/n. Three tests are simultaneously used as the data is valuable and deleting much data is unreasonable. Therefore, they are deleted first before further exploration.
Full model is the model in (1) and the power chosen in Box-Tidwell estimate Table.

By stepwise regression (not applied ridge regression and only OLS),
There are 15 variables (including interaction term) and only omitted single variable in the stepwise regression is Shell weight.
For simplicity, we define
Length as X_1, Diameter as X_2, Height as X_3, Whole weight as X_4, Shucked weightas X_5, Viscera weightas X_6 , Shell weight as X_7

Stepwise Regression(Criteria $= C_p$)

Estimate table by Stepwise regression

Variable	$\boldsymbol{X_1^{-22.75}}$	$X_2^{-2.75}$	$X_3^{0.20}$	$X_4^{0.50}$	$X_5^{0.40}$	$X_6^{1.00}$	$X_7^{11.0}$	$g_1 * \boldsymbol{X_1^{-22.75}}$	$g_1 * X_2^{-2.75}$	$g_1 * X_3^{0.20}$	$g_1 * X_4^{0.50}$
Coeff Estimate	3.26841E-26	-0.00095434	0.95656	2.93693	2.94554	1.27429	0	2.151E-19	-0.00203	0	0

Variable	$g_1 * X_5^{0.40}$	$g_1 * X_6^{1.00}$	$g_1 * X_7^{11.0}$	$g_2 * \boldsymbol{X_1^{-22.75}}$	$g_2 * X_2^{-2.75}$	$g_2 * X_3^{0.20}$	$g_2 * X_4^{0.50}$	$g_2 * X_5^{0.40}$	$g_2 * X_6^{1.00}$	$g_2 * X_7^{11.0}$
Coeff Estimate	-0.67599	0	0	0	-0.00669	0	-0.44118	-0.59413	0.46634	0

$\alpha_1 = 0.48909$, $\alpha_2 = 0.80815$ and β_0= 1.17373, R^2 = 0.6567 and C_p = 21.5469, AIC = -13922 and SBC = -13820

Best Subset(Criteria $= C_p$)

The Best C_p subset included all transformed 7 continuous variables, 2 sex and interaction terms including Sex = 1 with all continuous variables except Height and Sex=2 with diameter, whole weight, shucked weight, viscera weight and shell weight. That means there are 21 variables including interaction term and intercept in the model. R^2 = 0.65765 and C_p = 19.5822, AIC= -13923.61 and SBC = -13790.69.

Elastic Net

Elastic net method is used to select variable by shrinkage. It is used here but not the other because the graph and statistics showed that there is multicollinearity among most of variables. The Elastic net is one of method to lower dimension of the predictors so as to solve multicollinearity problem. It is more useful than ridge regression as Elastic net combines both ridge and LASSO regression. Ridge regression tends not to delete the variables but LASSO does.
The result from elastic net is that it includes 7 continuous variables , Sex = 1, Sex = 1 and interactions between Sex = 1 and all continuous variables except whole weight and interactions between Sex = 2 with length, diameter, shucked weight, Viscera weight.

Statistics

	# of variables	R^2	C_p	AIC	SBC
Stepwise Regression	15	0.6567	21.5469	-13922	-13820
Best Subset	20	0.65765	19.5822	-13923.61	-13790.69
Elastic Net	19	0.6545	22.40619	-9741.87425	-13761

From the table on the left, it showed the similar result for three selection method.

Therefore, it is suggested to choose elastic net for model selection as it combined LASSO and ridge regression. They relieved the multicollinearity problem even though AIC is higher among these methods.

Finally, check the influential point for regression model back. Using the same criteria, it is detected as an influential point wen it fulfilled simultaneously three thresholds of the influence criteria, such as DFFITS. Now, we check 21 observations are still influential point. Therefore, we put 12 observation back to regression model and refit the model.

Therefore, the optimal regression model chosen:

Variable	$\boldsymbol{X_1^{-22.75}}$	$X_2^{-2.75}$	$X_3^{0.20}$	$X_4^{0.50}$	$X_5^{0.40}$	$X_6^{1.00}$	$X_7^{11.0}$	$g_1 * \boldsymbol{X_1^{-22.75}}$	$g_1 * X_2^{-2.75}$	$g_1 * X_3^{0.20}$	$g_1 * X_4^{0.50}$
Coeff Estimate	4.27381E-26	-0.00111	1.108409	2.622621	-2.823633	-0.719818	-0.110056	1.826496E-19	0.001589	0.204502	0

Variable	$g_1 * X_5^{0.40}$	$g_1 * X_6^{1.00}$	$g_1 * X_7^{11.0}$	$g_2 * \boldsymbol{X_1^{-22.75}}$	$g_2 * X_2^{-2.75}$	$g_2 * X_3^{0.20}$	$g_2 * X_4^{0.50}$	$g_2 * X_5^{0.40}$	$g_2 * X_6^{1.00}$	$g_2 * X_7^{11.0}$
Coeff Estimate	-0.415343	-0.404304	0	-7.09487E-14	-0.001885	0	-0.44118	-0.556992	-0.337675	-0.114051

$\alpha_1 = 0.253363$, $\alpha_2 = 0.479406$and β_0= 1.168872

IV. Discussion

After fitting model, it showed that the residual does not follow normal by normality test. However, when linearity and equal variance probably holds, the LASSO or ridge regression model is still valid. Analysis of the residual indicated the direction of the modification of the model, especially the residual plot. However, the way to improve the model is that cross validation is suitable for applying estimation for parameter, which enables analysis based on data and not to be surjective.

It is observed that there are two unusual Box-Tidwell transformed variable, i.e. length and shell weight as the power terms are extreme. Analysis of this relationship maybe beneficial to learning the regression model. Viscera weight's power is near 1 which means not transformation so we consider it is simple independent variable to analyze the model. Regarding $X_2^{0.2}$ and X_2 between 0.1 to 0.3 usually, $0.1^{1/5}$ = 0.63100 and $0.3^{1/5}$ = 0.78600 which is small difference for range of Height so this variable looks like constant contributed to $log(y_i)$. It interpreted as it is important but it does not impact y_i significantly with coefficient only near 1.1084, which means height does not gives us clear picture for the age of abalone.

Instead of the model mentioned, there is other model which can also apply, such as cubic spline. However, the coefficient term as said increased a lot. When amount of coefficient increases, The model selection is hard to select those coefficient. At this moment, C_p is no longer appropriate for model selection. To reduce the model predictors which caused overfitting, we should choose a model selection method which tends to select fewer independent variables. For example, LASSO is one of potential regression model for both regression or model selection. In addition, there were advanced regression model in recent year which develop on the basis of LASSO. For prediction of age of abalone, regression tree or new methods may also apply for this data set.

References

[1] 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 Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)

[2] *Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). Multivariate Data Analysis (3rd ed). New York: Macmillan.*