Black-Lip Abalone Linear Discrimination Study of Sex Types

MSDS 6372-402

Group Project II

July 13, 2017

James Park

Mooyoung Lee

Purpose/Question

Determine if a Blacklip Abalone is Male, Female, or an Infant (M, F, or I) based on various dimensional and weight parameters. The data set utilized in this study is from the Machine Learning Repository from the University of California, Irvine (UCI). We would like to investigate if it is possible to differentiate (M, F, or I), via a MANOVA and LDA. Specifically to differentiate M and F so that food preparers may quickly identify in order to determine how to prepare. In addition we will differentiate I from M and F so that fisheries may quickly identify and throw back the Infant Abalone back in the water to further mature.

Background

There any many species of Abalone on cost lines across the world. Abalone are marine snails which are highly prized for their meat as well as their highly iridescent, nacre ('mother of pearl') shells used commonly in jewelry, buttons and inlay (figure 1). The flesh of abalones is widely considered to be a desirable food, and is consumed raw or cooked by a variety of cultures. Abalone mature to 9-10 centimeters in length in 3 to 6 years depending on the local conditions. Very few make it to maturity due to poaching and high demand.

Abalone shell

Abalone sashimi meat

Figure 1 (source: Wikipedia.org, wisegeek.com)

As in many fishing industries, demand is outstripping supply, and in order to maintain the species limits are placed by local authorities on the quantity and the size. In the 2016/2017 season in Australia an individual may obtain a 5 day (1 day per month over 5 months) fishing permit which allows up to 5 Abalone per day. The minimum size to keep is 13 centimeters in length, otherwise it must be thrown back.

The data investigated in this study is the Blacklip Abalone species found in the North Coast and Islands in the Bass Strait separating Tasmania from Australia (figure 2.)

Bass Strait between Tasmania and Australia	Blacklip Abalone

Figure 2, (Source Google maps and pics)

The Sex of an Abalone is determined by orienting the shell so that the meat is facing up, then moving the meat to the side exposing the gonads. An off white color indicates a male and a bluish color indicates a female (figure 3). This process is very manual and could be sped up if Abalones could be first filtered by the variables collected prior to identification. Abalone vary in size from 20 millimeters (0.79 in) (Haliotis pulcherrima) to 200 millimeters (7.9 in) while Haliotis rufescens is the largest of the genus at 12 inches (30 cm).

Abalone has been an important staple in native cultures around the world, specifically in Africa and on the North American West coast. The meat was used as food, and the shell was used as currency for many tribes. More recently Japan has steadily increased demand for Abalones.

Identification of an Abalone Sex

Figure 3 (source: fishconsult.org)

The Data

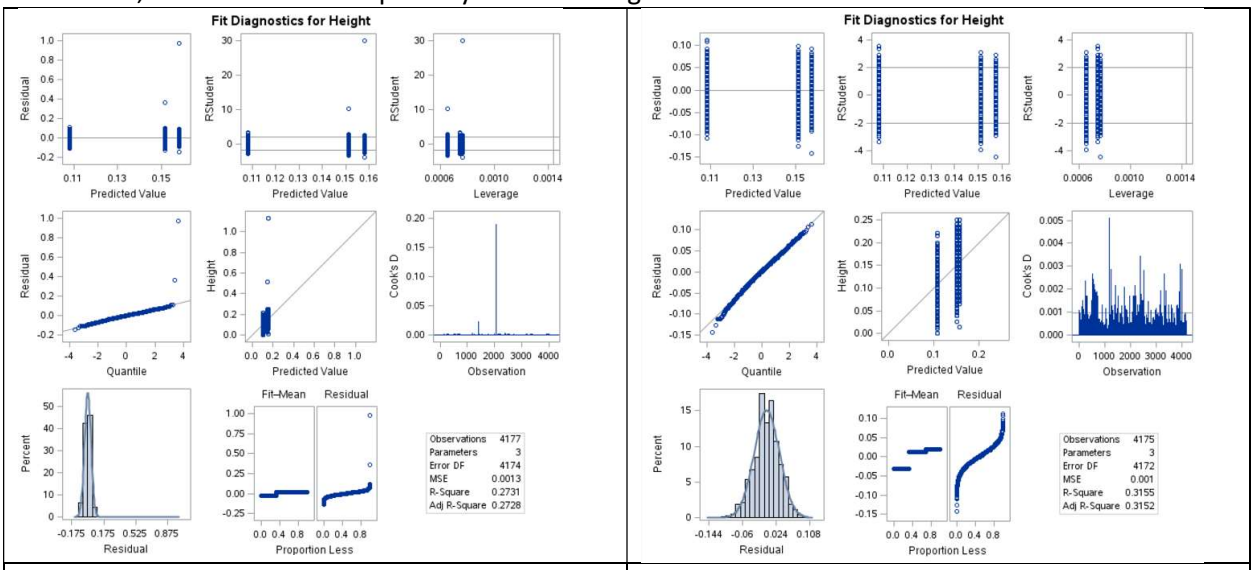
The Abalone data set and variable descriptions may be found in table 1. In addition to the variables in table 1, an additional variable "Volume" (Height*Diameter*Length) was added in an attempt to simplify the data set. The data set contains one categorical variable (Sex) with three levels (Male, Female, and Infant), one integer variable (number of Rings) and the remaining are continuous relating to shell dimensions and weight. One item of measurement should be noted here regarding the measurement scale of the Length, Diameter, Height and Weights. There were 4,177 samples in the dataset. The source of the data indicates the measurements are represented in millimeters and grams although ALL

other sources of Blacklip Abalone biology indicate the measurements should be on the order of centimeters and kilograms. Given the original data collection was in 1995, and for the purposes of this project assignment, we will assume that the dimensions have been scaled to an obscure multiplier to protect the proprietary nature of the original study. This will still serve the purposes of this project.

Name	Data Type	Meas.	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

Table 1 (source; uci.edu website)

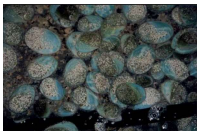
Two potential outliers were identified in early diagnostic plots. The SAS GLM procedure revealed the data shown in Figure x. The outliers are easily visualized in the Residual plots, RStudent plots, as well as the qq plots. In looking more closely at the raw data, for these two points, it was an obvious data collection error in the "Height" variable. This is most likely a typo or recording error as similar data was either an order of magnitude different or a large amount off (>50% compared to all other data). These two point were removed leaving the total number of samples reduced to 4,175. Original data, corrections, and raw data excerpts may be seen in Figure 4.



Original Data | Corrected Data

Outlier	ID	Sex	Length	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
Yes	1418	M	0.705	0.565	0.515	2.21	1.1075	0.4865	0.512	10
No	1419	M	0.705	0.555	0.215	2.141	1.0465	0.383	0.528	11
Yes	2052	F	0.455	0.355	1.13	0.594	0.332	0.116	0.1335	8
No	2053	M	0.46	0.345	0.12	0.4935	0.2435	0.1175	0.132	8

Figure 4



Data Visualization (initial plots)

Prior to any deep data exploration the data was first visually explored on a high level to gain perspective and to provide a basic understanding of relative magnitudes and distributions. Exploration techniques included (See Appendix A for full results and code):

- Means Output: the calculation of means, minimums, maximums, range and standard deviations via a SAS means procedure (see Appendix A)
- Frequency and percent of Sex (M, F, I) via SAS frequency procedure. Frequencies are important in conducting and Linear Discriminant Analysis (LDA) so that the prior proportions may be taken into account. Frequency data is shown below in Figure 5. The data is relatively evenly distributed meaning about a third of the studied population is in each group.

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	1306	31.28	1306	31.28
I	1342	32.14	2648	63.43
M	1527	36.57	4175	100.00

Figure 5, frequency of Sex levels

- Boxplot of each variable grouped by Sex via SAS univariate procedure. As an example the "Length" variable by Sex (F, I, M) is shown below in Figure 6. The "Length" variable was typical of all other variables with respect to boxplot output, meaning the Infant was easily differentiated visually from the Male and Female samples and the Male and Female appeared fairly close. The remaining box-plots and associated code may be found in Appendix A.

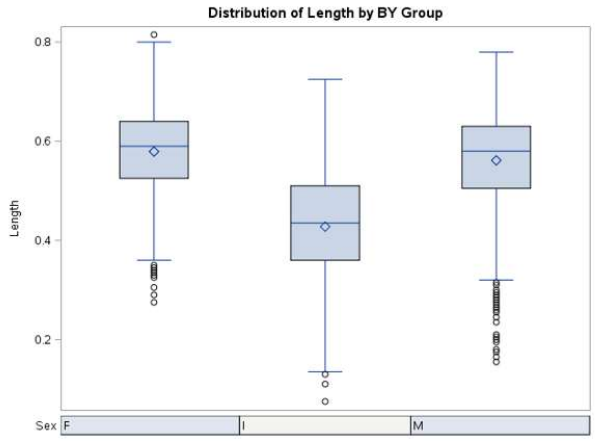


Figure 6, boxplot of example variable (Length) by Sex

- Scatter and Histogram plots of all continuous and integer variables involved via SAS sgscatter procedure were also viewed. As one can see in Figure 7, there is a fair amount of association of variables. Although this should not be an issue in conducting an LDA further residual examination will be explored in order meet LDA assumptions of normality of residuals.

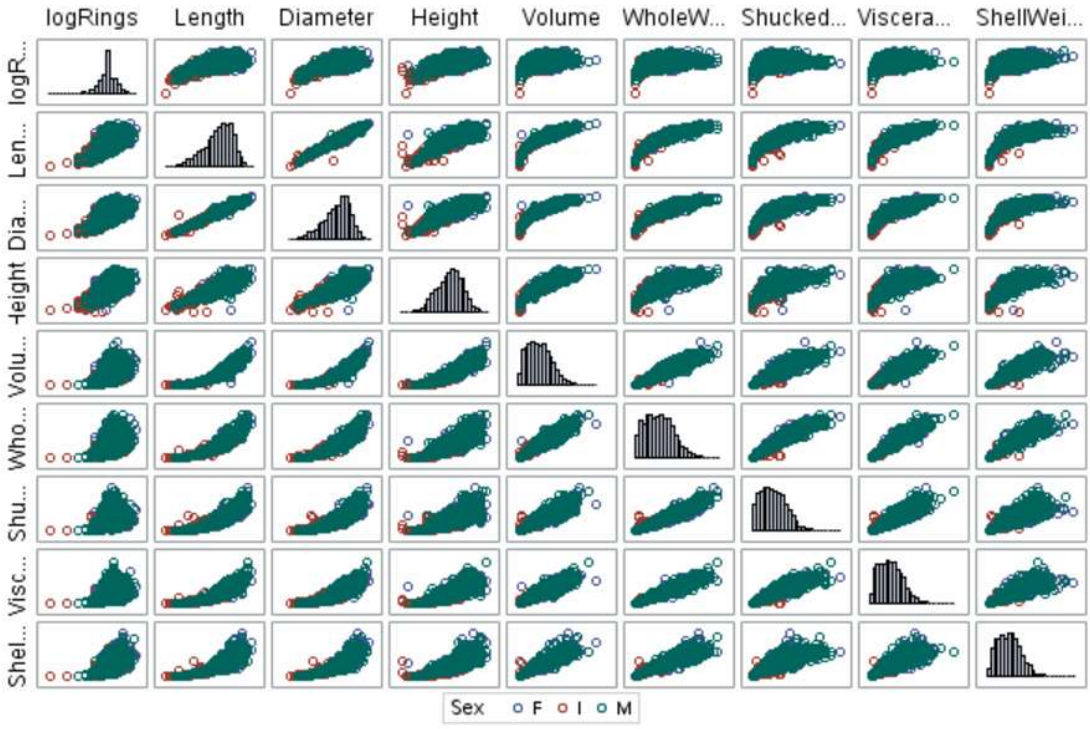
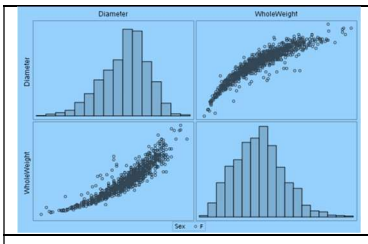
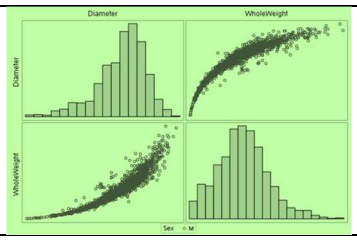
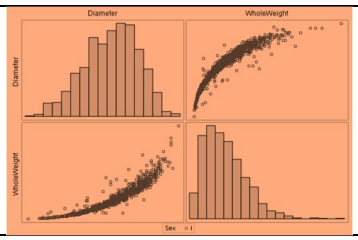


Figure 7, initial scatter plot

- Additionally scatter plots were more closely examined by Sex in order to visually examine if equal variance/covariance looked similar as part of the investigation if homogeneity is maintained. As an example in, in figure 8, the "Wholeweight" variable was compared to the "diameter" variable via scatter-plots and histograms by sex. The visual output would indicate that the variance/covariance may be similar. This was the case for all other variables investigated. While the full number of combinations is exhaustive several others and the associated SAS code used to generate are included in Appendix A.





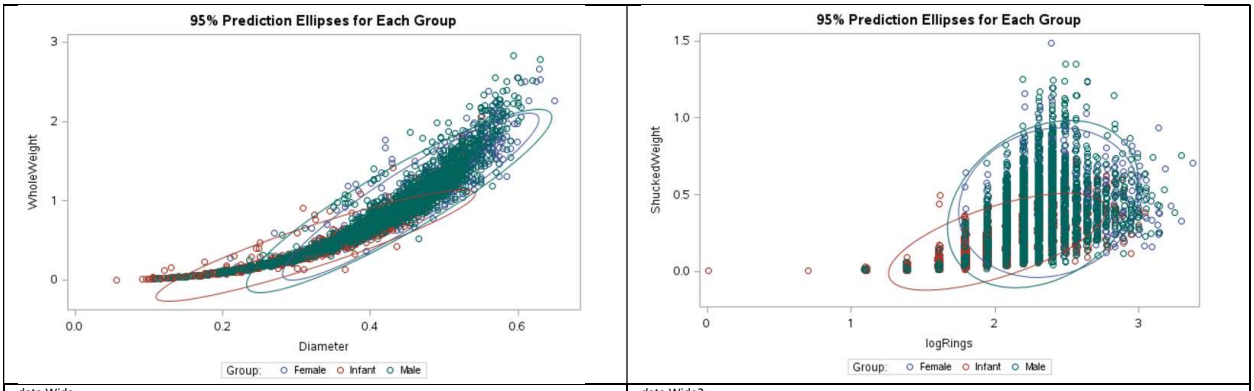


Female	Male	Infant
Similar variation per visuals		

Figure 8, variable comparison by Sex

Next, 95% ellipse prediction plots were generated for variable pairs in order to assess potential differences in covaraince matrices. Variables selected and the associated prediction ellipses may be viewed in figure 9. The ellipses indicate a difference in the covariance matrix between the Infant and other Sex categories (M and F). They also indicate similar shape and orientation between the M and F groups. Several variable pairs were randomly selected and they all exhibited similar trends. In the

interest of saving space and moving on in this study the other 95% prediction ellipses are not provided in this report.




```
data Wide;
/*   |-- Diameter --| |--- WholeWeight ---|  */
keep L_AA L_BB L_CC W_AA W_BB W_CC;  /* names of new variables */
merge abalone2(where=(Sex="F")
          rename=(Diameter=L_AA WholeWeight=W_AA))
    abalone2(where=(Sex="I")
          rename=(Diameter=L_BB WholeWeight=W_BB))
    abalone2(where=(Sex="M")
          rename=(Diameter=L_CC WholeWeight=W_CC));
run;

title "95% Prediction Ellipses for Each Group";
proc sgplot data= Wide;
 scatter x=L_AA  y=W_AA  / jitter name="AA"  legendlabel="Female";
 scatter x=L_BB y=W_BB / jitter name="BB" legendlabel="Infant";
 scatter x=L_CC y=W_CC / jitter name="CC" legendlabel="Male";
 ellipse x=L_AA  y=W_AA  / lineattrs=GraphData1;
 ellipse x=L_BB y=W_BB / lineattrs=GraphData2;
 ellipse x=L_CC y=W_CC / lineattrs=GraphData3;

 keylegend "AA" "BB" "CC" / title="Group:";
 xaxis label="Diameter";
 yaxis label="WholeWeight";
run;
```

```
data Wide2;
/*   |-- logRings --| |--- ShuckedWeight ---|  */
keep X1 X2 X3 L_DD Y1 Y2 Y3 W_DD;  /* names of new variables */
merge abalone2(where=(Sex="F")
          rename=(logRings=X1 ShuckedWeight=Y1))
    abalone2(where=(Sex="I")
          rename=(logRings=X2 ShuckedWeight=Y2))
    abalone2(where=(Sex="M")
          rename=(logRings=X3 ShuckedWeight=Y3));
run;

title "95% Prediction Ellipses for Each Group";
proc sgplot data= Wide2;
 scatter x=X1  y=Y1  / jitter name="1"  legendlabel="Female";
 scatter x=X2 y=Y2 / jitter name="2" legendlabel="Infant";
 scatter x=X3 y=Y3 / jitter name="3" legendlabel="Male";
 ellipse x=X1  y=Y1  / lineattrs=GraphData1;
 ellipse x=X2 y=Y2 / lineattrs=GraphData2;
 ellipse x=X3 y=Y3 / lineattrs=GraphData3;

 keylegend "1" "2" "3" / title="Group:";
 xaxis label="logRings";
 yaxis label="ShuckedWeight";
```

Figure 9, 95% prediction ellipses

Assumptions (MANOVA/LDA)

Although visual indications were insightful in understanding the variables additional investigation was conducted to determine if the assumptions for a MANOVA/LDA were met. In principle the MANOVA will test whether or not the mean vectors of the three response levels are common and to what level. These investigations are outlined below and include rationale and some selected graphics output with associated code from SAS. The full SAS output and code reference may be found in Appendix B. The assumptions investigated include normality, variance/covariance, independence, and common means vector.

