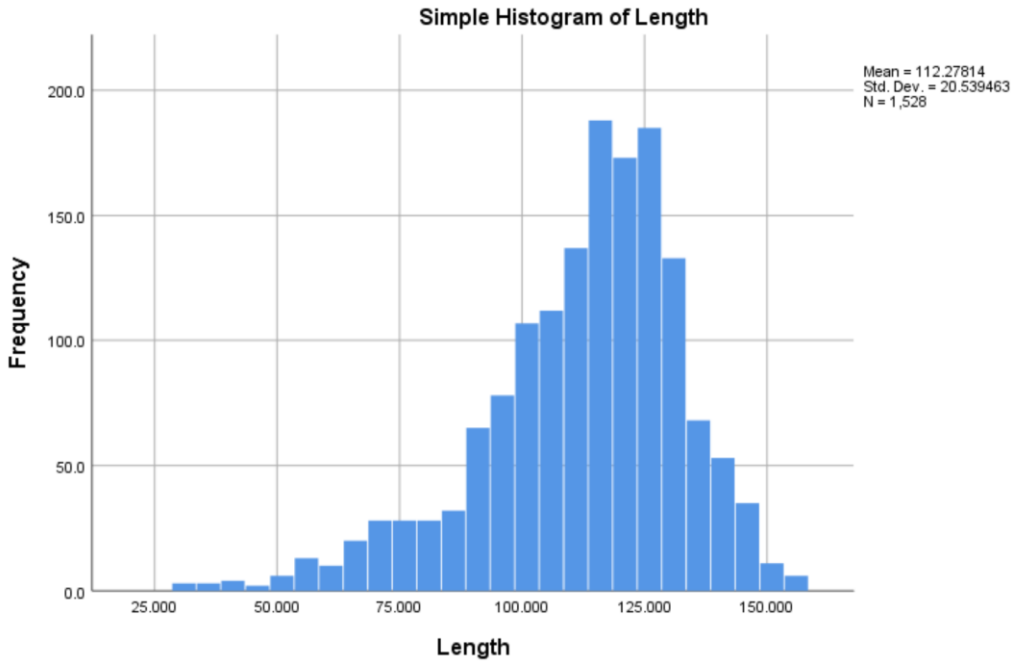
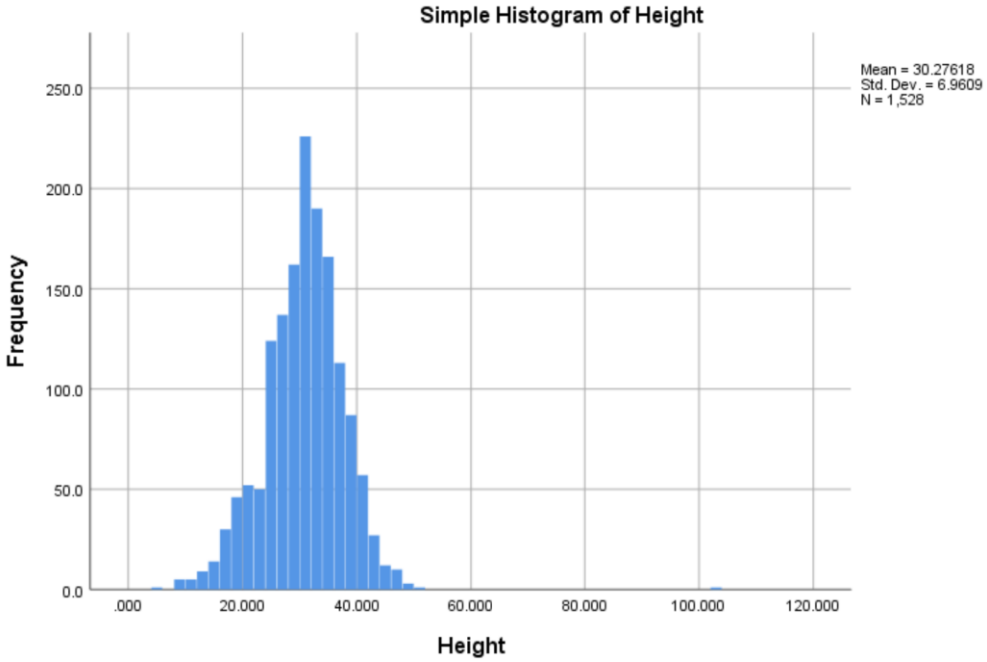
Assignment 1
44468423
Gautham Meenakshisundaram

a) Construct histograms of all of the numeric variables, and comment on their shapes.

