Abalone Regression Analysis

Evan Garcia

2024-03-16

Introduction

Abalone are a type of shellfish that resemble mussels or oysters and are found along coastal waters; their meat is considered a delicacy and they are farmed or fished around the world. The Abalone dataset contains information collected in 1994 of 4177 observed organisms with 8 features: Sex, length, diameter, height, Whole weight, shucked weight, viscera weight, shell weight, and the response variable rings. The sex variable corresponds to the sex of the Abalone observed, as a Male, Female, or Infant, where the sex of the organism is unclear. The continuous variables have all been scaled by neural network prior to download, dividing the measurements by a factor of 200 where the volumetric measurements are done in millimeters and weight measurements conducted in grams. Whole weight refers to the total weight of the Abalone and the Shucked weight refers to the weight of the edible meat of an Abalone. Viscera Weight refers to the weight of the organs and the shell weight to the weight of the shell. Analyzing this data and fitting a regression model with respect to the age of Abalone may be useful since it could allow for age predictions without having to manually cut open each organism and count the number of rings present.

Data Processing

Fortunately all missing values were removed prior to the publishing of the data, so the values were convenient to work with. The rings variable is simply a way to describe the age, where adding 1.5 to the rings becomes the age of an Abalone since thats how many years it takes for rings to being to show in the species. The Age variable becomes the replacement for rings in the table. After fitting and initial linear model to check for notable predictors it becomes apparent the Male sex is not a significant predictor of age. It also becomes clear that Length does not significantly affect the model so that will be looked into as well.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.3946414	0.2915665	18.5022693	0.0000000
SexI	-0.8248763	0.1023953	-8.0558017	0.0000000
SexM	0.0577157	0.0833465	0.6924791	0.4886751
Length	-0.4583354	1.8091246	-0.2533465	0.8000129
Diameter	11.0751025	2.2272802	4.9724783	0.0000007
Height	10.7615367	1.5362029	7.0052833	0.0000000
Whole_Weight	8.9754446	0.7254039	12.3730299	0.0000000
Shucked_Weight	-19.7868669	0.8173500	-24.2085592	0.0000000
Viscera_Weight	-10.5818270	1.2937489	-8.1791972	0.0000000
Shell_Weight	8.7418058	1.1247315	7.7723492	0.0000000

Analyzing the differences in average values for each predictor based on sex, there is a marginal difference between the males and the females, however the status of an infant demonstrates a clear difference in the values. As such, a more appropriate replacement for the sex category is the feature “Is_infant,” which will be a true and false value for each instance.

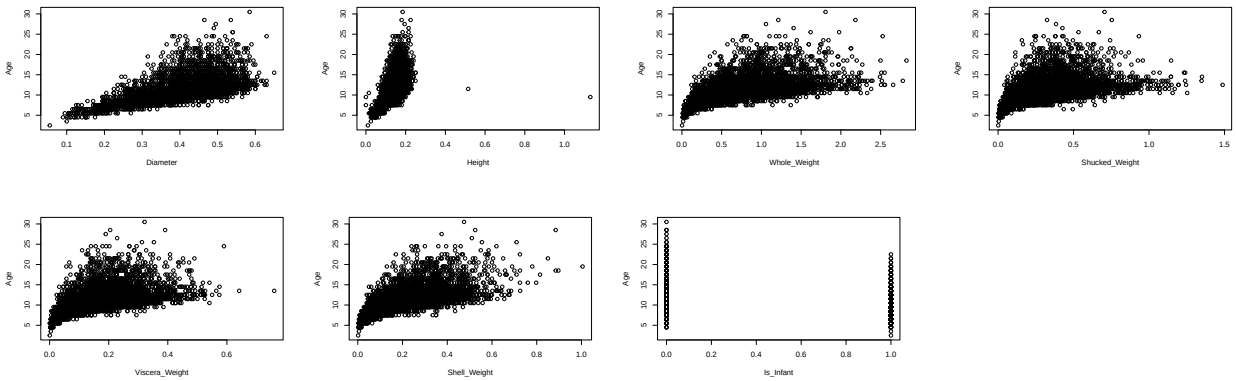
	Length	Diameter	Height	Whole_Weight	Shucked_Weight	Viscera_Weight	Shell_Weight	Age
F	0.57909	0.45473	0.15801	1.04653	0.44619	0.23069	0.30201	12.62930
I	0.42775	0.32649	0.10800	0.43136	0.19104	0.09201	0.12818	9.39046
M	0.56139	0.43929	0.15138	0.99146	0.43295	0.21554	0.28197	12.20550