Normality: Virgin, log-transformed, square-transformed, and mean-center-transformed data were all investigated in order to attempt to further improve the normality of the residual data. The SAS GLM procedure was used to generate individual ANOVAs in order to generate fit diagnostics plots for each variable. This was executed for virgin, log-transformed, mean-center transformed as well as mean-center transformed data. Primarily the residuals, histogram and qq plots were utilized to determine normality. Given the large sample size and the central limit theorem we will proceed with the raw data with the exception of the "Rings" variable which remains log transformed for the remainder of this study. The "Rings" variable output (pre and post log) is shown in figure 10, while several other investigations may be viewed in Appendix B. The improvement from "Rings" to "LogRings" is primarily

evident in the residual plots, qq plots and in the adjusted R-squared. For all other variables the qq-plots seemed to invert causing an equally "non-normal" condition. Mean-Center and Square transformations did not yield improvement to the normality of the residuals. In many cases the same inversion occurred on the qq plots or the fit simply did not improve. These efforts and associated code may also be viewed in Appendix B.

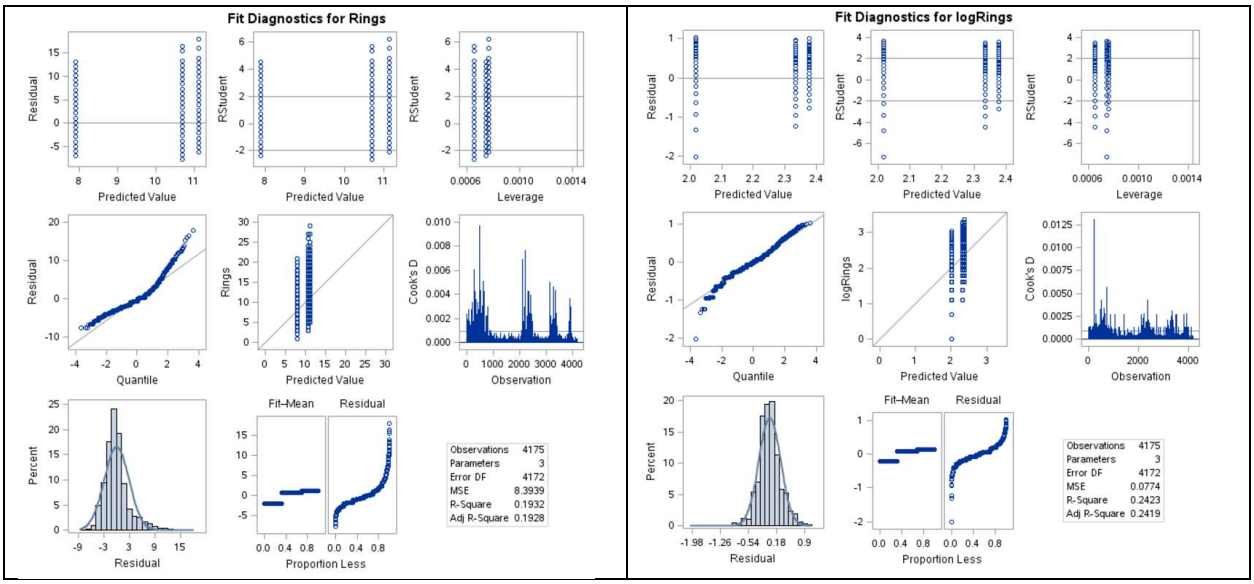


Figure 10, transformation example for Rings variable

Variance/Covariance: Once the state of variable transformation was determined the next assumption to be verified was the variance/covariance assumption. Prior 95% prediction ellipses indicated a difference in the Infant level to the two other levels (Male and Female). It also indicated similar variance/covariance between the Male and Female groups. The SAS code and output noted in figure 11 provide evidence that variance/covariance matrices may not be the same as indicated by the p-value on the Chi-Square test of less than 0.0001. Although the matrices cannot be confirmed to be the same (even though visual indications indicate they may be), this does not rule out the use of an LDA as SAS will use a quadratic, versus liner, method which may assume different matrices.

```
proc discrim data=abalone2 pool=test;
  class sex;
  var logRings Length Diameter Volume Height WholeWeight ShuckedWeight VisceraWeight ShellWeight; run;
```

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
2465.893308	20	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

There is a difference among the covariance matrices. Also noted from SAS output is an indication that the within covariance matrices will be used in the discriminate function.

Figure 11, Variance/Covariance equality test

Independence: Since there were separate shells measured and none were repeated we will assume the independence assumption is made. All samples were gathered from the Bass Strait in an area north or Tasmania. Therefore the independence is limited to this region of data collection and may not be

necessarily extended outside of this region. For purposes of this study, all data is assumed to be randomly collected as there is no provided rationale to assume that they are not.

Although variance/covariance are not confirmed common we will proceed with MANOVA F-Tests, individual ANOVA's and LDA to further understand any differences between the Sex group mean vectors.

MANOVA

A MANOVA was run to understand where the differences in the Sex group mean vectors. This knowledge of differences may be used to understand contributors in the LDA model. All of the test criteria noted in the SAS output in figure 12 indicate a rejected Ho (all p-values are <.0001) that the mean vectors are the same. This means that there is difference in the Sex levels per the p-values.

```
/* MANOVA after Log Transformation */
Proc GLM Data=abalone2 plots = all;
 Class Sex;
 Model logRings Length Diameter Height Volume WholeWeight
ShuckedWeight VisceraWeight ShellWeight = Sex;
 Manova H=_All_ / PrintE PrintH Canonical;
Run;
```

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Sex Effect
H = Type III SSCP Matrix for Sex
E = Error SSCP Matrix

S=2 M=3 N=2081

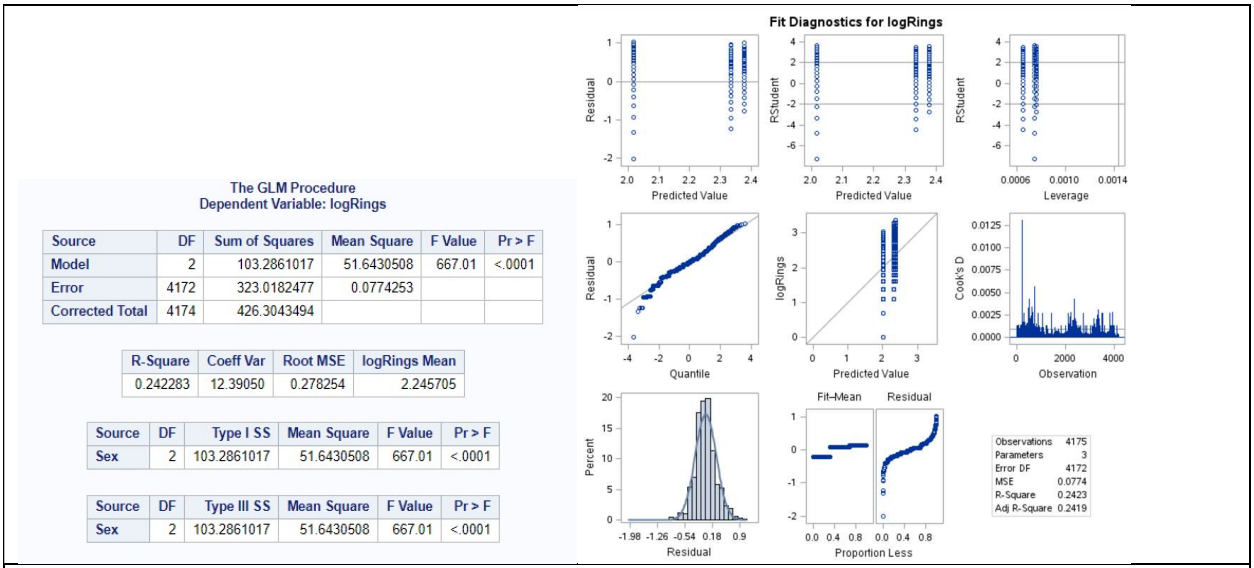
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.61800157	125.87	18	8328	<.0001
Pillai's Trace	0.38633119	110.79	18	8330	<.0001
Hotelling-Lawley Trace	0.61110793	141.34	18	6936.2	<.0001
Roy's Greatest Root	0.59941158	277.39	9	4165	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

Figure 12, MANOVA comparing Sex level mean vectors. There is a difference.

Since there is a difference in mean vectors we proceed to several one-way ANOVA F-Tests to find which variables are potentially driving the differences. The SAS code and sample output may be seen in figure 13 (the SAS code and remaining output may be viewed in Appendix C). Per the output, all variables are significantly and individually influencing (p-value <0.0001) the difference in response (Sex) levels.



Example output of "logRings" variable significance to Sex level

Figure 13, ANOVA F-Test example

Based on the MANOVA and individual ANOVA tests there must be a group mean difference so the individual mean test is performed using LSMEANS procedure. The SAS code and output may be viewed in figure 14. The Bonferroni adjustment is used in order to match Type-I error rate as 0.05 with multiple pair wise tests on a single set of data. In this test, every pair-wise test for every dependent variable shows that the differences are highly significant except one case. For the ShuckedWeight variable, Sex level 1 and 3 (which is Male and Female) shows no significant difference (p-value 0.1601). There is, however, a significant difference (p-value <0.0001) between the Infant and others (Male and Female).

```
Proc GLM Data=abalone2 plots = all plots=all;
 Class Sex;
 Model logRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight ShellWeight = Sex;
 Manova H=_All_ / PrintE PrintH Canonical;
 lsmeans Sex / adjust=bon cl; Run;
```

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: logRings

i/j	1	2	3
1		<.0001	0.0002
2	<.0001		<.0001
3	0.0002	<.0001	

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Length

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		<.0001
3	<.0001	<.0001	

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Diameter

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		<.0001
3	<.0001	<.0001	

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Height

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		<.0001
3	<.0001	<.0001	

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Volume

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		<.0001
3	<.0001	<.0001	

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: WholeWeight

i/j	1	2	3
1		<.0001	0.0007
2	<.0001		<.0001
3	0.0007	<.0001	

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: ShuckedWeight

i/j	1	2	3
1		<.0001	0.1601
2	<.0001		<.0001
3	0.1601	<.0001	

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: VisceraWeight

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		<.0001
3	<.0001	<.0001	

Least Squares Means for effect Sex
Pr > |t| for H0: LSMean(i)=LSMean(j)

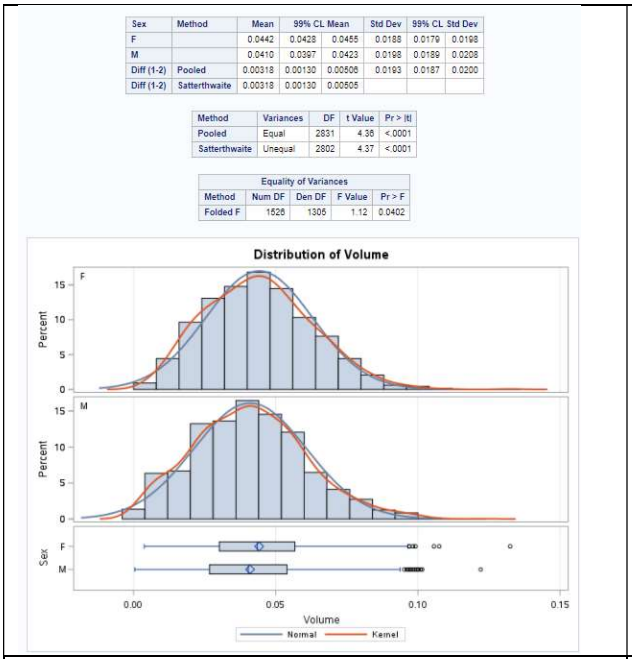
Dependent Variable: ShellWeight

i/j	1	2	3
1		<.0001	<.0001
2	<.0001		<.0001
3	<.0001	<.0001	

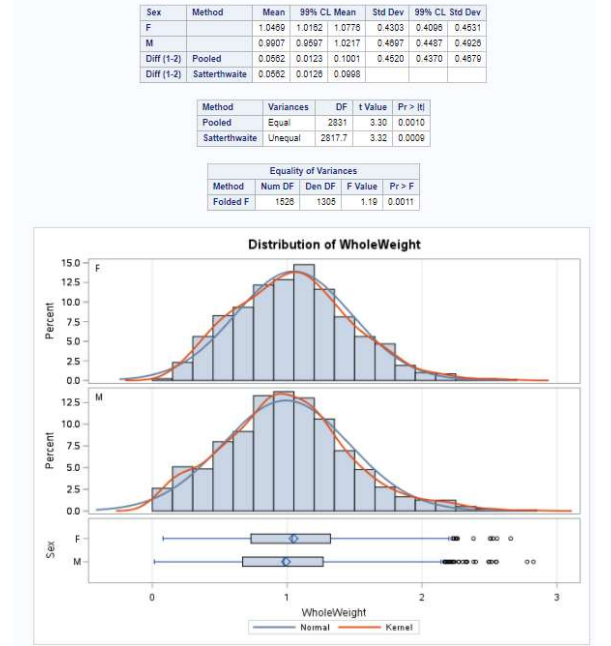
Figure 14, SAS code and output for LSMEANS testing.

Based on the significance of the MANOVA, ANOVA, and LSMEANS tests and the indications in the visual inspection of variables we will use all the original variables in the initial LDA including Volume and log transformed “Rings”.

Purely out of academic interest and to help visualize practical versus statistical significance between the Male and Female levels we also decided to perform two simple t-tests. The output may be viewed in figure 15. One t-test compares the “Volume” between the Male and Female levels. The second t-test compares the “Wholeweight” between the Male and Female levels. In both examples, the difference in means is apparent and statistically significant. However, there is significant overlap of the two histograms and boxplots indicating the difference may not be practical. For example, one could gather Wholeweight measurements that fall in the overlapping area and not be able to discern whether it came from the Male or Female. We will take this into consideration in looking at the LDA results.



t-test (Male Volume compared to Female Volume)



t-test (Male Wholeweight compared to Female Wholeweight)

Figure 15, t-test for practical significance understanding

Method for LDA

Based on the above visual and statistical examinations we will proceed with the Linear Discriminant Analysis (LDA) and cross-validation to determine if Male and Female Abalones may be distinguished and if the Infant Abalones may be distinguished by collecting measurements consistent with the data set described earlier in this paper. This Baseline data set will include the original data including "Volume" and a log transformed "Rings" variable. Once the LDA is run we will report the error rates and interpret the results with respect to the questions of interest. The LDA was run using the SAS DISCRIM procedure. The code and output may be seen in figure 16. The Baseline LDA indicates an overall error rate of 46.11%, meaning that it is correct, per the cross-validation, 53.89% of the time in the Bass Strait region of Ocean for the black-lip Abalone. As the t-test visuals and previous box-plots indicate there may be statistical significance in the mean vectors, however the large overlaps cause a practical concern where many of the variables collected are in the overlap region and may easily fall into the 46.11% error side of the discrimination.

```
PROC DISCRIM data = abalone2 CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=test wcov pcov;
CLASS sex;
VAR logRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight ShellWeight;
PRIORS EQUAL;
run;
quit;
```

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.ABALONE2
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Sex

From Sex	F	I	M	Total
F	871 66.69	328 25.11	107 8.19	1306 100.00
I	164 12.22	1147 85.47	31 2.31	1342 100.00
M	916 59.99	466 30.52	145 9.50	1527 100.00
Total	1951 46.73	1941 46.49	283 6.78	4175 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Sex

	F	I	M	Total
Rate	0.3331	0.1453	0.9050	0.4611
Priors	0.3333	0.3333	0.3333	

46.11% Error Rate – Baseline variable set

Figure 16, LDA and percent error

Let's look a little closer at the questions of interest.

Can we discriminate Male from Female Black-Lip Abalones? The error of discriminating an Abalone in the Female group is 33.31% and in the Male group is 90.5%, meaning that the Female discrimination is correct 66.69% of the time and Male discrimination is correct only 9.5% of the time. This method of discriminating Male and Female Black-Lip Abalones in the Bass Strait region of ocean would not be very reliable or practical. In fact, the number of measurements, even if collected autonomously, would probably not rival the more manual method of checking the gonad color.