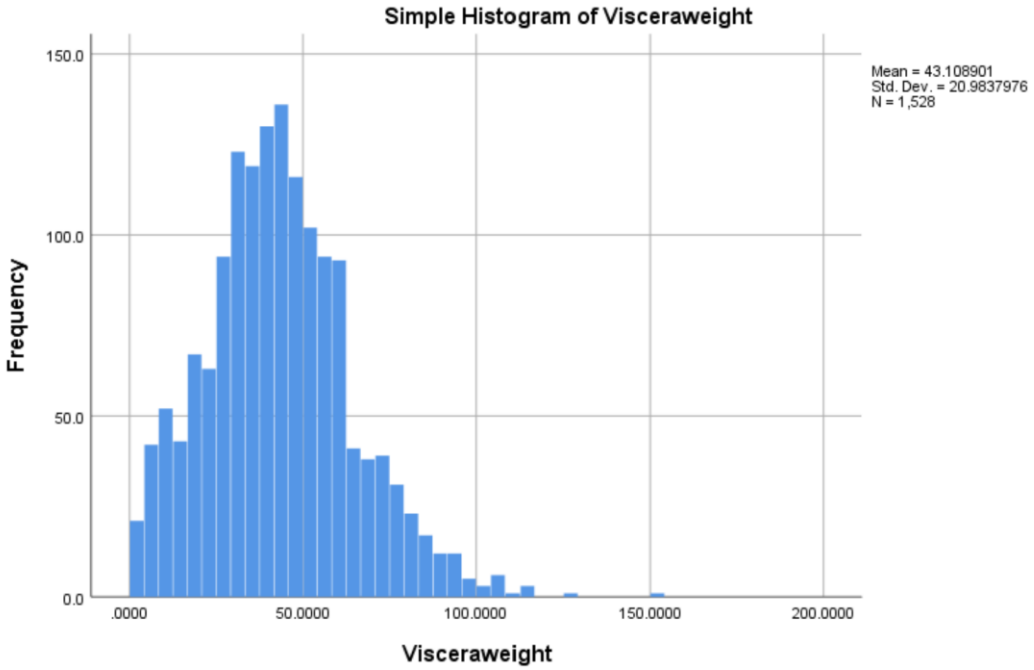


The histogram of Length seems to be slightly left skewed.