In addressing the significance of the length variable, feature selection is in order to assuredly determine which of the variables should be included in a proper model of the Age prediction. Running an exhaustive model of all possible combinations of the features, a table is constructed to show highligh the adjusted R^2, BIC, and CP for each set of features. It becomes clear that the 7-variable model that excludes length presents the best model with respect to accounting for the variance of the data, so going forward length will be excluded. It is worth noting that the multicollinearity of the predictor variables is high as shown in the table below, however the exhaustive search of variables trumps this fact and reassures that the set of predictors selected will result in the model based on the most accurate predictors, which is further confirmed in the significance levels shown later.

	Length	Diameter	Height	Whole	Shucked	Viscera	Shell	Infant	AdjR2	BIC	CP
1-Var							*		0.3937	-2074.442	1293.4805
2-Var					*		*		0.4737	-2657.910	573.3437
3-Var		*			*		*		0.5020	-2881.528	319.0641
4-Var		*			*		*	*	0.5146	-2981.588	206.0929
5-Var		*		*	*		*	*	0.5248	-3063.020	115.1122
6-Var		*		*	*	*	*	*	0.5317	-3116.847	53.8848
7-Var		*	*	*	*	*	*	*	0.5370	-3157.153	7.0615
8-Var	*	*	*	*	*	*	*	*	0.5369	-3148.878	9.0000
VIF	40.943	42.349	3.579	109.764	28.452	17.417	21.263	1.518	NA	NA	NA

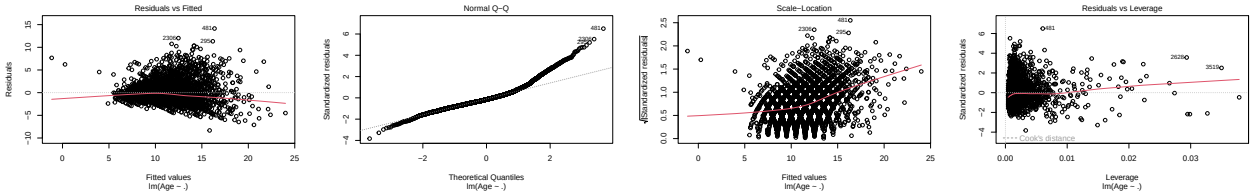
Descriptive Analysis or Statistics

Analyzing some summary statistics of the model, it becomes apparent that the maximum height value is particularly large and may need to be removed, so upon closer inspection, the 1.1 value in the height plot is a particularly egregious leverage point and thus removed. Beyond the height inconsistency, the rest of the points appear to be within a reasonable range of their means. The models of the variables directly compared against the age also show there is a possibility of a linear relationship between the variables, as the age seems to increase with all the measurement variables compared to their low values.



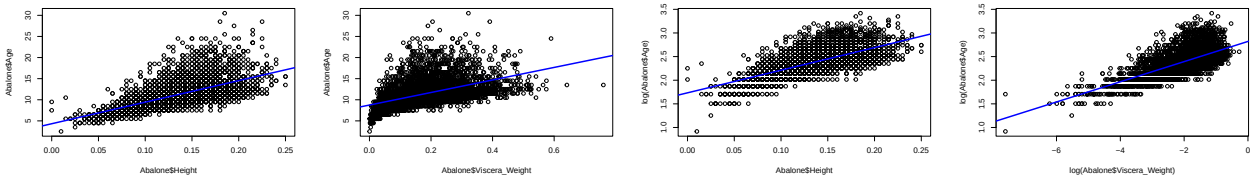
	Diameter	Height	Whole_Weight	Shucked_Weight	Viscera_Weight	Shell_Weight	Age
Min.	0.0550	0.0000	0.0020	0.0010	0.0005	0.0015	2.50
1st Qu.	0.3500	0.1150	0.4415	0.1860	0.0935	0.1300	9.50
Median	0.4250	0.1400	0.7995	0.3360	0.1710	0.2340	10.50
Mean	0.4079	0.1395	0.8287	0.3594	0.1806	0.2388	11.43
3rd Qu.	0.4800	0.1650	1.1530	0.5020	0.2530	0.3290	12.50
Max	0.6500	1.1300	2.8255	1.4880	0.7600	1.0050	30.50

Assumptions of Regression Model



Observing the diagnostic plots of the initial model, there seems to be a slight pattern in the residuals vs fitted as well as a violation of the normality assumption, so actions must be taken to correct these violations.