Can we discriminate Infant from Male and Female Black-Lip Abalones? The error of discriminating an Abalone in the Infant group is 14.53%, meaning that it is correct, per the cross-validation, 85.47% of the time in the Bass Strait region of ocean for the black-lip Abalone. This may be very useful to fisherman attempting to discriminate if they should throw the Abalone back or keep it. In fact, this is what

fisherman do (take a measurement and decide). Further investigation may likely reveal that the error occurs in Infants that are near the size and weight of the matured Male or Female Abalones. In this case many may simply be checked by eye and thrown back without question, thus speeding up the process and improving profits through eliminating non-value added and time consuming measurements. In the event the price of Abalones continues to rise automation may further improve this process. This level of discrimination is useful with respect to the original question posed.

Auto-selection LDA

For academic purposes and in order to determine if a simplified LDA model could discriminate Male/Female as well or better with a reduced data set, the data set was reduced to only include Male and Female data entries and then the SAS STEPDISC procedure was utilized to auto-select variables for use in a revised LDA. The thought is to attempt to build either an improved model (less error) or a much simpler approach to future data collection utilizing fewer variables. For reference STEPDISC selection functions in the following manner (ref: support.sas.com website):

- selection methods include forward selection, backward elimination, and stepwise selection
- variables are chosen to enter or leave the model according to one of two criteria:
  - the significance level of an *F* test from an analysis of covariance, where the variables already chosen act as covariates and the variable under consideration is the dependent variable
  - the squared partial correlation for predicting the variable under consideration from the CLASS variable, controlling for the effects of the variables already selected for the model

The STEPDISC procedure resulted in a much smaller subset of variables (Diameter, Height, ShuckedWeight, and VisceraWeight) listed in the SAS output in figure 17.

```
Code reference for stepwise selection and results.
*automatic variable selection using stepdisc for LDA variable selection;
proc stepdisc data=abaloneMFonly bsscp tsscp;
  class Sex;
  var LogRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight ShellWeight; run;
```

The STEPDISC Procedure
Stepwise Selection: Step 5

Statistics for Removal, DF = 1, 2828

Variable	Partial R-Square	F Value	Pr > F
Diameter	0.0077	21.82	<.0001
Height	0.0008	2.29	0.1300
ShuckedWeight	0.0142	40.67	<.0001
VisceraWeight	0.0017	4.94	0.0264

No variables can be removed.

Figure 17, Stepwise auto-selection of LDA variables. Final step shown.

The percent error rate comparing the original LDA to the auto-selected LDA may be viewed in figure 18. The error rate was degraded by 0.4% (from 46.11% to 46.51%) while the number of variables was reduced from ten down to the four auto-selected by SAS.

```
PROC DISCRIM data = abalone2 CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=test wcov pcov;
CLASS sex;
VAR logRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight ShellWeight;
PRIORS EQUAL;
run;
quit;
```

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.ABALONE2
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Sex

From Sex	F	I	M	Total
F	871 66.69	328 25.11	107 8.19	1306 100.00
I	164 12.22	1147 85.47	31 2.31	1342 100.00
M	916 59.99	466 30.52	145 9.50	1527 100.00
Total	1951 46.73	1941 46.49	283 6.78	4175 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Sex

	F	I	M	Total
Rate	0.3331	0.1453	0.9050	0.4611
Priors	0.3333	0.3333	0.3333	

46.11% error rate using all variables

```
*rerun LDA using auto-selected variables;
PROC DISCRIM data = abaloneMFonly CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=test wcov pcov;
CLASS sex;
VAR Diameter Height ShuckedWeight VisceraWeight;
PRIORS EQUAL;run;quit;
```

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.ABALONE
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Sex

From Sex	F	I	M	Total
F	756 57.89	360 27.57	190 14.55	1306 100.00
I	151 11.25	1150 85.69	41 3.06	1342 100.00
M	782 51.21	487 31.89	258 16.90	1527 100.00
Total	1689 40.46	1997 47.83	489 11.71	4175 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Sex

	F	I	M	Total
Rate	0.4211	0.1431	0.8310	0.4651
Priors	0.3333	0.3333	0.3333	

46.51% error rate using fewer variables via auto-selection procedure in SAS (STEPDISC)

Figure 18, Comparison of error rate in original LDA to auto-selected LDA variables

Conclusions

We initially set out to determine if Blacklip Abalones could be discriminated by Infant versus Male and Female and also if they could be discriminated between Male and Female based on several measurement and weight variables. Through a series of both visual and statistical methods it has been determined that the Infant may be discriminated 85.69% (14.31% error) of the time using a Cross-validated and auto-selected data set and SAS's DISCRIM LDA procedure. This means that the four auto-selected measurements would differentiate the Infant the majority of the time and could be useful, especially if the measurement and sorting was automated and occurred upon collection. The Male and Female discrimination was more problematic, meaning the error rates ranged between 33.31% and 90.5% for the base and auto-selected models attempted. Although the MANOVA and one-way ANOVA's indicated statistical significance, the difference was simply not practical as all of the one-way ANOVA's exhibited significant overlap areas. So, a measurement found in this area could potentially be either a Male or Female. Auto-selection techniques did not change the high combined error rates. Our recommendation is to simply continue to utilize current methods to determine the gender of the Blacklip Abalone.

Appendix A Data Visualization and Assumptions

Variable Summary Stats

```
proc means data =  abalone n mean max min range std fw=8;
var _numeric_ ;
output out = meansout mean = mean std = std;
title 'Summary Stat';
run;
```

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
4	Diameter	Num	8	BEST12.	BEST32.
5	Height	Num	8	BEST12.	BEST32.
1	ID	Num	8	BEST12.	BEST32.
3	Length	Num	8	BEST12.	BEST32.
11	Rings	Num	8	BEST12.	BEST32.
2	Sex	Char	1	$1.	$1.
10	ShellWeight	Num	8	BEST12.	BEST32.
8	ShuckedWeight	Num	8	BEST12.	BEST32.
9	VisceraWeight	Num	8	BEST12.	BEST32.
6	Volume	Num	8	BEST12.	BEST32.
7	WholeWeight	Num	8	BEST12.	BEST32.

The MEANS Procedure

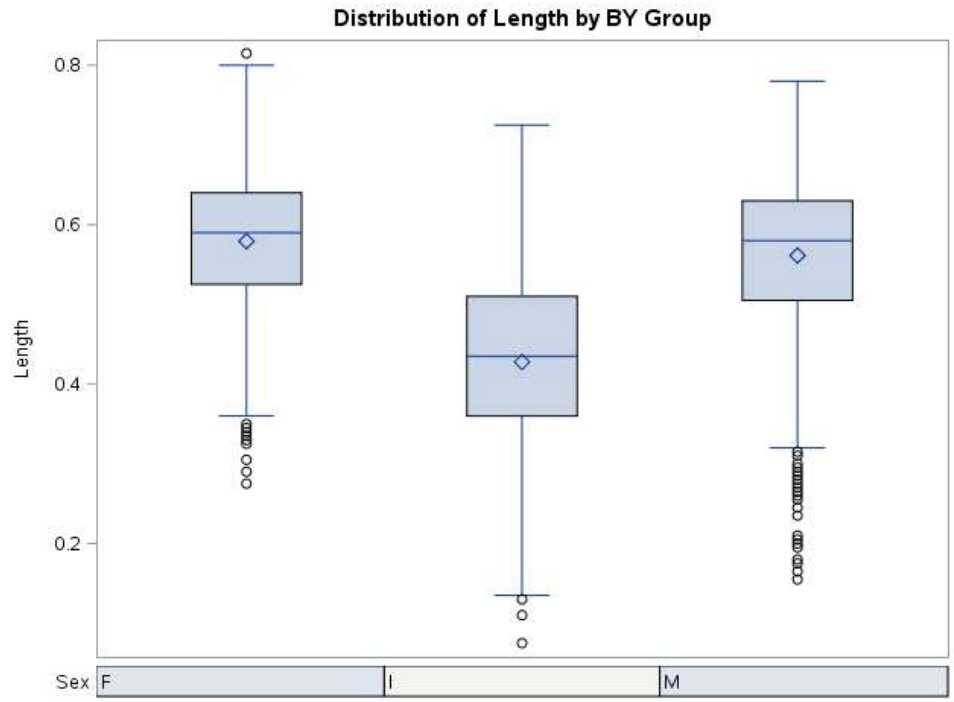
Variable	N	Mean	Maximum	Minimum	Range	Std Dev
ID	4175	2089.2	4177.0	1.0000	4176.0	1206.2
Length	4175	0.5240	0.8150	0.0750	0.7400	0.1201
Diameter	4175	0.4079	0.6500	0.0550	0.5950	0.0992
Height	4175	0.1392	0.2500	0	0.2500	0.0385
Volume	4175	0.0346	0.1324	0	0.1324	0.0209
WholeWeight	4175	0.8285	2.8255	0.00200	2.8235	0.4900
ShuckedWeight	4175	0.3592	1.4880	0.00100	1.4870	0.2217
VisceraWeight	4175	0.1805	0.7600	0.000500	0.7595	0.1095
ShellWeight	4175	0.2388	1.0050	0.00150	1.0035	0.1392
Rings	4175	9.9341	29.0000	1.0000	28.0000	3.2248

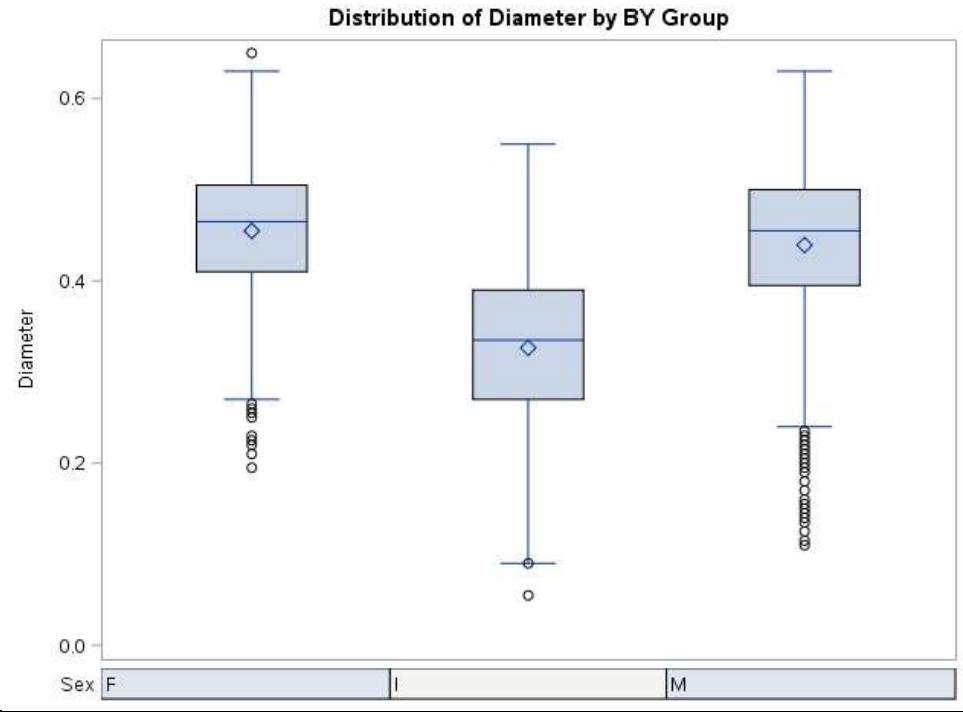
```
proc freq data = abalone;
run;
```

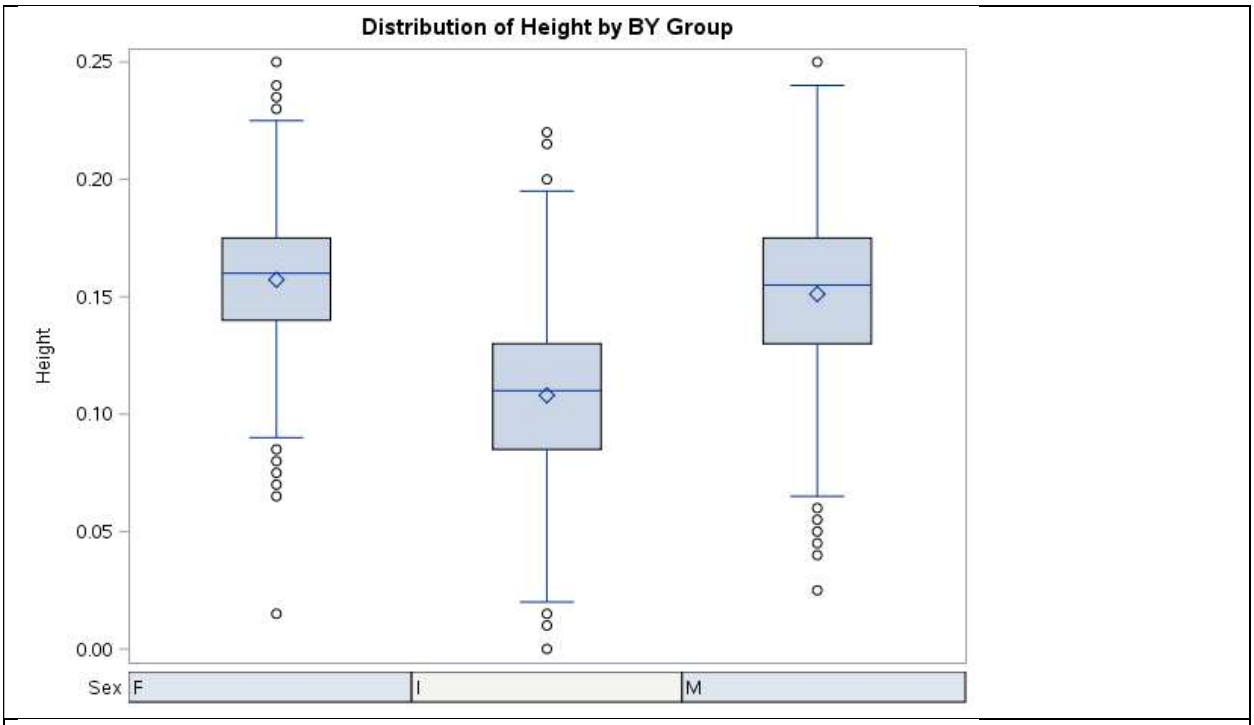
Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	1306	31.28	1306	31.28
I	1342	32.14	2648	63.43
M	1527	36.57	4175	100.00