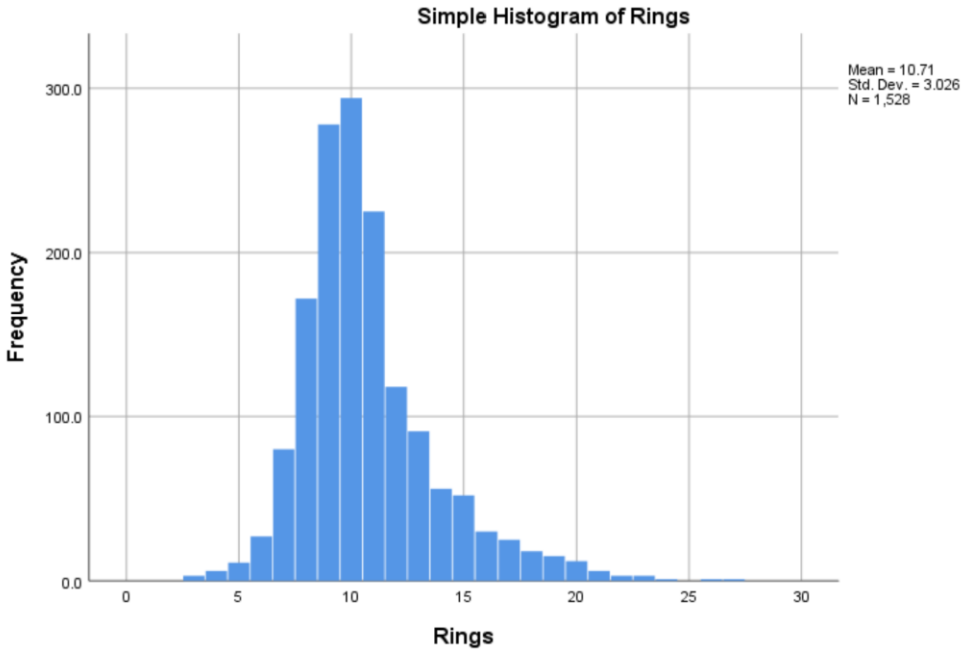


The histogram of Height seems to be symmetric.

Assignment 1
44468423
Gautham Meenakshisundaram



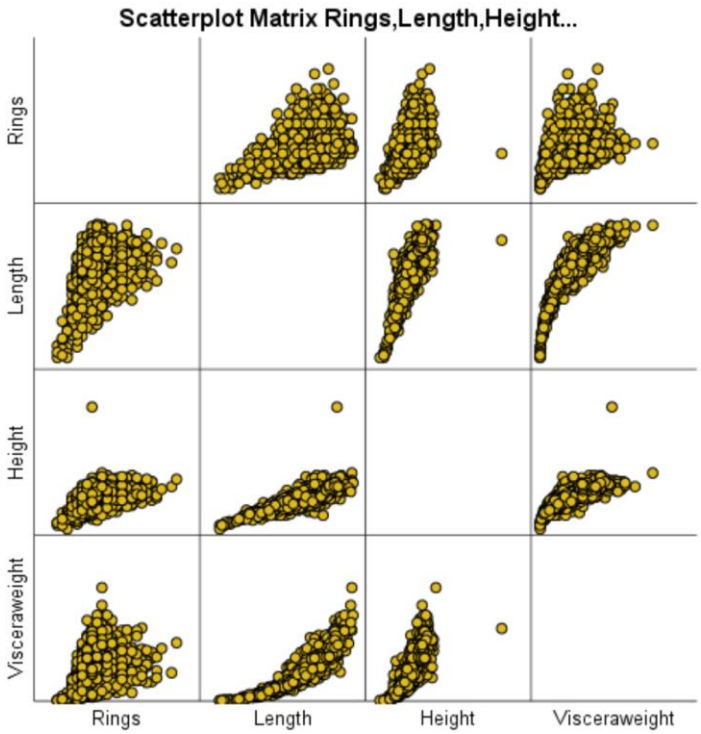
The histogram of Viscera weight seems to be right skewed with an outlier around 150 grams.



The histogram of Rings seems to be right skewed.

Assignment 1
44468423
Gautham Meenakshisundaram

b) Construct a scatterplot matrix, and comment on the relationships between the variables.



Rings ~ Length :
fanned out in the shape of a funnel, indicating non-constant variance and heteroscedasticity

Rings ~ Height :
one obvious outlier; fanned out in the shape of a funnel, indicating non-constant variance and heteroscedasticity

Rings ~ Viscera weight :
fanned out in the shape of a funnel, indicating non-constant variance and heteroscedasticity