The weight variables all seem to share a similar pattern relative to age so by inspection, a log transformation seems in order to help the linearity model. Additionally, although it may not be as apparent, transforming the Age variable with a log function also seems to aid in the model's ability to account for variance as it helped increase the adjusted R^2 value.

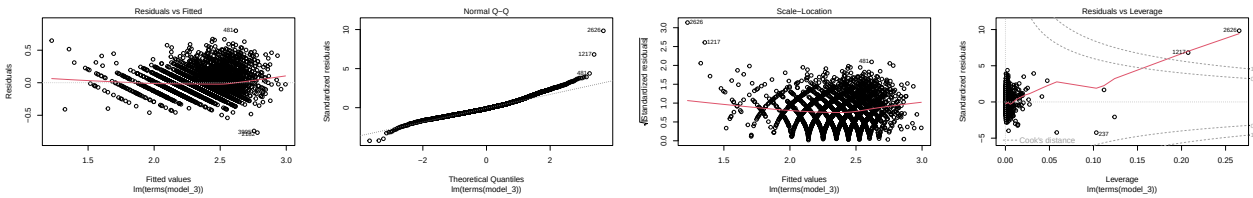


Furthermore, to aid in the model's fit of the values, weighted least squares is employed to help negate the the multitude of points that may be skewing the effectiveness of the model. The weights in this model are estimated based on the residuals of a linear model with the fitted values of the original model squared.

Linear regression Model

After applying transformation of the variables, in tandem with the application of weighted least squares, the new and improved regression model demonstrates a higher adjusted R^2 value.

As well, the F-Statistic is much higher, so another look at the diagnostic plots is appropriate.



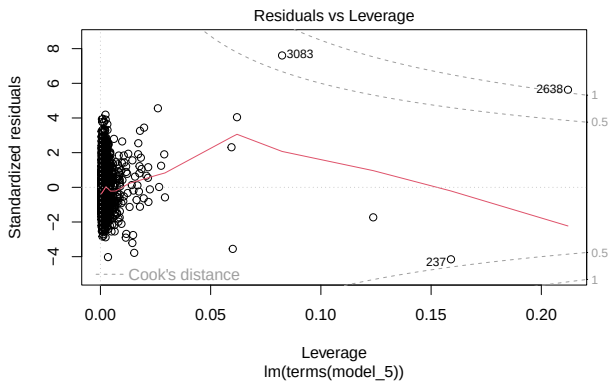
As a result of the weighted least squares application combined with the transformation, the normality assumption is much more valid, and the residuals seem to have more homoscedasticity. However, analyzing the residuals vs Leverage plot, a few of them lie outside of the cooks distance, so they are removed from the model.

Table 1: Before transformation and weighted least squares: F-statistic 709.4 and Adj. R2 = 0.54298

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.05056	0.27032	18.68349	0
Diameter	7.95265	1.04317	7.62357	0
Height	23.14281	2.27155	10.18813	0
Whole_Weight	8.85634	0.72087	12.28559	0
Shucked_Weight	-19.39830	0.81009	-23.94589	0
Viscera_Weight	-11.23103	1.28117	-8.76625	0
Shell_Weight	7.71613	1.12576	6.85416	0
Is_InfantTRUE	-0.82615	0.08896	-9.28702	0

Table 2: After transformation and weighted least squares: F-Statistic 1118 and Adj. R2 = 0.65249

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.71569	0.06531	41.58448	0.00000
Height	1.06404	0.17208	6.18340	0.00000
Is_Infant	-0.05794	0.00671	-8.63507	0.00000
Diameter	-0.52047	0.10050	-5.17855	0.00000
Log_Whole_Weight	0.22043	0.03456	6.37853	0.00000
Log_Shucked_Weight	-0.31339	0.01655	-18.94127	0.00000
Log_Viscera_Weight	-0.04545	0.01563	-2.90783	0.00366
Log_Shell_Weight	0.37759	0.01870	20.19616	0.00000



With these removals, the adjusted R^2 improves slightly. However the diagnostic plots reveals there are many more points like those removed, and it would take qmany more removals to lower the scale of the cook's distance plot to a reasonable level,harming the integrity of the data, so the rest of the observations