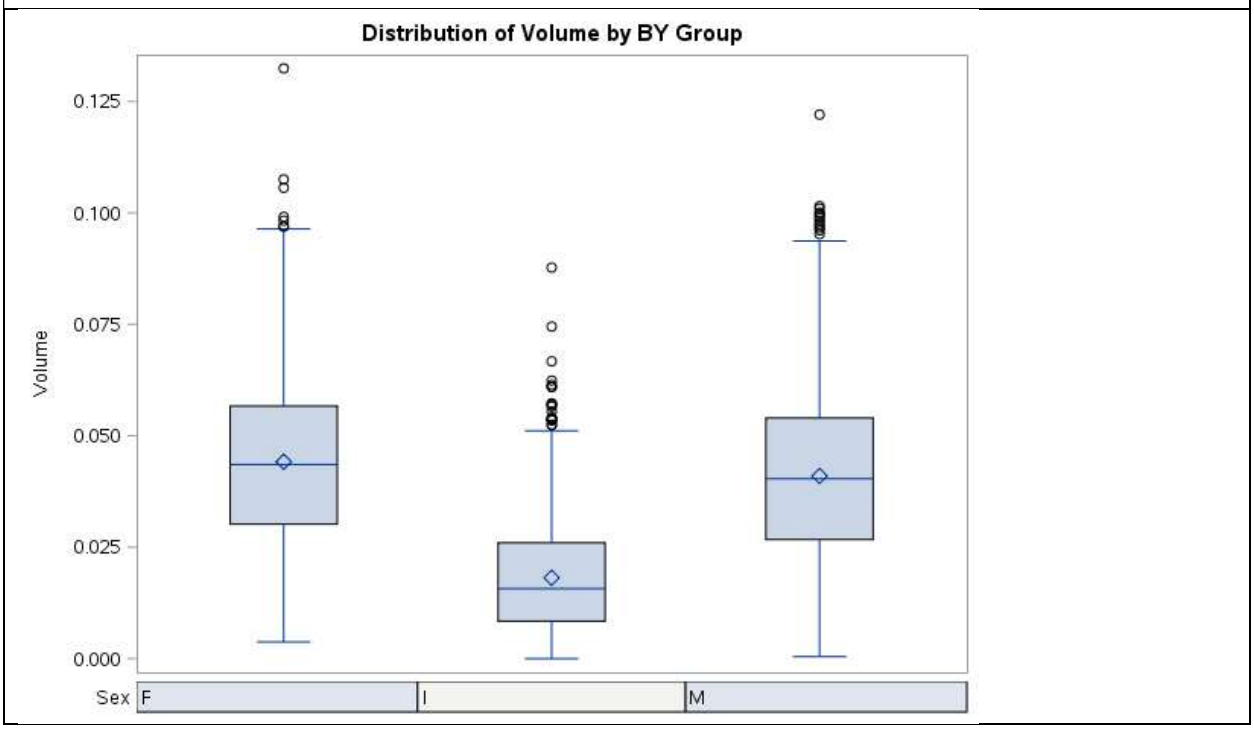
Box Plot by Sex

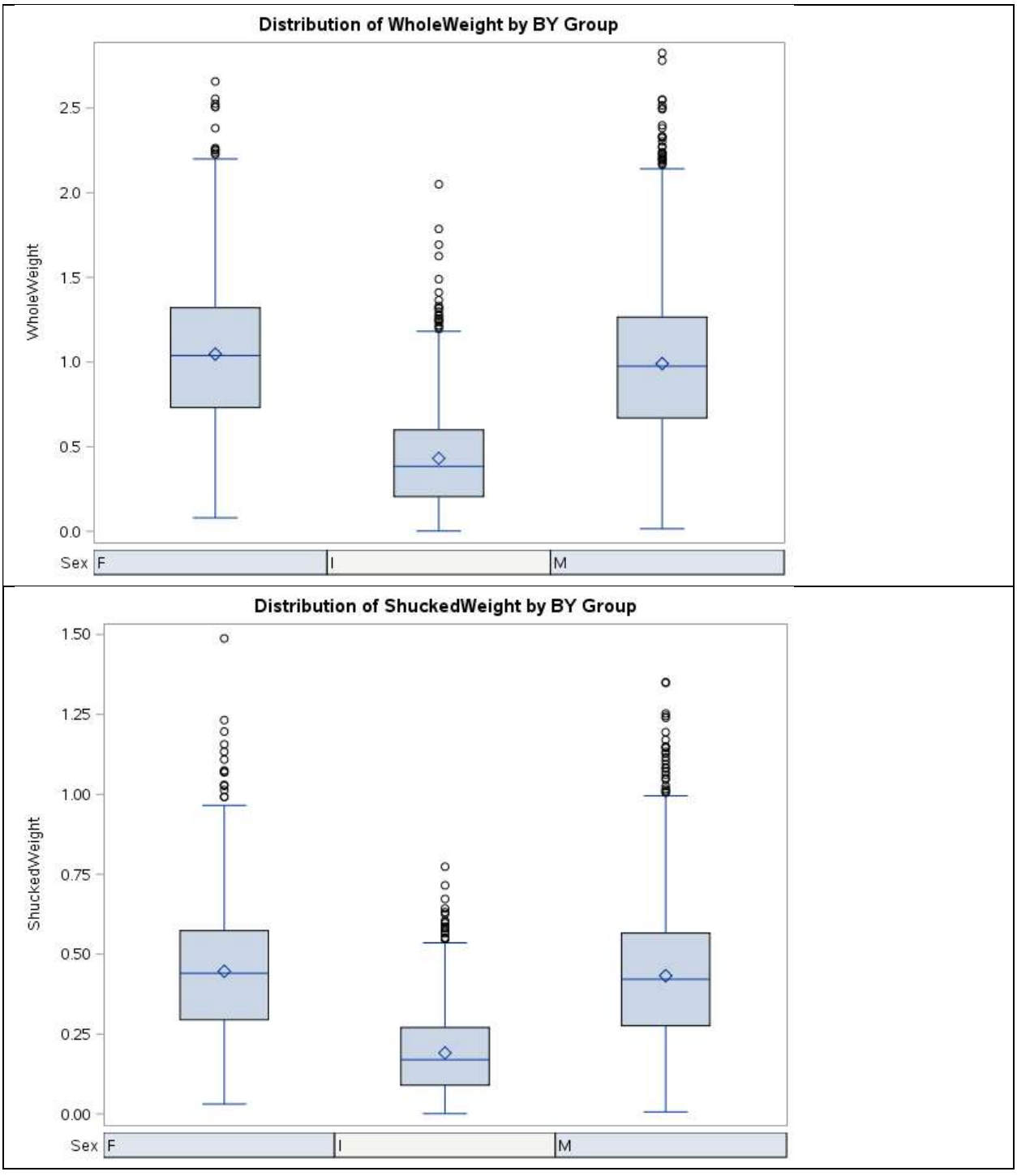
```
proc univariate data = abalone2 plots;
by Sex;
run;
```

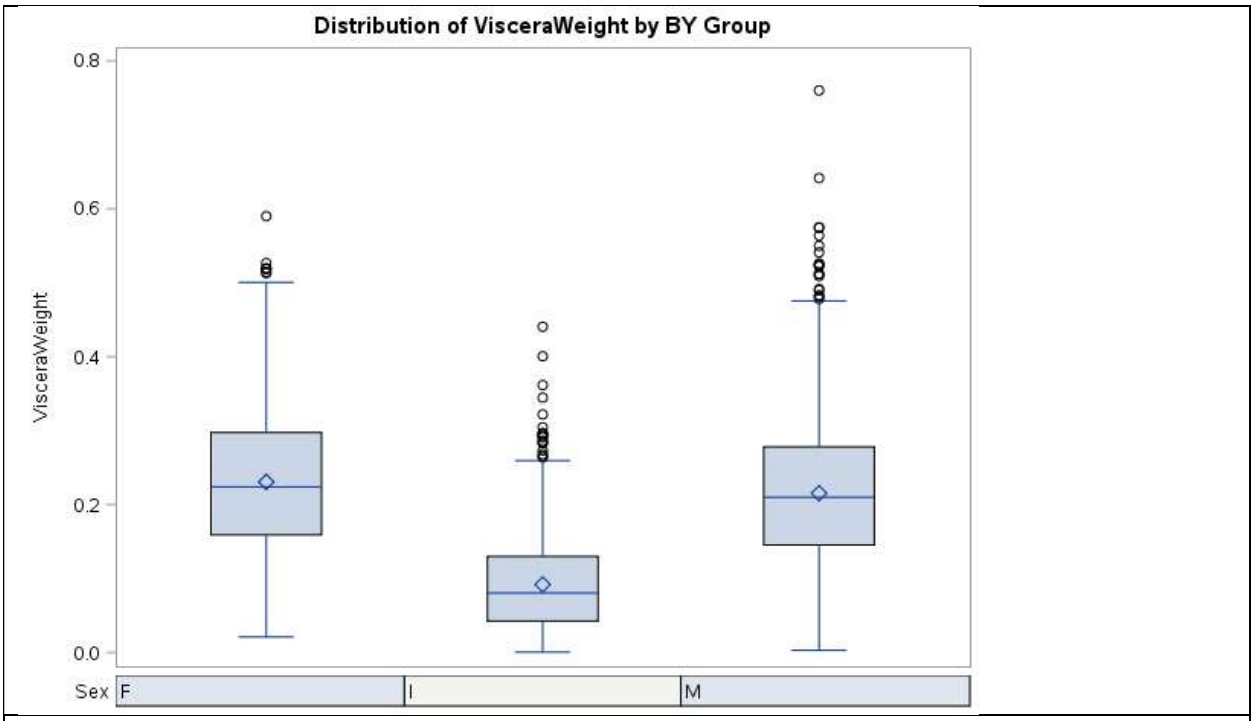


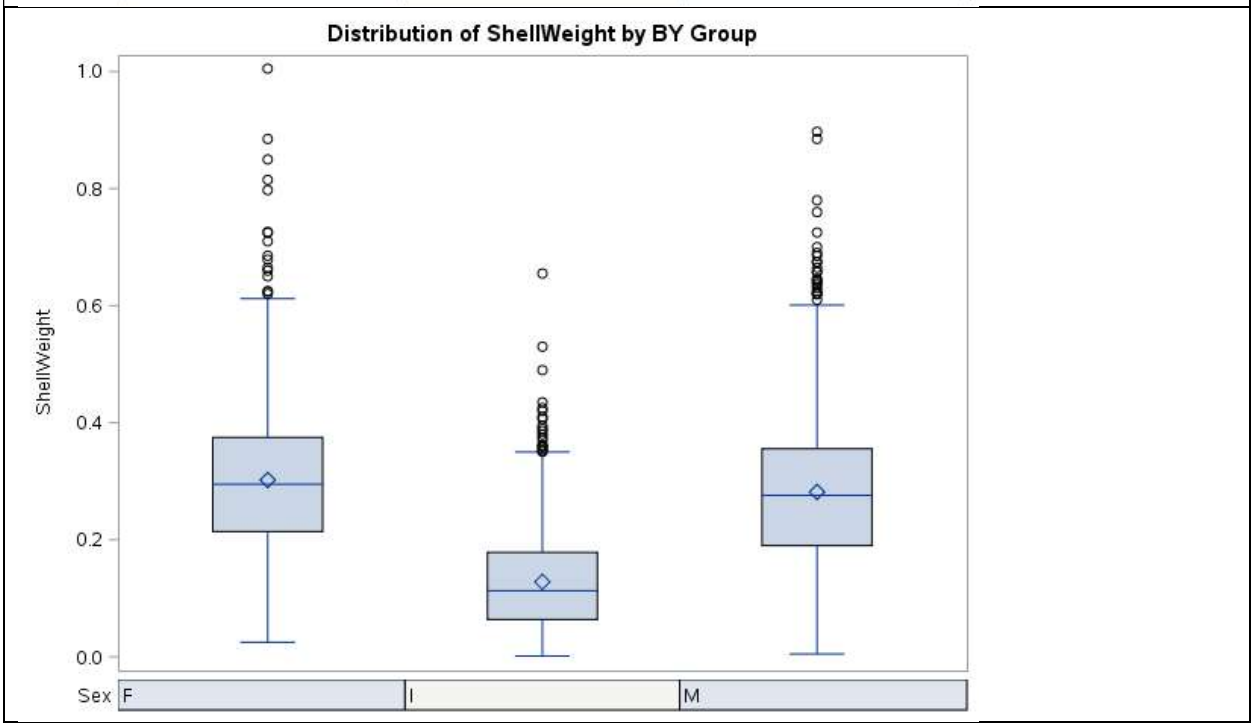


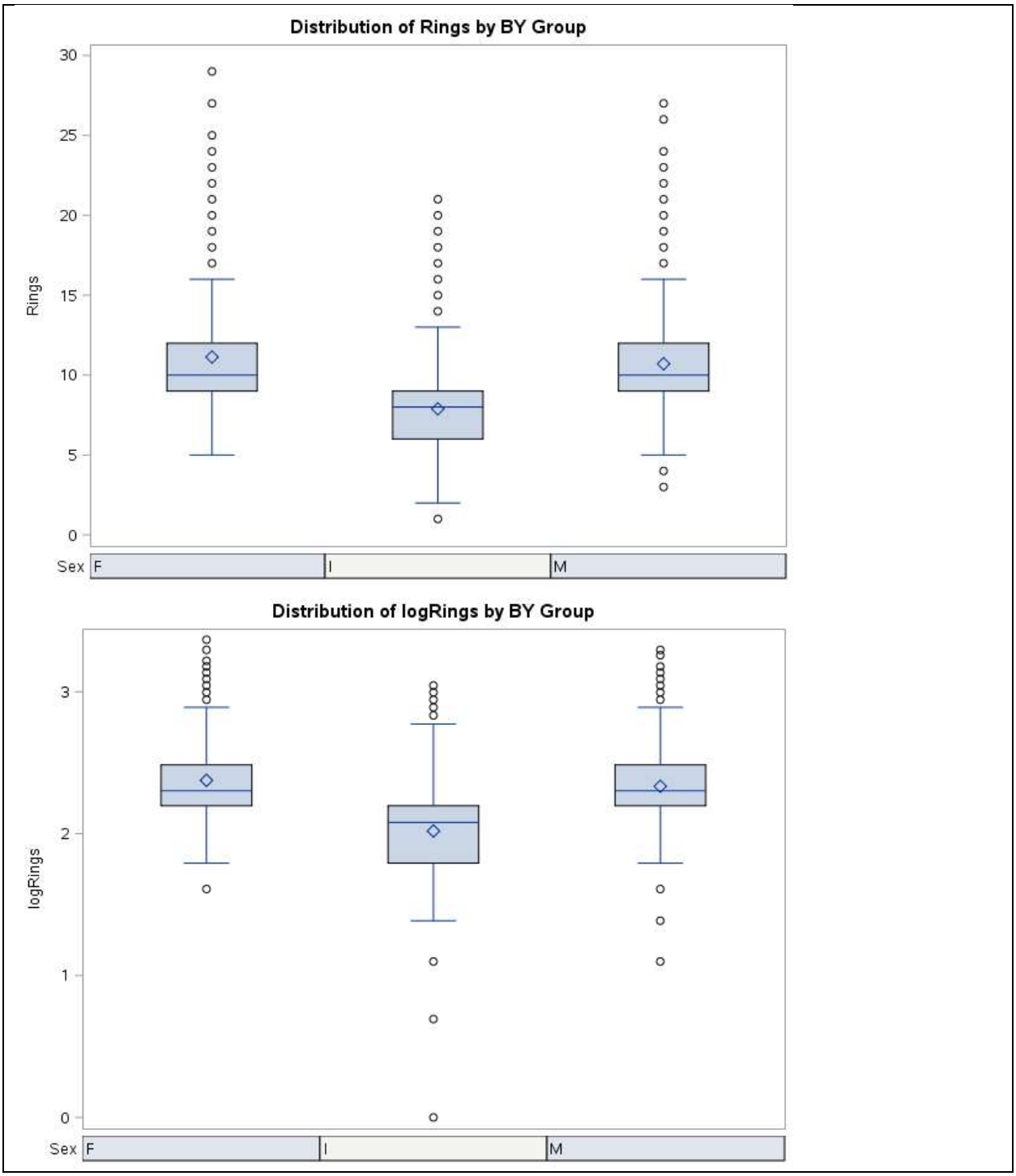

Distribution of Height by BY Group
Height
0.25
0.20
0.15
0.10
0.05
0.00
Sex
F
I
M


Distribution of Volume by BY Group
Volume
0.125
0.100
0.075
0.050
0.025
0.000
Sex
F
I
M


Distribution of WholeWeight by BY Group
WholeWeight
2.5
2.0
1.5
1.0
0.5
0.0
Sex
F
I
M
Distribution of ShuckedWeight by BY Group
ShuckedWeight
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Sex
F
I
M

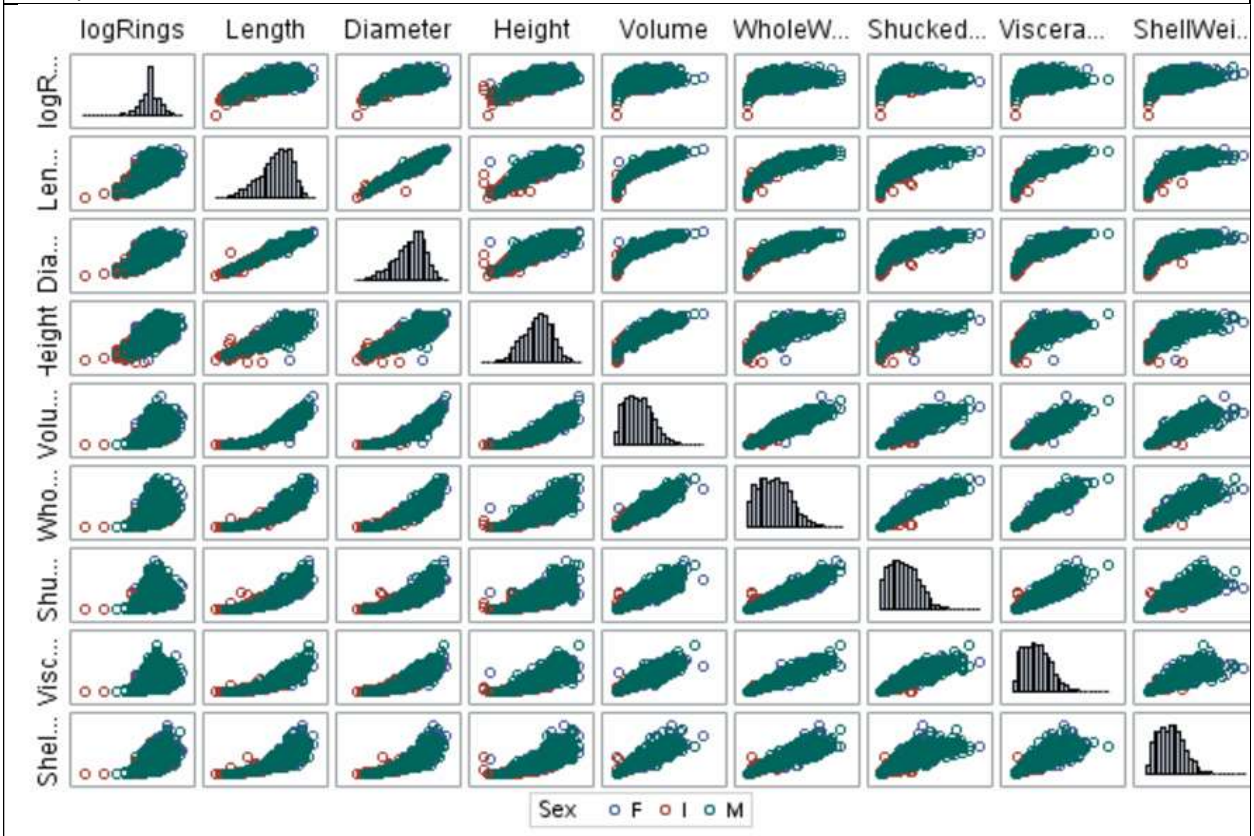

Distribution of VisceraWeight by BY Group
VisceraWeight
0.8
0.6
0.4
0.2
0.0
Sex
F
I
M