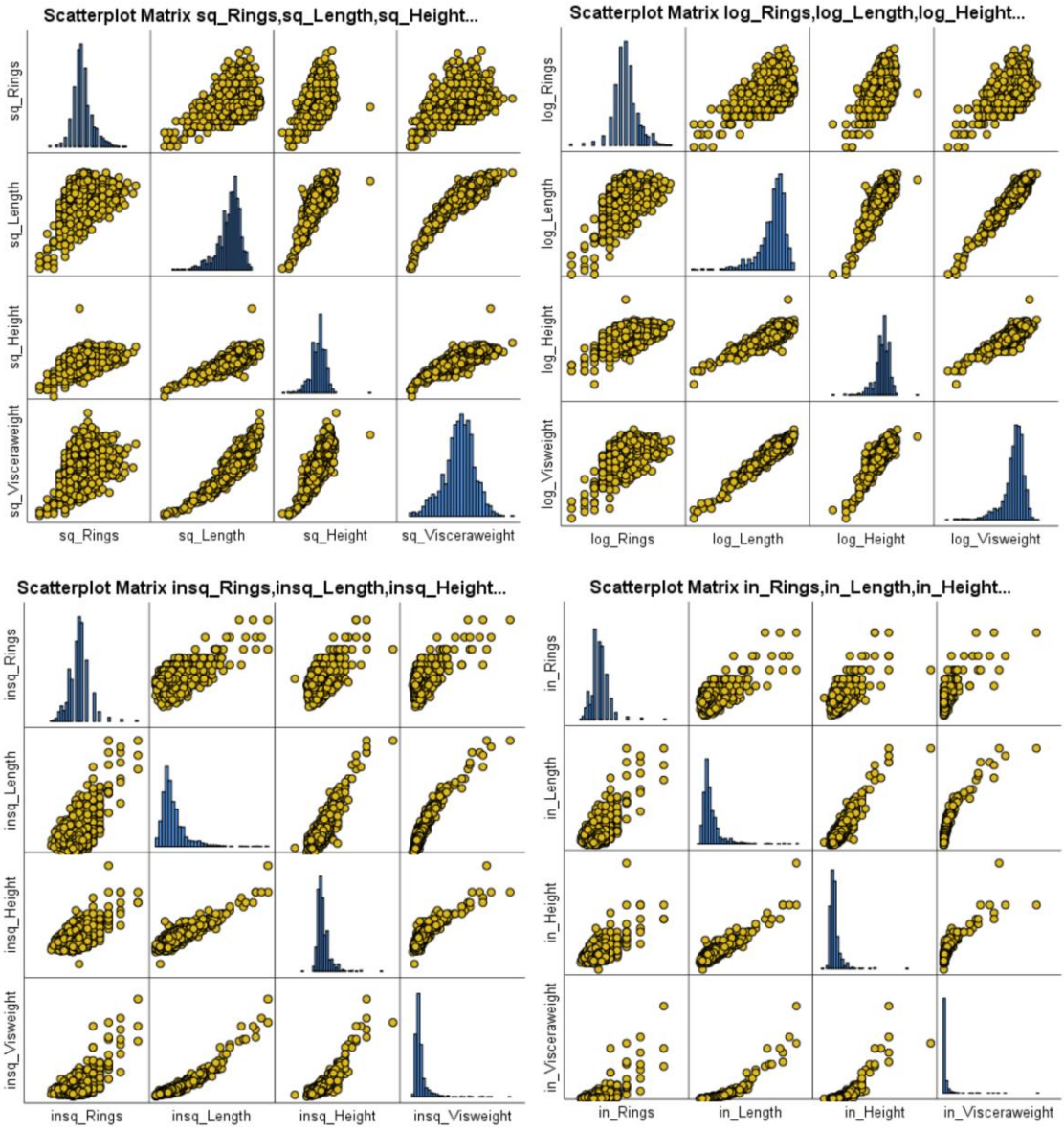
Length ~ Height :
one obvious outlier; moderate positive linear relationship

Length ~ Viscera weight:
one possible outlier; curvature indicating non-linear relationship

Height ~ Viscera weight:
couple of obvious outliers; curvature indicating non-linear relationship

Assignment 1
44468423
Gautham Meenakshisundaram

c) Using either SPSS or ARC, determine which transformation(s), if any, are needed in order to perform the regression.



A comparison of square root, log, inverse square root and inverse transformations on all the variables is shown above.

The variables in square root transformation graph has features of fanning out in the shape of a funnel indicating heteroscedasticity, and curvature indicating non-linear relationship. Moreover, the graph is similar to the scatterplot matrix shown previously without any transformations, indicating that square root transformation has not had much impact on the variables.

The variables in the log transformation graph show either moderate or strong positive linear relationship.

Most of the variables in inverse square root and inverse transformation graphs show non-constant variance.