Table 3: Final Adj. R2 = 0.67771

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.49768	0.06469	38.61268	0.00000
Abalone$Height	0.80275	0.16992	4.72435	0.00000
Abalone$Is_InfantTRUE	-0.04886	0.00655	-7.45897	0.00000
Abalone$Diameter	-0.42765	0.09682	-4.41719	0.00001
log(Abalone$Whole_Weight)	0.45716	0.03691	12.38643	0.00000
log(Abalone$Shucked_Weight)	-0.48657	0.02092	-23.26222	0.00000
log(Abalone$Viscera_Weight)	-0.03318	0.01486	-2.23208	0.02566
log(Abalone$Shell_Weight)	0.30693	0.01844	16.64079	0.00000

will remain included in the model.

Finally the anova test all of the parameters are significant, passing the f-test. Since the independent variables pass both the t-test and the f-test, these are the significant predictors in the model predicting age of Abalone.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Height	1	11664.42517	11664.42517	6785.78426	0.000000
Is_Infant	1	306.03021	306.03021	178.03321	0.000000
Diameter	1	273.40567	273.40567	159.05386	0.000000
Log_Whole_Weight	1	538.27229	538.27229	313.14013	0.000000
Log_Shucked_Weight	1	1777.18766	1777.18766	1033.87967	0.000000
Log_Viscera_Weight	1	19.70074	19.70074	11.46092	0.000717
Log_Shell_Weight	1	476.00476	476.00476	276.91597	0.000000
Residuals	4165	7159.42756	1.71895	NA	NA

Conclusion

After applying a few different techniques to aid in the linear model's accuracy and validity of assumptions, the R^2 value did not reach particularly high percentages. This indicates that throughout the data set the noise term in the model is quite impactful and causes a reasonable amount of variance that simply cannot be explained. The relationship however is still strong between the measurements of an Abalone relative to its age. Overall the diagnostic plots of the adjusted model meet the assumptions of a linear model fairly well, retaining a reasonable normality level and displaying minimal, if any pattern in the residuals. The original predictors of Length and sex were not found to be significant and are not useful in determining the age of an Abalone.

Because all the results were scaled by a factor of 1/200 before the data was processed and the fact that the units of measurement for volume were all in mm, the coefficients between the predictors can be directly compared. Based on the best model's coefficient results, height has the greatest positive impact on determining the age of an Abalone. To re-transform the information into interpretable results, the predictor coefficients that did not receive a log transformation must be exponentiated. In the case for height, the coefficient becomes 2.231, meaning for every increase of 200mm in height, the age of an Abalone is predicted to increased by a factor of 2.231. Other predictor variables that were not log scaled follow a similar logic, so for the status of being an infant and the diameter of the Abalone, the predicted age decreases by a factor of 0.952 and 0.652 for every 200mm increase in diameter respectively. The weight variables of the Abalone were all log scaled, meaning that for every 1% increase of 200g, the predicted age will be affected will change by the factor of the coefficient. So for the whole weight, it is predicted if the whole weight were to increase by 2g, then the predicted age increased by a factor of about 45%, similarly if the shell weight were to increase by 2g then the predicted age would increase by about 30%. With regard to the negative weights, if shucked weight or viscera weight increase by 2g, then the predicted age decrease by about 48% and 33% respectively.

Given the nature of how things grow with age, the coefficients for height, whole weight, and shell weight intuitively make sense, since as the organism ages then its body will become larger. Additionally, the status of being too young to have an identified gender and simply being labeled an infant would naturally be correlated with a younger age in the abalone. However more reasoning is necessary for the other negative coefficients, such as how a larger diameter is predicted to relate to a smaller age. From this negative relation, it could be reasoned that younger abalone start off wider in their life cycle and become more elongated as they age, losing diameter length. Shucked weight as a negative predictor in the age of abalone suggests that after a certain point in aging, the usable meat that can be harvested from abalone decreases in weight; a reasonable assumption as some species redistribute their mass to different parts of the body as they age. Finally for the negative predictor of viscera, the weight of the organs seems to decrease with an increased age, where one could reason that the organs of an abalone simply get weaker over time and thus lose mass.

For future models, a more complex series of transformations may be able to further improve the linearity assumptions of the model, and a better look into the data collection and explanation of outliers would help explain the findings of a few peculiar observations.