Distribution of ShellWeight by BY Group
ShellWeight
1.0
0.8
0.6
0.4
0.2
0.0
Sex
F
I
M


Distribution of Rings by BY Group
Rings
30
25
20
15
10
5
0
Sex
F
I
M
Distribution of logRings by BY Group
logRings
3
2
1
0
Sex
F
I
M

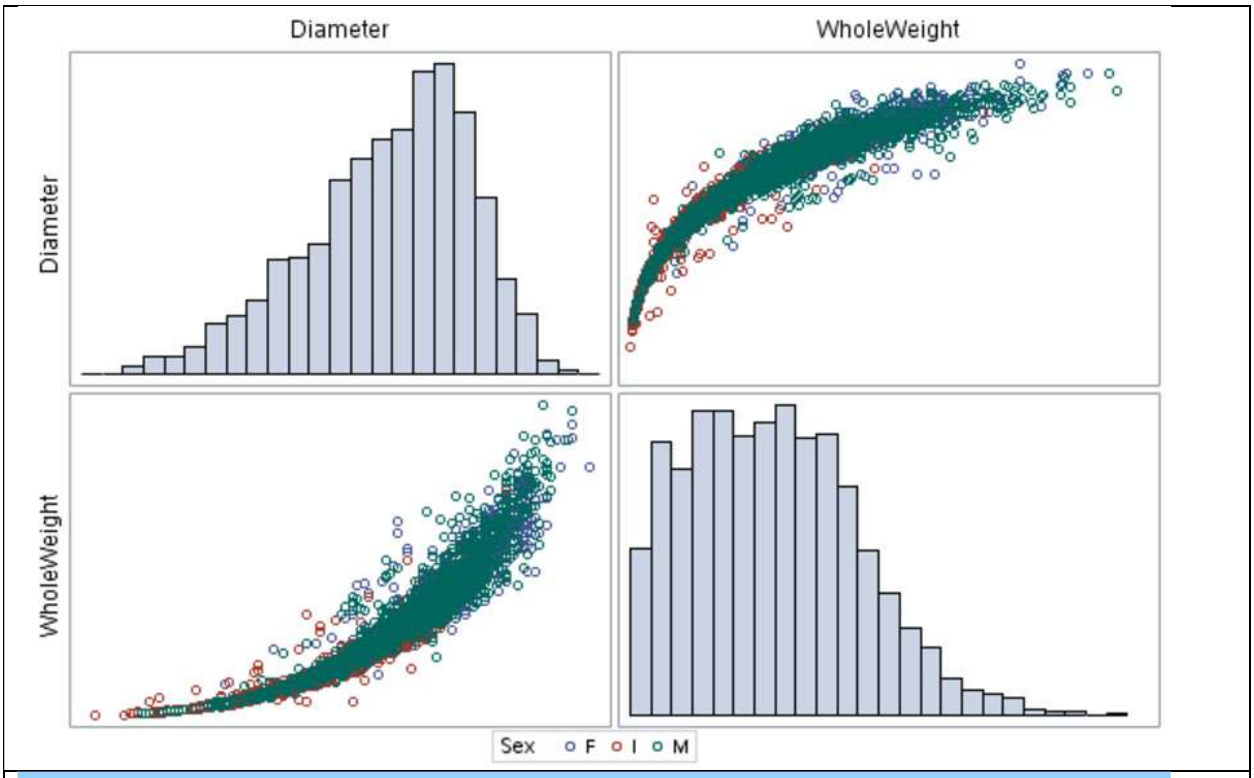
Scatter plot by Sex: All Variables

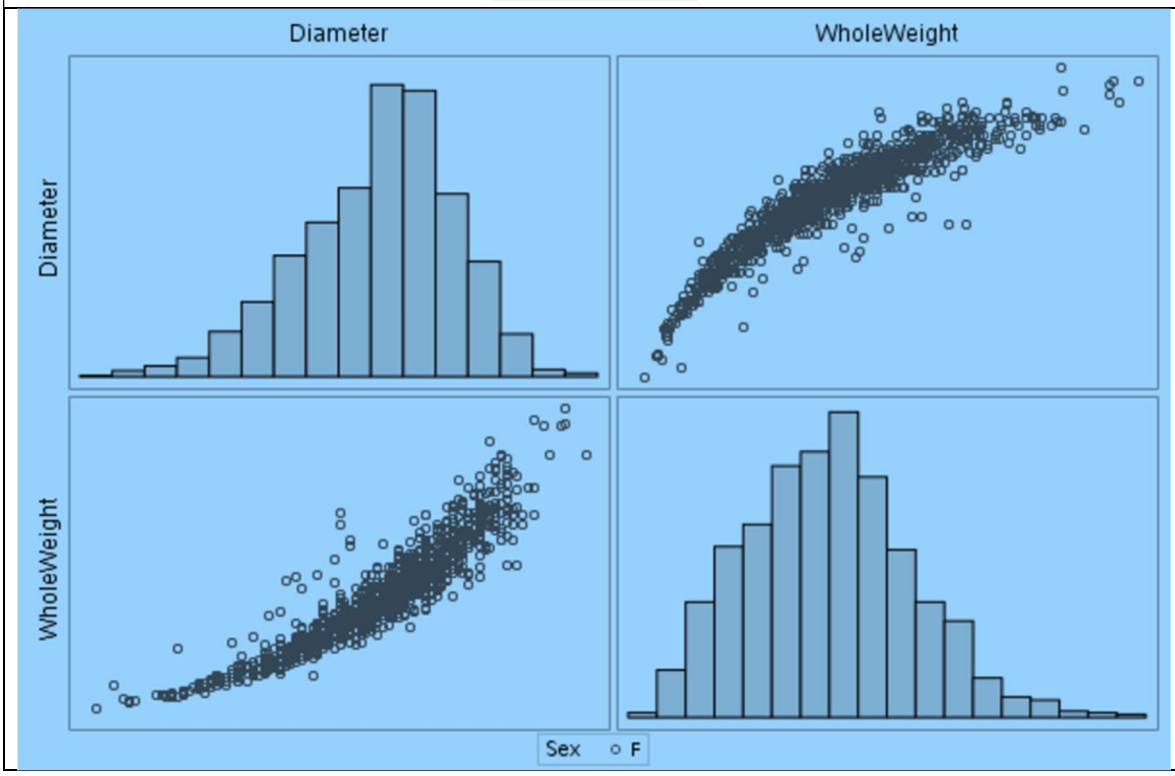
```
proc sgscatter data = abalone2;
matrix logRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight
ShellWeight/ diagonal=(histogram) group = Sex;
run;quit;
```

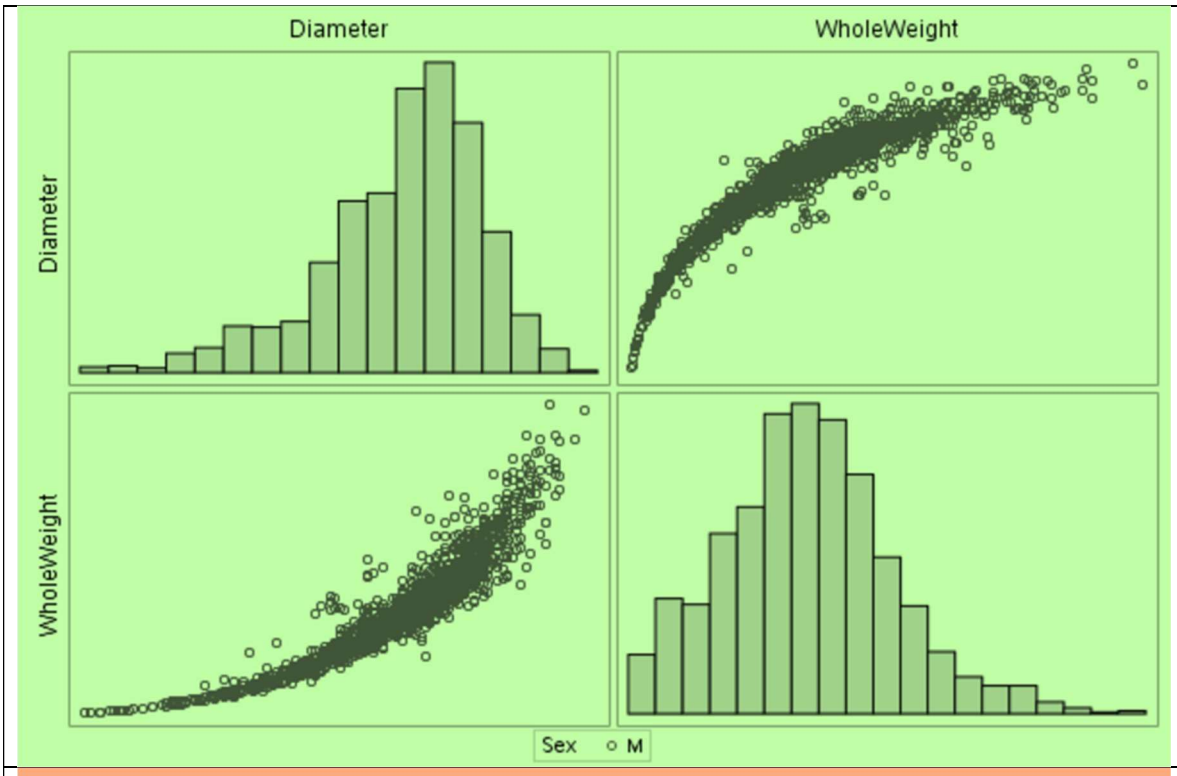


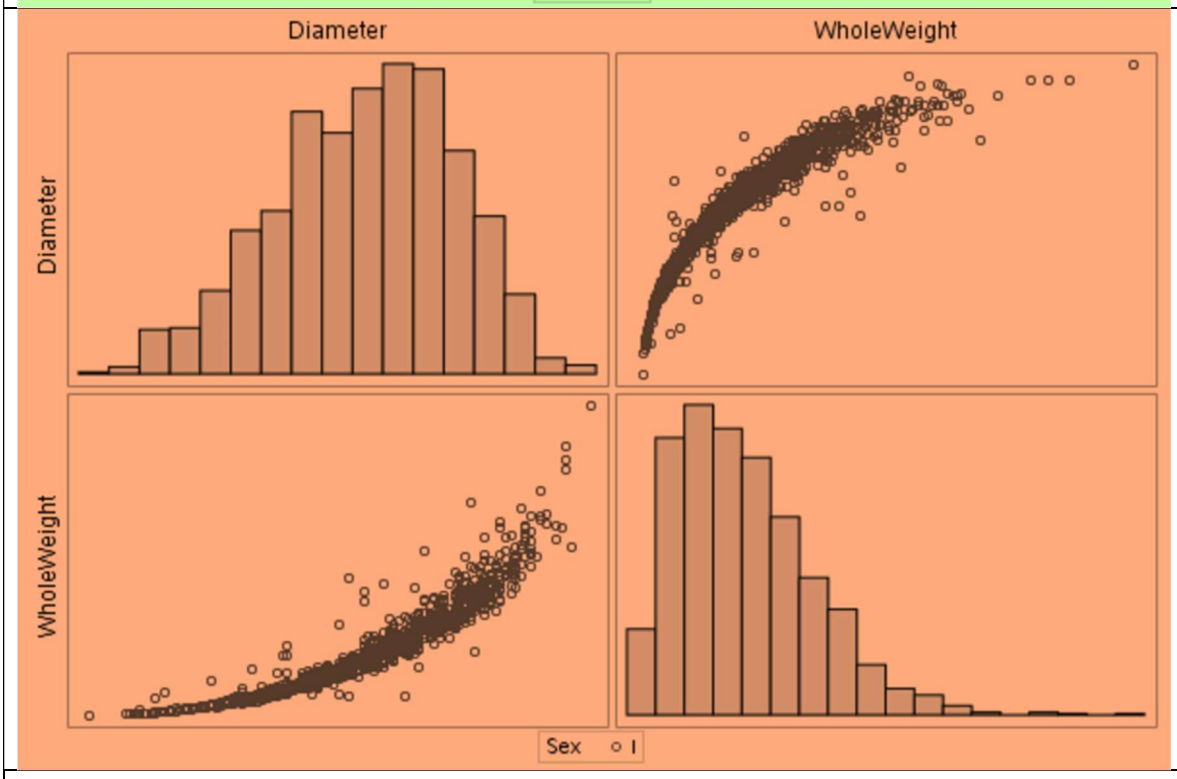
Scatter plot by Sex: { Diameter WholeWeight }

```
proc sgscatter data = abalone2;
matrix Diameter WholeWeight/ diagonal=(histogram) group = Sex;
run;quit;
```