To conclude, log transformation on all variables might be required to perform regression with this data.

d) Construct the linear model for ring with length, height and viscera weight as predictors (i.e. the three predictors in one model). Do not transform any variables. Perform significance tests for the three predictors, individually. Perform diagnostic checking on the model.

Model 1:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.437[a]	.191	.190	2.724

a. Predictors: (Constant), Viscera weight, Height, Length

b. Dependent Variable: Rings

R Square value of 0.191 represents that the model can explain only 19% of variation in the response variable rings out of the total variance explained.

It is to be noted that this model has a standard error of 2.724.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.373	.584		5.774	.000
	Length	.023	.008	.153	2.793	.005
	Height	.196	.019	.450	10.344	.000
	Viscera weight	-.026	.008	-.182	-3.472	.001

a. Dependent Variable: Rings

Significance test – Length:

Null Hypothesis H_0 : B_1 = significant
Alternate Hypothesis H_1 : B_1 = not significant

$t = (B_1 / S.e(B_1)) \sim t_{df=n-p-1}$

where

B_1 is the coefficient of Length at 0.023
$S.e(B_1)$ is the standard error of Length at 0.008
n is the number of observations at 1528
p is the number of predictors at 3
df is the degress of freedom at 1524

$t_{1524} = 2.793$

Since p-value of B_1 is 0.005 which is greater than the benchmark value 0.001, we reject the null hypothesis H_0 and conclude that B_1 is not significant.

Significance test – Height:

Null Hypothesis H_0 : B_2 = significant
Alternate Hypothesis H_1 : B_2 = not significant

$t = (B_2 / S.e(B_2)) \sim t_{df=n-p-1}$

where

B_2 is the coefficient of Height at 0.196
$S.e(B_2)$ is the standard error of Height at 0.019
n is the number of observations at 1528
p is the number of predictors at 3
df is the degress of freedom at 1524

$t_{1524} = 10.344$

Since p-value of B_2 is 0.000 which is lesser than or equal to the benchmark value 0.001, we do not reject the null hypothesis H_0 and conclude that B_2 is significant.

Significance test – Visceraweight:

Null Hypothesis H_0 : B_3 = significant
Alternate Hypothesis H_1 : B_3 = not significant

$t = (B_3 / S.e(B_3)) \sim t_{df=n-p-1}$

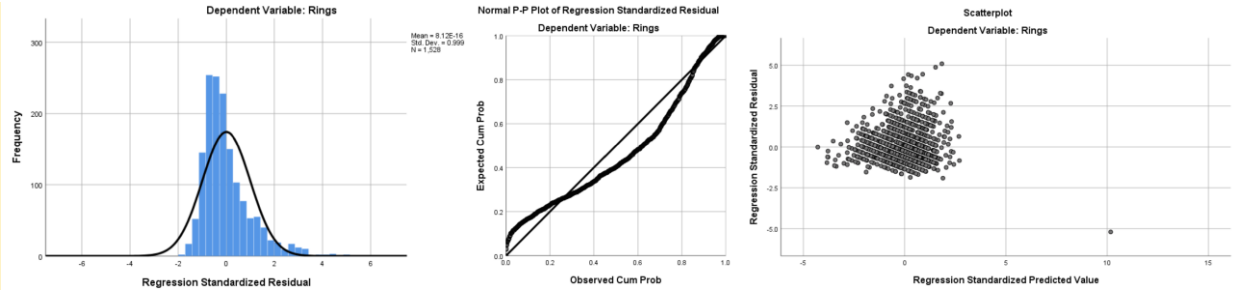
where

B_3 is the coefficient of Visceraweight at -0.026
$S.e(B_3)$ is the standard error of Visceraweight at 0.008
n is the number of observations at 1528
p is the number of predictors at 3
df is the degress of freedom at 1524