Diameter
WholeWeight
Diameter
WholeWeight
Sex F I M


Diameter
WholeWeight
Diameter
WholeWeight
Sex F

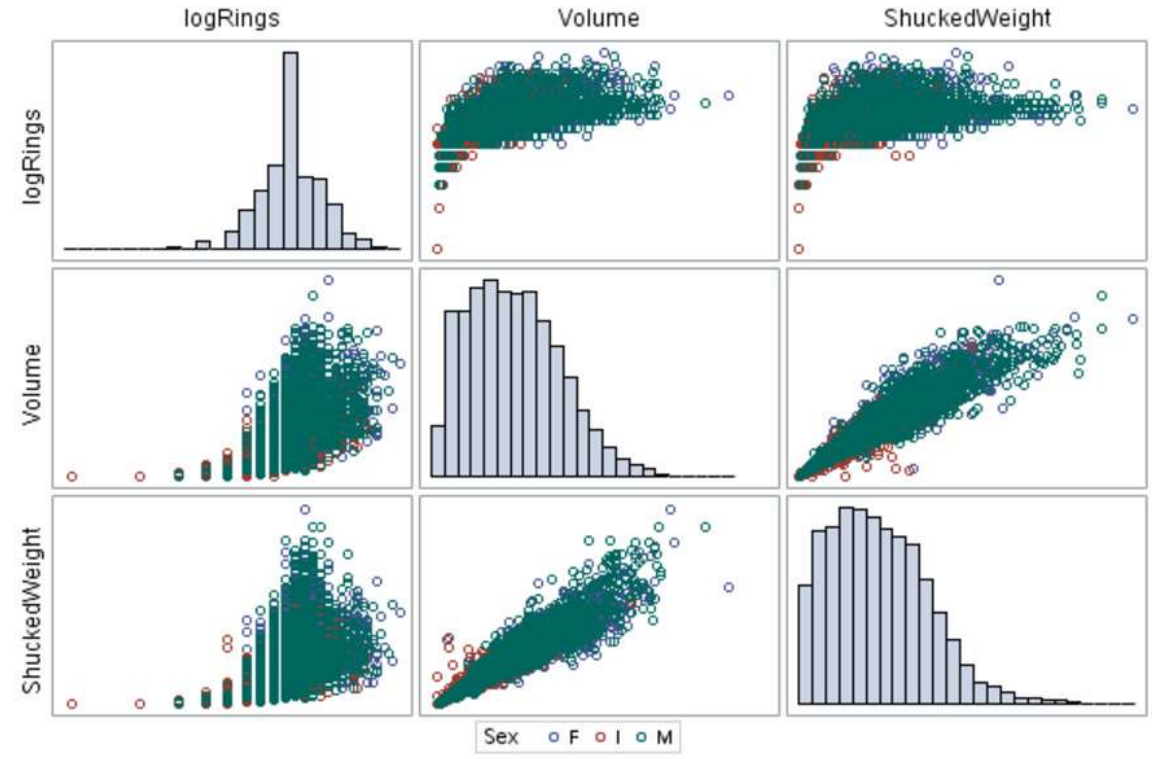


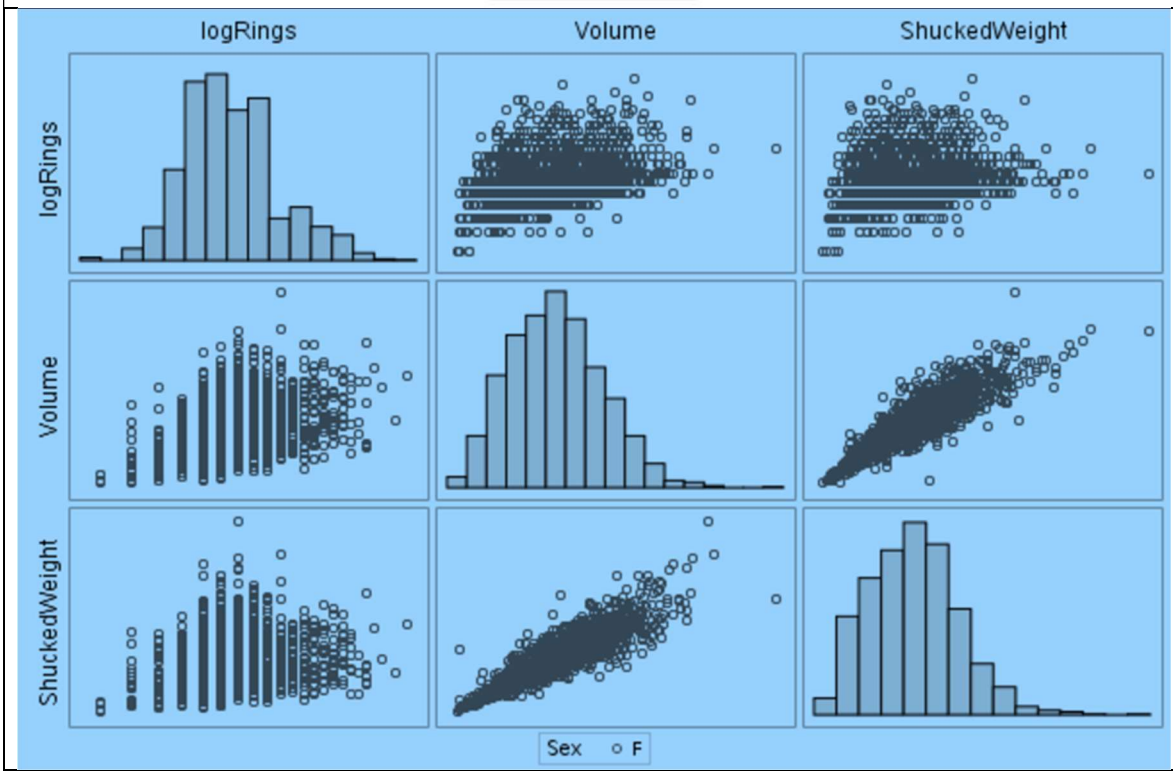


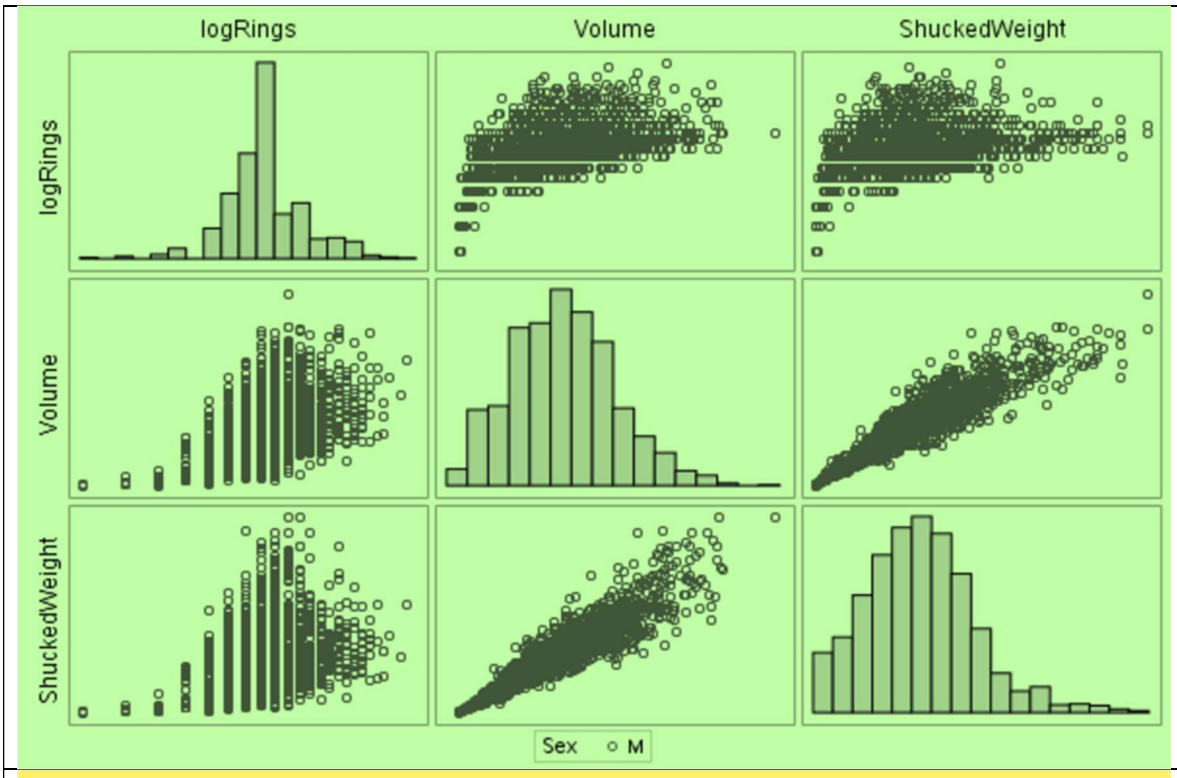
The shape/direction of spread is about same for all Sex levels. This is an indication of constant covariance matrix.

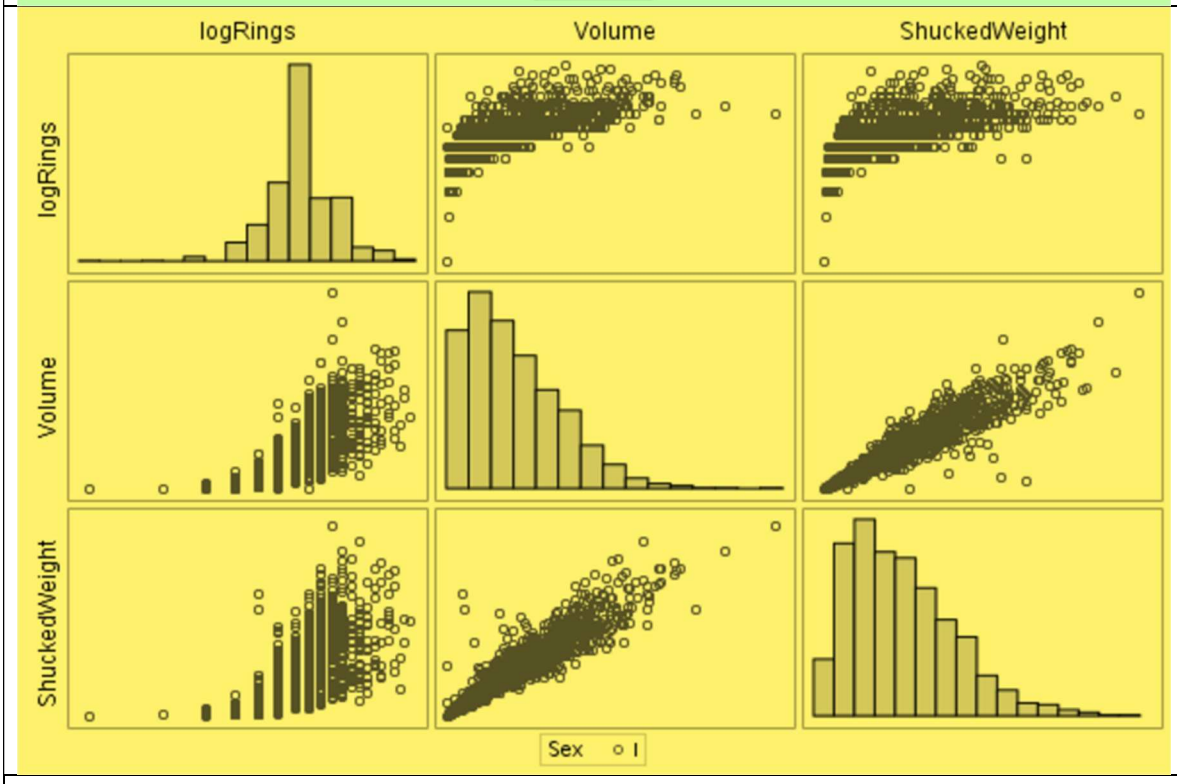
Scatter Plot by Sex: { logRings Volume ShuckedWeight }

```
proc sgscatter data = abalone2;
matrix logRings Volume ShuckedWeight / diagonal=(histogram) group = Sex;
run;quit;
```







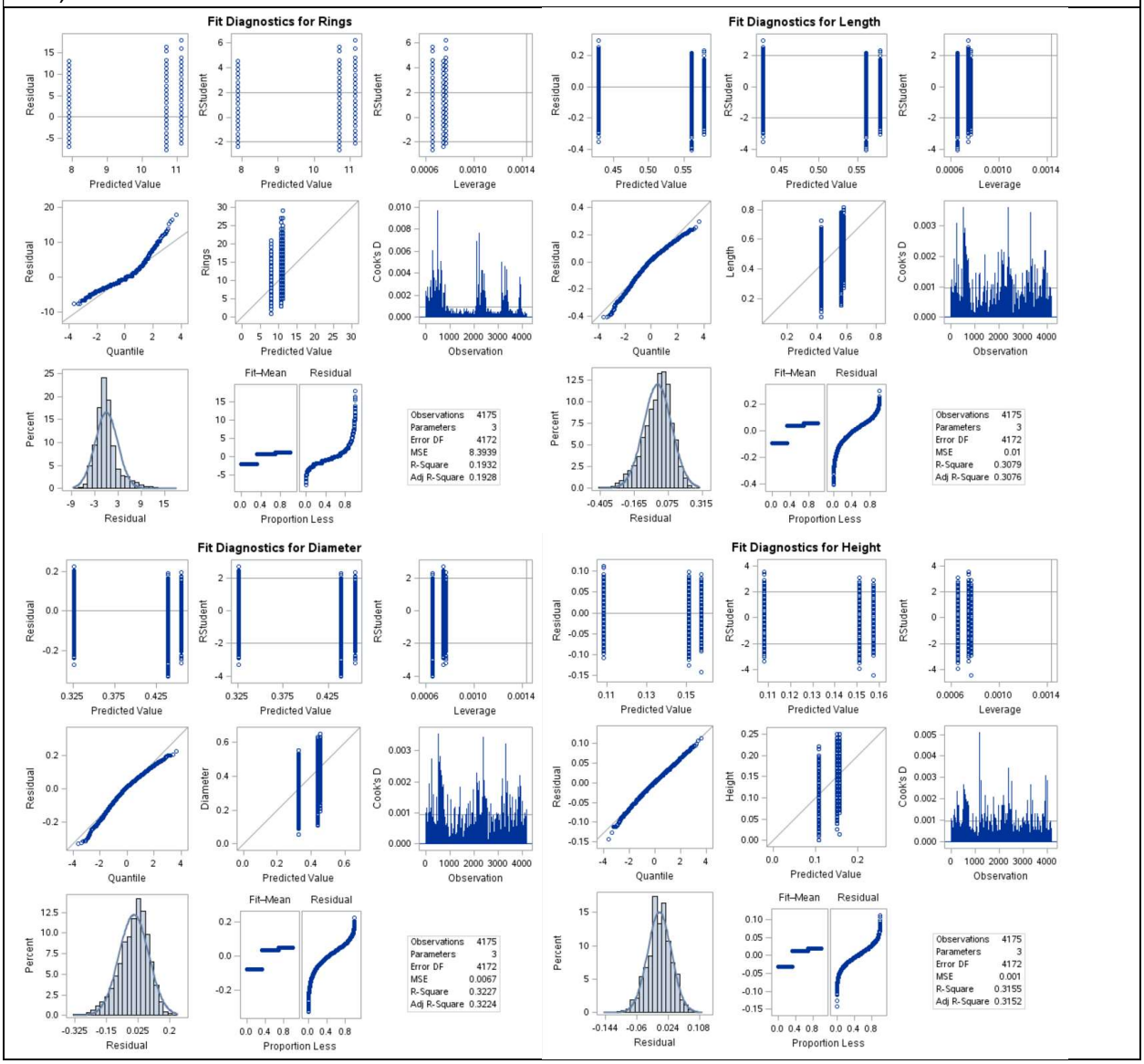


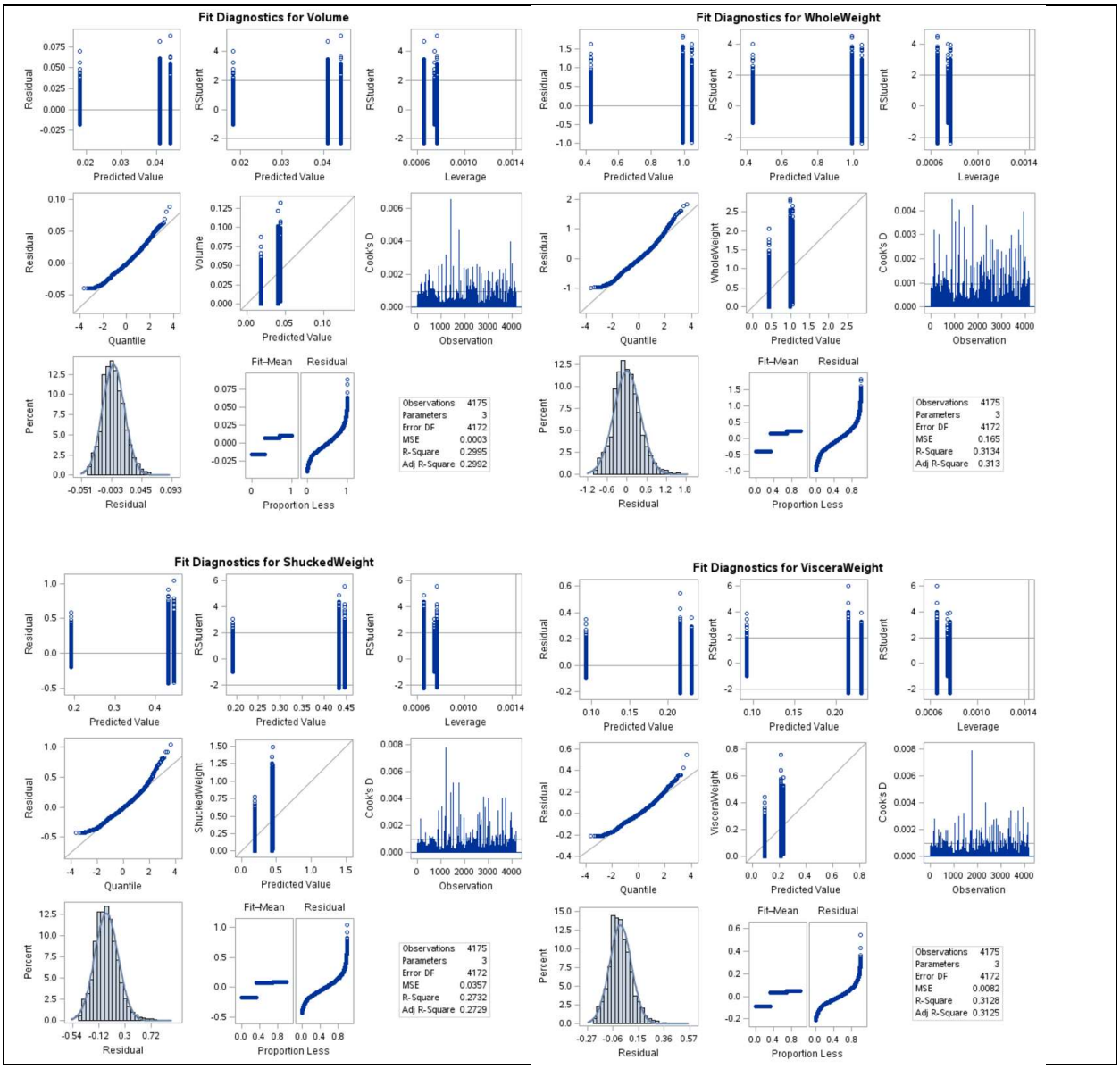
The shape/direction of spread is about same for all Sex levels. This is an indication of constant covariance matrix.

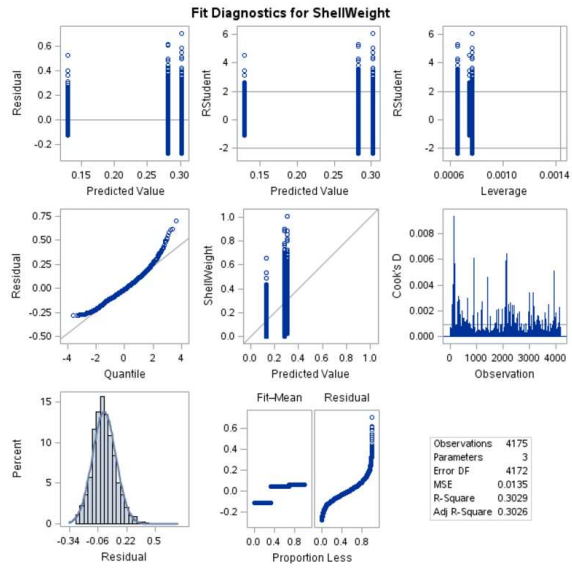
Appendix B (Assumptions investigation output and SAS code)

MANOVA is used to see the model error distributions.

```
Proc GLM Data=abalone2 plots = all;
 Class Sex;
 Model Rings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight
ShellWeight = Sex;
 Manova H=_All_ / PrintE PrintH Canonical;
Run;
```




Fit Diagnostics for Volume
Residual
RStudent
RStudent
Predicted Value
Predicted Value
Leverage
Residual
Volume
Cook's D
Quantile
Predicted Value
Observation
Percent
Fit–Mean
Residual
Residual
Proportion Less
Observations 4175
Parameters 3
Error DF 4172
MSE 0.0003
R-Square 0.2995
Adj R-Square 0.2992
Fit Diagnostics for WholeWeight
Residual
RStudent
RStudent
Predicted Value
Predicted Value
Leverage
Residual
WholeWeight
Cook's D
Quantile
Predicted Value
Observation
Percent
Fit–Mean
Residual
Residual
Proportion Less
Observations 4175
Parameters 3
Error DF 4172
MSE 0.165
R-Square 0.3134
Adj R-Square 0.313
Fit Diagnostics for ShuckedWeight
Residual
RStudent
RStudent
Predicted Value
Predicted Value
Leverage
Residual
ShuckedWeight
Cook's D
Quantile
Predicted Value
Observation
Percent
Fit–Mean
Residual
Residual
Proportion Less
Observations 4175
Parameters 3
Error DF 4172
MSE 0.0357
R-Square 0.2732
Adj R-Square 0.2729
Fit Diagnostics for VisceraWeight
Residual
RStudent
RStudent
Predicted Value
Predicted Value
Leverage
Residual
VisceraWeight
Cook's D
Quantile
Predicted Value
Observation
Percent
Fit–Mean
Residual
Residual
Proportion Less
Observations 4175
Parameters 3
Error DF 4172
MSE 0.0082
R-Square 0.3128
Adj R-Square 0.3125

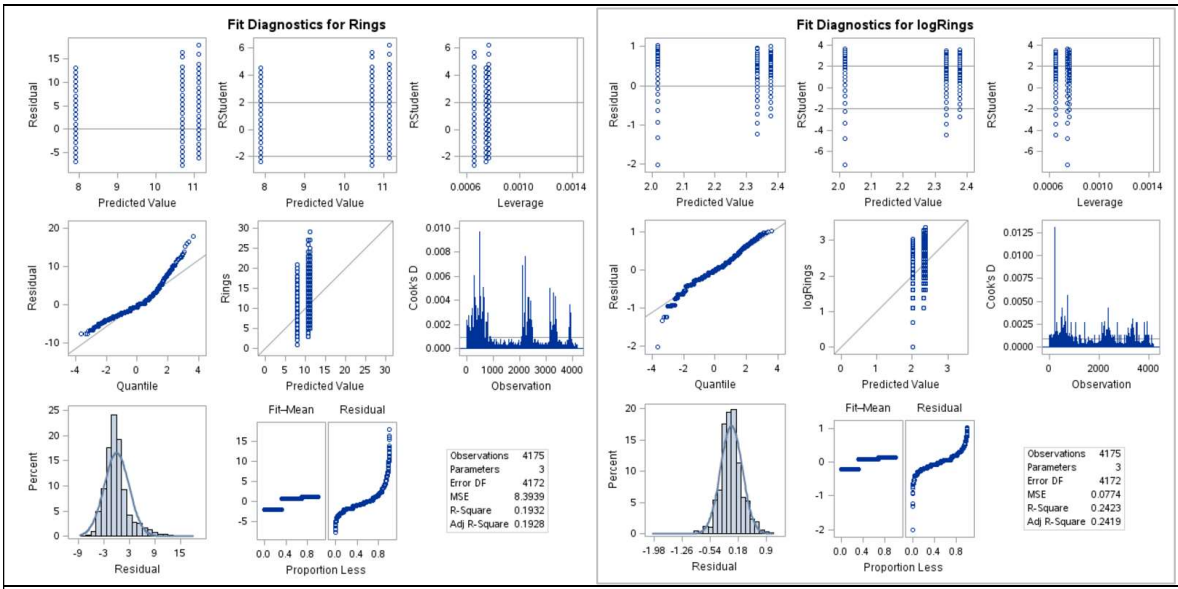


Based on the Residual vs. Quantile plots for each variable, following variables were not following the normal distribution.

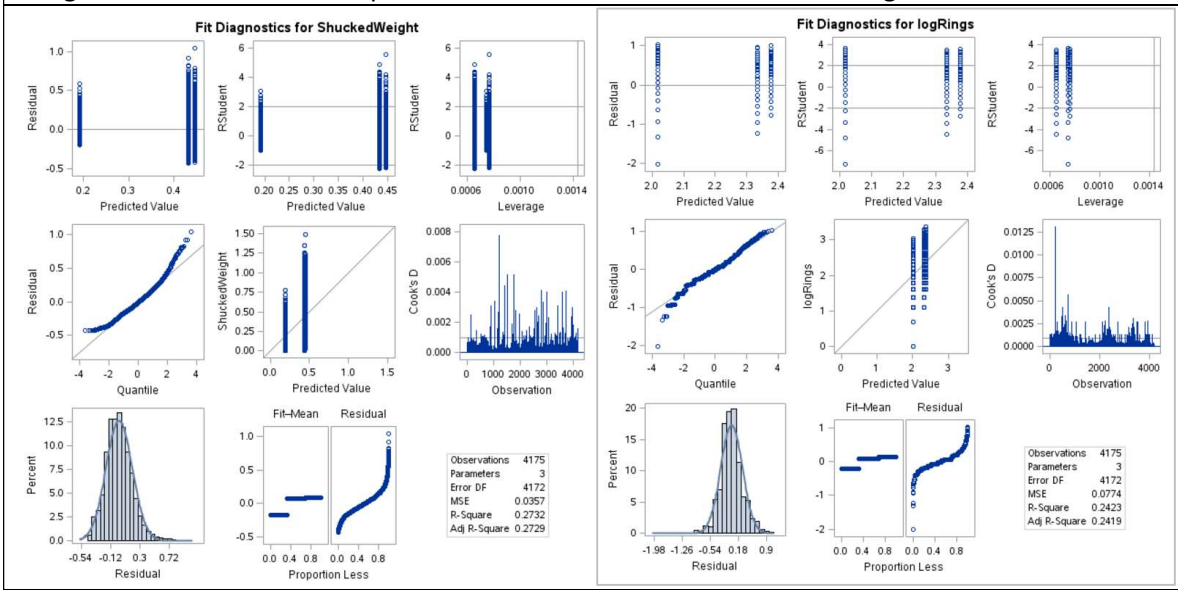
[Rings, Volume, WholeWeight, ShuckedWeight, VisceraWeight, ShellWeight]

Log Transformations

```
data abalone2;
set abalone;
logRings = log(Rings);
logVolume = log(Volume);
logWholeWeight = log(WholeWeight);
logShuckedWeight = log(ShuckedWeight);
logVisceraWeight = log(VisceraWeight);
logShellWeight = log(ShellWeight);
run;
```



'Rings' variable showed an improvement in error distribution after the log transformation.



Rest variables showed worse error distributions after the log transformation as shown above as an example. The scale of 'Residual vs. Quantile' plot is changed and the error is increased.

Other Transformations: Squared transformation and mean centering transformation were tried but there was no improvements in error distributions.

```
*squared transformation;
data abaloneSquare;
set abalone;
SRings = Rings*Rings;
SLength = Length*Length;
SDiameter = Diameter*Diameter;
SHeight = Height*Height;
```

```
SVolume = Volume*Volume;
SWholeWeight = WholeWeight*WholeWeight;
SShuckedWeight = ShuckedWeight*ShuckedWeight;
SVisceraWeight = VisceraWeight*VisceraWeight;
SShellWeight = ShellWeight*ShellWeight;
run;
proc print data = abaloneSqaure; run;

*Mean Centering Transformation;
data abaloneMC;
set abalone;
*MRings = Rings*Rings;
*MLength = Length*Length;
*MDiameter = Diameter*Diameter;
*MHeight = Height*Height;
*MVolume = Volume*Volume;
MWholeWeight = (WholeWeight-0.8285)**2;
MShuckedWeight = (ShuckedWeight-0.3591)**2;
MVisceraWeight = (VisceraWeight-0.1085)**2;
MShellWeight = (ShellWeight-0.2388)**2;
run;

*using Mean Centering;
proc means data =  abalone n mean max min range std fw=8;
var _numeric_ ;
output out = meansout mean = mean std = std;
title 'Summary Stat';
run;


proc glm data=abaloneMC plots=all;
class sex;
model MWholeWeight     MShuckedWeight MVisceraWeight MShellWeight    = Sex;
output out=resides r=ral rfe rmg rca rna;
run;


* using squared transformatin;
proc glm data=abaloneSquare plots=all;
class sex;
model SRings SLength SDiameter SVolume SHeight SWholeWeight
SShuckedWeight SVisceraWeight SShellWeight    = Sex;
output out=resides r=ral rfe rmg rca rna;
run;
```

Appendix C MANOVA and ANOVA output

```
Proc GLM Data=abalone2 plots = all;
 Class Sex;
 Model logRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight
ShellWeight = Sex;
 Manova H=_All_ / PrintE PrintH Canonical;
Run;
```

ANOVA F-test shows variables are significantly the sex levels for all variables. P-values from all F-tests are less than 0.0001.

Weight related variables shows slightly off-normal error distributions. At the same time the residual variance for each Sex level seem slightly different which is a violation of the ANOVA assumption.

… another assumption violation … tried many transformations above but there was no effective method to fix the unequal variance and residual normality.

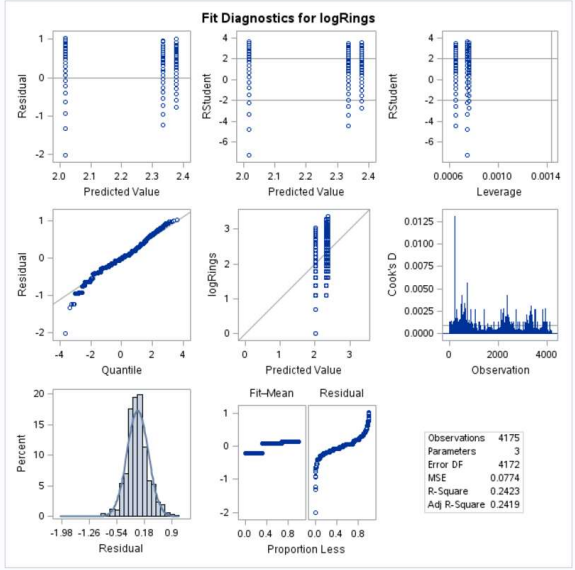
The GLM Procedure
Dependent Variable: logRings

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	103.2861017	51.6430508	667.01	<.0001
Error	4172	323.0182477	0.0774253		
Corrected Total	4174	426.3043494			

R-Square	Coeff Var	Root MSE	logRings Mean
0.242283	12.39050	0.278254	2.245705

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	103.2861017	51.6430508	667.01	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	103.2861017	51.6430508	667.01	<.0001



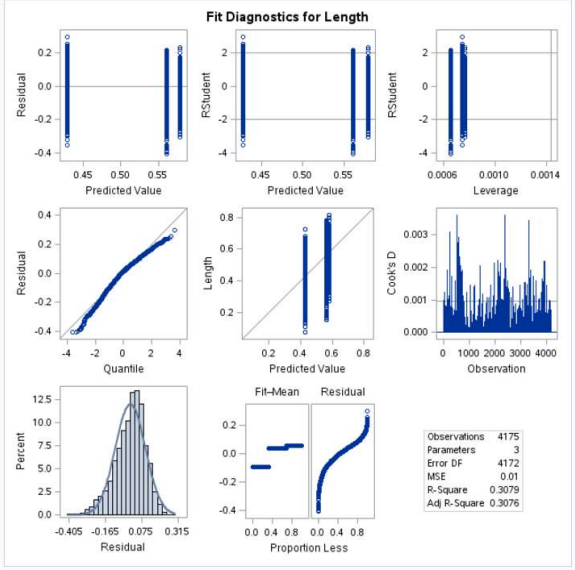
The GLM Procedure
Dependent Variable: Length

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	18.53530134	9.26765067	928.22	<.0001
Error	4172	41.65472863	0.00998435		
Corrected Total	4174	60.19002996			

R-Square	Coeff Var	Root MSE	Length Mean
0.307946	19.07030	0.099922	0.523965

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	18.53530134	9.26765067	928.22	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	18.53530134	9.26765067	928.22	<.0001



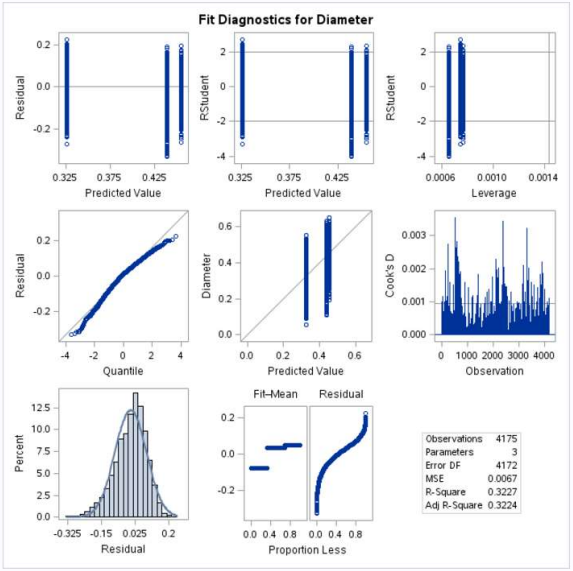
The GLM Procedure
Dependent Variable: Diameter

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	13.26347392	6.63173696	993.93	<.0001
Error	4172	27.83658985	0.00667224		
Corrected Total	4174	41.10006377			

R-Square	Coeff Var	Root MSE	Diameter Mean
0.322712	20.02759	0.081684	0.407856

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	13.26347392	6.63173696	993.93	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	13.26347392	6.63173696	993.93	<.0001



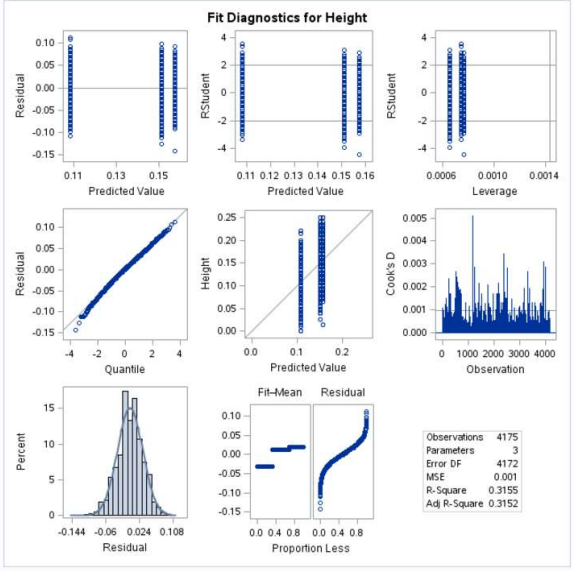
The GLM Procedure
Dependent Variable: Height

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.95080039	0.97540020	961.43	<.0001
Error	4172	4.23263012	0.00101453		
Corrected Total	4174	6.18343051			

R-Square	Coeff Var	Root MSE	Height Mean
0.315468	22.88376	0.031852	0.139189

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	1.95080039	0.97540020	961.43	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	1.95080039	0.97540020	961.43	<.0001



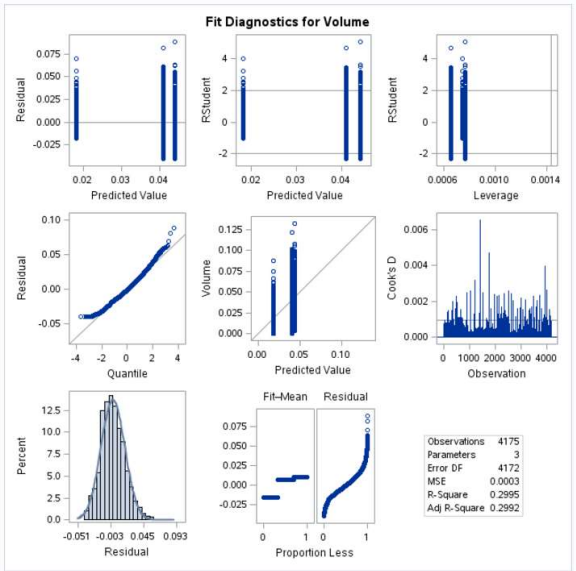
The GLM Procedure
Dependent Variable: Volume

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.54444984	0.27222492	892.01	<.0001
Error	4172	1.27321717	0.00030518		
Corrected Total	4174	1.81766700			

R-Square	Coeff Var	Root MSE	Volume Mean
0.299532	50.43300	0.017469	0.034639

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	0.54444984	0.27222492	892.01	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	0.54444984	0.27222492	892.01	<.0001



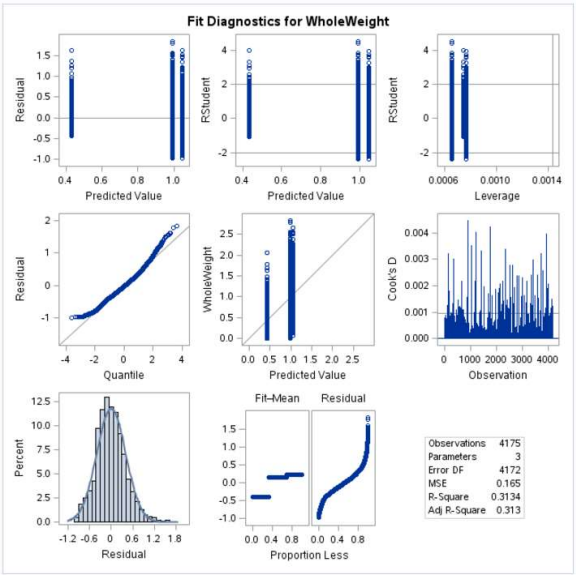
The GLM Procedure
Dependent Variable: WholeWeight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	314.094419	157.047209	952.06	<.0001
Error	4172	688.192570	0.164955		
Corrected Total	4174	1002.286989			