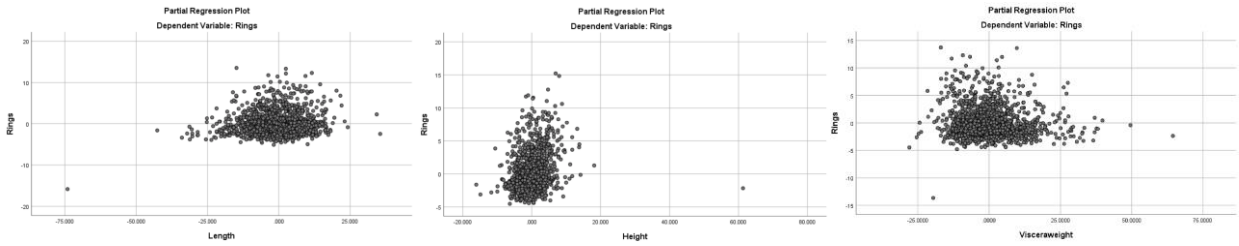
$t_{1524} = -3.472$

Since p-value of B_3 is 0.001 which is lesser than or equal to the benchmark value 0.001, we do not reject the null hypothesis H_0 and conclude that B_3 is significant.

Diagnostic Checking



The histogram seems to be right skewed.
The normal probability plot shows lot of curvature and is nowhere close to the linear line.
The fitted vs residuals plot shows residuals fanning out in the shape of a funnel, indicating heteroscedasticity.



All the partial regression plots show residuals fanning out in the shape of a funnel, indicating heteroscedasticity.

e) Based on your conclusions in previous parts of the question, construct a better model for abalone age (or rings), explaining what changes you have made. You should simplify the model by removing any non-significant predictors. Perform diagnostic checking on this model, and comment on the model fit. Write down carefully your final model.

As observed earlier, log transformation on all the variables is required to perform regression with this data.

Model 2:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.521[a]	.271	.270	.23025

a. Predictors: (Constant), log_Visweight, log_Height, log_Length

b. Dependent Variable: log_Rings

R Square value of 0.271 represents that the model can explain only 27% of variation in the response variable rings out of the total variance explained. The R Square value is better compared to previous model's 0.190.

It is to be noted that this model has a standard error of 0.23025 which is better compared to previous model's 2.724.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.223	.363		.615	.539
	log_Length	.103	.103	.082	1.002	.316
	log_Height	.491	.052	.468	9.360	.000
	log_Visweight	-.009	.034	-.023	-.268	.789

a. Dependent Variable: log_Rings

Significance test – log(Length):

Null Hypothesis H_0 : log(B_1) = significant
Alternate Hypothesis H_1 : log(B_1) = not significant

$t = (\log(B_1) / S.e(\log(B_1))) \sim t_{df=n-p-1}$

where

log(B_1) is the coefficient of log(Length) at 0.103
S.e(log(B_1)) is the standard error of log(Length) at 0.103
n is the number of observations at 1528
p is the number of predictors at 3
df is the degress of freedom at 1524

$t_{1524} = 1.002$

Since p-value of log(B_1) is 0.316 which is greater than the benchmark value of 0.001, we reject the null hypothesis H_0 and conclude that log(B_1) is not significant.

Significance test – log(Height):

Null Hypothesis H_0 : log(B_2) = significant
Alternate Hypothesis H_1 : log(B_2) = not significant

$t = (\log(B_2) / S.e(\log(B_2))) \sim t_{df=n-p-1}$

where

log(B_2) is the coefficient of log(Height) at 0.491
S.e(log(B_2)) is the standard error of log(Height) at 0.052
n is the number of observations at 1528
p is the number of predictors at 3
df is the degress of freedom at 1524

t_{1524} = 9.360

Since p-value of log(B_2) is 0.000 which is lesser than or equal to the benchmark value of 0.001, we do not reject the null hypothesis H_0 and conclude that log(B_2) is significant.

Significance test – log(Visceraweight):

Null Hypothesis H_0 : log(B_3) = significant
Alternate Hypothesis H_1 : log(B_3) = not significant

$t = (\log(B_3) / S.e(\log(B_3))) \sim t_{df=n-p-1}$

where

log(B_3) is the coefficient of log(Visceraweight) at -0.009
S.e(log(B_3)) is the standard error of log(Visceraweight) at 0.034
n is the number of observations at 1528
p is the number of predictors at 3
df is the degress of freedom at 1524

t_{1524} = -0.268

Since p-value of log(B_3) is 0.789 which is greater than the benchmark value of 0.001, we reject the null hypothesis H_0 and conclude that log(B_3) is not significant.

Model 3:

Since log(Length) and log(Visceraweight) are not significant, a linear regression model is fitted between log(Rings) and log(Height).

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.520[a]	.271	.270	.23025

a. Predictors: (Constant), log_Height

b. Dependent Variable: log_Rings

R Square value of 0.271 represents that the model can explain only 27% of variation in the response variable rings out of the total variance explained.

The R Square value is same compared to previous model, indicating that it is correct to remove log(Length) and log(Visceraweight) from the model, and that they do not explain any variation in the response variable Rings and are indeed not significant.

It is to be noted that this model has a standard error of 0.23025 which is same as previous model.

Assignment 1
44468423
Gautham Meenakshisundaram

Influential Observations:

There are no leverage values more than 0.03 which does not satisfy the required moderate leverage values of 0.2 – 0.5 and high leverage values of > 0.5.

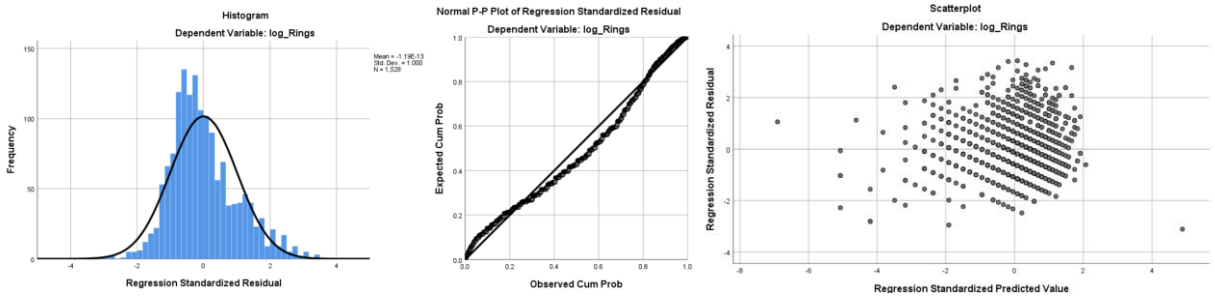
There are no Cook's distance values greater than neither the F statistic value of 565.849 nor the benchmark value of 1.

$DFBETA_{ij} > 2/sqrt(n)$
$DFBETA_{ij} > 2/sqrt(1528) = 0.05116$

There are quite a few standardized DFBETA values greater than the absolute value of 0.05116 and none of these values seems to be extreme.

To conclude, there doesn't seem to be any influential observations.

Diagnostic Checking:



The histogram seems to be symmetric and much better than the previous model.
The normal probability plot still shows a bit of curvature but is closer to the linear line.
The fitted vs residuals plot shows random scatter with couple of obvious outliers.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	.487	.078		6.257	.000
	log_Height	.546	.023	.520	23.788	.000

a. Dependent Variable: log_Rings

$\log y_i = 0.487 + 0.546 \log x_{1i}$

where for each ring i,

y_i = Rings of abalone
x_{1i} = Height of abalone

When height increases by one millimeter, the rings of abalone increases by 0.546.

Assignment 1
44468423
Gautham Meenakshisundaram

f) Give the predicted age, and 95% confidence interval for the prediction, of an abalone having length 50mm, height 40mm and viscera weight 100g.

$\log y_i = 0.487 + 0.546 \log x_{1i}$
$= 0.487 + 0.546 \log(40)$
$y_0 = 1.368$

We know that age can be arrived by adding 1.5 to rings and hence the predicted age is $y_0 = 2.9$.

The 95% confidence interval for the prediction = $y_0 +- z_{(1-a/2)} * S.e(y_o)$
where,

$y_0 = 2.9$
$z_{(1-a/2)} = 1.96$
$S.e(y_o) = 0.23025$

$= 2.9 +- (1.96*0.23025)$
$= 2.9 +- 0.45129$
$= (2.4, 3.4)$