R-Square	Coeff Var	Root MSE	WholeWeight Mean
0.313378	49.02384	0.406147	0.828468

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	314.0944190	157.0472095	952.06	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	314.0944190	157.0472095	952.06	<.0001



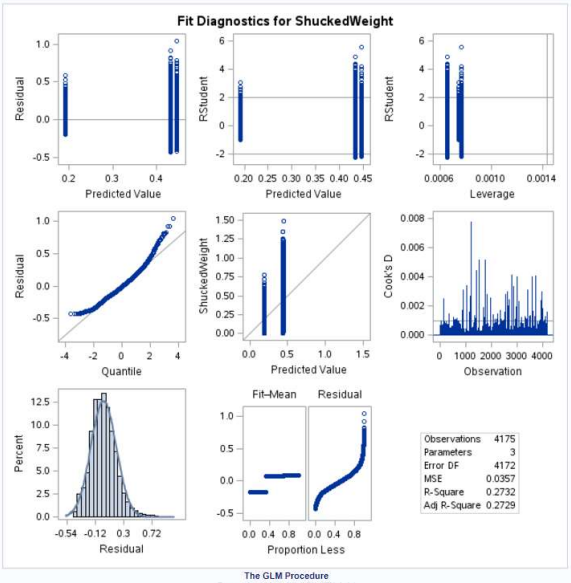
The GLM Procedure
Dependent Variable: ShuckedWeight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	56.0586164	28.0293082	784.18	<.0001
Error	4172	149.1220999	0.0357436		
Corrected Total	4174	205.1807162			

R-Square	Coeff Var	Root MSE	ShuckedWeight Mean
0.273216	52.63429	0.189060	0.359195

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	56.05861637	28.02930819	784.18	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	56.05861637	28.02930819	784.18	<.0001



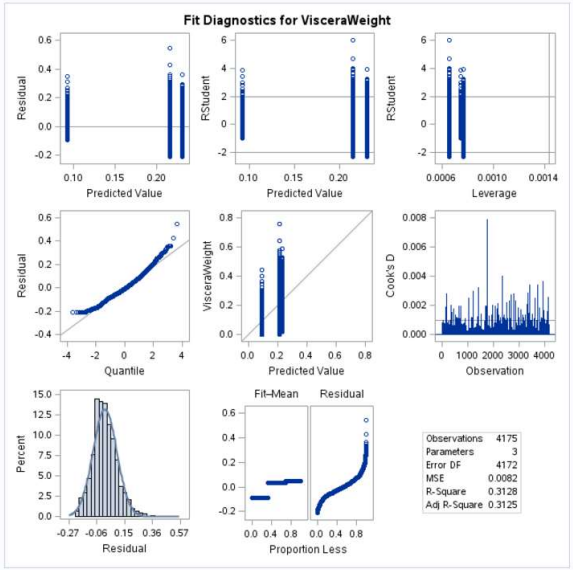
The GLM Procedure
Dependent Variable: VisceraWeight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	15.66607656	7.83303828	949.65	<.0001
Error	4172	34.41198384	0.00824832		
Corrected Total	4174	50.07806040			

R-Square	Coeff Var	Root MSE	VisceraWeight Mean
0.312833	50.30595	0.090820	0.180536

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	15.66607656	7.83303828	949.65	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	15.66607656	7.83303828	949.65	<.0001



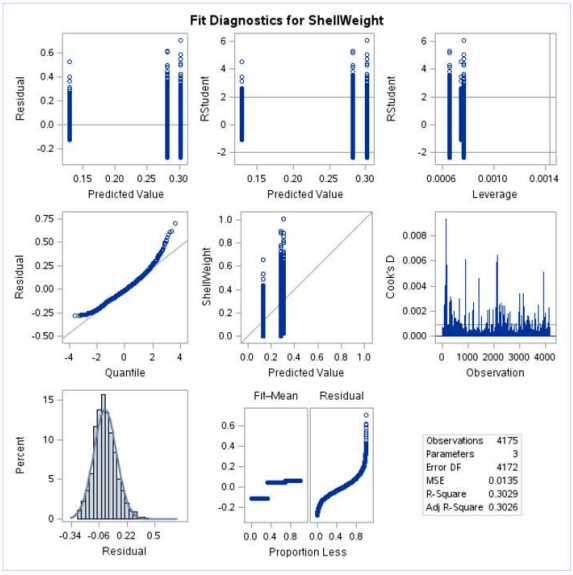
The GLM Procedure
Dependent Variable: ShellWeight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24.48642405	12.24321203	906.49	<.0001
Error	4172	56.34780548	0.01350619		
Corrected Total	4174	80.83422954			

R-Square	Coeff Var	Root MSE	ShellWeight Mean
0.302921	48.66862	0.116216	0.238791

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	2	24.48642405	12.24321203	906.49	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	2	24.48642405	12.24321203	906.49	<.0001



Appending D – Full Code including all trials

```
/* MSDS6372 Project 2 Abalone Problem */
/* James Park 08JULY2017 */


/* Loading CSV */
FILENAME REFFILE '/home/jmpark1/Abalone  Clean Data.csv';

PROC IMPORT DATAFILE=REFFILE
		DBMS=csv
		OUT=abalone;
		GETNAMES=yes;
RUN;
PROC CONTENTS DATA=abalone; RUN;
PROC PRINT DATA = abalone;run;

/* Data Exploring */

data abaloneM;
set abalone;
if Sex in ('M');
proc print data = abeloneM; run;

data abaloneF;
set abalone;
if Sex in ('F');
proc print data = abeloneF; run;

data abaloneI;
set abalone;
if Sex in ('I');
proc print data = abeloneI; run;

*log variables;
data abalone2;
set abalone;
logRings = log(Rings);
logLength = log(Length);
logDiameter = log(Diameter);
logHeight = log(Height);
logVolume = log(Volume);
logWholeWeight = log(WholeWeight);
logShuckedWeight = log(ShuckedWeight);
logVisceraWeight = log(VisceraWeight);
logShellWeight = log(ShellWeight);
Volume = Length*Diameter*Height;
run;
proc print data = abalone2; run;

*squared transformation;
data abaloneSquare;
set abalone;
SRings = Rings*Rings;
SLength = Length*Length;
SDiameter = Diameter*Diameter;
SHeight = Height*Height;
SVolume = Volume*Volume;
SWholeWeight = WholeWeight*WholeWeight;
SShuckedWeight = ShuckedWeight*ShuckedWeight;
SVisceraWeight = VisceraWeight*VisceraWeight;
SShellWeight = ShellWeight*ShellWeight;
run;
proc print data = abaloneSqaure; run;

*Mean Centering Transformation;
data abaloneMC;
set abalone;
*MRings = Rings*Rings;
*MLength = Length*Length;
*MDiameter = Diameter*Diameter;
*MHeight = Height*Height;
*MVolume = Volume*Volume;
MWholeWeight = (WholeWeight-0.8285)**2;
MShuckedWeight = (ShuckedWeight-0.3591)**2;
```

```
MVisceraWeight = (VisceraWeight-0.1085)**2;
MShellWeight = (ShellWeight-0.2388)**2;
run;

*using Mean Centering;
proc glm data=abaloneMC plots=all;
class sex;
model MWholeWeight    MShuckedWeight          MVisceraWeight          MShellWeight              = Sex;
output out=resides r=ral rfe rmg rca rna;
run;



* using squared transformatin;
proc glm data=abaloneSquare plots=all;
class sex;
model SRings SLength     SDiameter SVolume SHeight              SWholeWeight              SShuckedWeight          SVisceraWeight          SShellWeight
= Sex;
output out=resides r=ral rfe rmg rca rna;
run;

proc means data =  abalone n mean max min range std fw=8;
var _numeric_ ;
output out = meansout mean = mean std = std;
title 'Summary Stat';
run;

data abaloneFemaleMale;
set abalone2;
keep Sex WholeWeight Volume;
if Sex in ('M', 'F'); run;
proc print data = abaloneFemaleMale; run;

proc ttest data = abaloneFemaleMale ALPHA = .01;
CLASS sex;
var Volume WholeWeight;
run;


proc means data = abalone n mean max min range std fw=8;
var _numeric_;
by sex;
output out = meansout mean = mean std = std;
title 'Summary Stat by Sex';
run;

proc corr data = abalone;
run;

/* Scatter Plot */
proc sgscatter data=abalone ;
matrix Rings Length        Diameter Volume Height WholeWeight              ShuckedWeight          VisceraWeight          ShellWeight
/ diagonal=(histogram) ;
run;

* Frequency distribution check;
proc freq data = abalone; run;

proc univariate data = abalone plots;
by sex;
Run;

proc univariate data = abalone alpha = 0.05 mu0=0 normal plot;
 class sex;
 var Rings Length          Diameter Volume Height WholeWeight              ShuckedWeight          VisceraWeight          ShellWeight;
 histogram /normal
 ;inset mean std median min max;
 run;quit;

*Initial Check of normality and covariance matrix;

proc glm data=abalone2 plots=all;
class sex;
model logRings Length     Diameter Volume Height WholeWeight              ShuckedWeight          VisceraWeight          ShellWeight  = Sex;
output out=resides r=ral rfe rmg rca rna;
```

```
run;

proc glm data=abalone2 plots=all;
class sex;
model logRings logLength logDiameter logHeight logVolume logWholeWeight logShuckedWeight logVisceraWeight
        logShellWeight = Sex;
output out=resides r=ral rfe rmg rca rna;
run;

* Check for equal variance and covariance matrix;
proc discrim data=abalone2 pool=test;
  class sex;
  var logRings Length Diameter Volume Height WholeWeight ShuckedWeight VisceraWeight ShellWeight;
 run;

* Rerun Check for equal variance and covariance matrix with fewer variables;
proc discrim data=abalone pool=test;
  class sex;
  var Volume WholeWeight;
 run;


/* Scatter Plot with log data*/
proc sgscatter data=abalone2 ;
matrix logRings logLength logDiameter logHeight logWholeWeight logShuckedWeight logVisceraWeight logShellWeight
/ diagonal=(histogram) ;
run;

*PCA first attempt with all data;
ods rtf ; *places results in a word file;
ods graphics on;
proc princomp plots=all data=abalone cov out=pca;
    var Rings Length Diameter Height WholeWeight ShuckedWeight VisceraWeight ShellWeight;
    run;
proc print data=pca; run;
proc gplot data=pca;
        plot prin2*prin1=location;
        run;
ods graphics off;
ods rtf close; *completes word file;

*PCA second attempt with only male data;
ods rtf ; *places results in a word file;
ods graphics on;
proc princomp plots=all data=abaloneM cov out=pcaM;
    var Rings Length Diameter Height WholeWeight ShuckedWeight VisceraWeight ShellWeight;
    run;
proc print data=pcaM; run;
proc gplot data=pcaM;
        plot prin2M*prin1M=locationM;
        run;
ods graphics off;
ods rtf close; *completes word file;

*PCA third attempt with only female data;
ods rtf ; *places results in a word file;
ods graphics on;
proc princomp plots=all data=abaloneF cov out=pcaF;
    var Rings Length Diameter Height WholeWeight ShuckedWeight VisceraWeight ShellWeight;
    run;

data pcaF2;
set pcaF;
prin1F=prin1;
prin2F=prin2;
keep prin1F prin2F sex;
run;
proc print data=pcaF2; run;

data pcaM2;
set pcaM;
prin1M=prin1;
prin2M=prin2;
keep prin1M prin2M sex;
run;
```

```
proc print data=pcaM2; run;

proc gplot data=pcaF2;
                plot prin2F*prin1F=locationF;
                run;
ods graphics off;
ods rtf close; *completes word file;

*pca combined;
 data pcaCombined;
 set pcaF2 pcaM2;
  run;
 proc print pcaCombined; run;

 data pcaCombined2;
 set pcaCombined;
 if prin1M='.' then prin1C = prin1F;
   else if prin1F='.' then prin1C= prin1M;
  if prin2M='.' then prin2C = prin2F;
   else if prin2F='.' then prin2C= prin2M;
 run;
 proc print data=pcaCombined2; run;

*if we wanted to build a model to predict age;
proc glm data = abalone plots = all;
class sex;
model Rings = sex LengthDiameter     Height        WholeWeight              ShuckedWeight            VisceraWeight            ShellWeight / solution;
run;quit;

*LDA STARTING CODE;
*First run a one-way ANOVA for each response;
Proc GLM Data=abalone plot=diagnostics;
 Class Sex;
 *Model height Length    = Sex;
 Model Rings Length       Diameter Height          WholeWeight              ShuckedWeight            VisceraWeight            ShellWeight = Sex;
 lsmeans sex / pdiff tdiff adjust=bon;
 Output Out=Errs R=Eact Eant; *to look at residuals;
 Means Sex / CLM T Alpha=0.0167; *code if we assume we are ok on assumptions.  Adjusted Bonferonni;
 Means Sex / CLDIFF T Alpha=0.0167; *code if we assume we are ok on assumptions.  Adjusted Bonferonni;
 *Contrast 'Sex contrast' sex -0.5 0 0.5;  *order would be F, I, M;
 *Estimate 'Sex estimate' sex -0.5 0 0.5;  *order would be F, I, M;
 Manova H=_All_ / PrintE PrintH Canonical;  ***Must have this to actually function as a multiple response test;
 Contrast 'Sex contrast F M' sex -0.5 0 0.5;  *order would be F, I, M;
 Contrast 'Sex contrast F I' sex -0.5 0.5 0;  *order would be F, I, M;
 ODS Output Cancoef=CanonicalCoefficients;
Run;
Quit;

*Second Run with log data one-way ANOVA for each response;
Proc GLM Data=abalone2 plot=diagnostics;
 Class Sex;
 *Model height Length    = Sex;
 Model logRings logLength              logDiameter logHeight    logWholeWeight           logShuckedWeight         logVisceraWeight         logShellWeight = Sex;
 lsmeans sex / pdiff tdiff adjust=bon;
 Output Out=Errs R=Eact Eant; *to look at residuals;
 Means Sex / CLM T Alpha=0.0167; *code if we assume we are ok on assumptions.  Adjusted Bonferonni;
 Means Sex / CLDIFF T Alpha=0.0167; *code if we assume we are ok on assumptions.  Adjusted Bonferonni;
 *Contrast 'Sex contrast' sex -0.5 0 0.5;  *order would be F, I, M;
 *Estimate 'Sex estimate' sex -0.5 0 0.5;  *order would be F, I, M;
 Manova H=_All_ / PrintE PrintH Canonical;  ***Must have this to actually function as a multiple response test;
 Contrast 'Sex contrast F M' sex -0.5 0 0.5;  *order would be F, I, M;
 Contrast 'Sex contrast F I' sex -0.5 0.5 0;  *order would be F, I, M;
 ODS Output Cancoef=CanonicalCoefficients;
Run;
Quit;

*using only principle components;
Proc GLM Data=pcaCombined2 plot=diagnostics;
 Class Sex;
 Model prin1C prin2C= Sex;
 lsmeans sex / pdiff tdiff adjust=bon;
 Output Out=Errs R=Eact Eant; *to look at residuals;
 *Means Sex / CLM T Alpha=0.0125; *code if we assume we are ok on assumptions.  Adjusted Bonferonni;
 *Means Sex / CLDIFF T Alpha=0.0125; *code if we assume we are ok on assumptions.  Adjusted Bonferonni;
 *Contrast 'Sex contrast' sex -1 1;  *order would be F, M;
```

```
 *Estimate 'Sex estimate' sex -1 1;  *order would be F, M;
 Manova H=_All_ / PrintE PrintH Canonical;  ***Must have this to actually function as a multiple response test;
 Contrast 'Sex contrast F M' sex -1 1;  *order would be F M;
 *Contrast 'Sex contrast F I' sex -1 1;  *order would be F M;
 ODS Output Cancoef=CanonicalCoefficients;
Run;
Quit;

*using only Sex M and F data with wholeweight and volume;
Proc GLM Data=abaloneFemaleMale plot=diagnostics;
 Class Sex;
 Model WholeWeight Volume = Sex;
 lsmeans sex / pdiff tdiff adjust=bon;
 Output Out=Errs R=Eact Eant; *to look at residuals;
 *Means Sex / CLM T Alpha=0.0125; *code if we assume we are ok on assumptions.  Adjusted Bonferonni;
 *Means Sex / CLDIFF T Alpha=0.0125; *code if we assume we are ok on assumptions.  Adjusted Bonferonni;
 *Contrast 'Sex contrast' sex -1 1;  *order would be F, M;
 *Estimate 'Sex estimate' sex -1 1;  *order would be F, M;
 Manova H=_All_ / PrintE PrintH Canonical;  ***Must have this to actually function as a multiple response test;
 Contrast 'Sex contrast F M' sex -1 1;  *order would be F M;
 *Contrast 'Sex contrast F I' sex -1 1;  *order would be F M;
 ODS Output Cancoef=CanonicalCoefficients;
Run;
Quit;

proc glm data=abalone; * order=data;
class sex;
model Rings Length Diameter Height WholeWeight ShuckedWeight VisceraWeight ShellWeight = sex;
lsmeans sex / pdiff tdiff adjust=bon;
means sex / lsd bon tukey scheffe;
*contrast 'lin' sex -1 1 0;*data order would be F, I, M;
estimate ' I M Comparison' Sex -1 0 1;
run;

PROC DISCRIM data = abalone2 CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=test wcov pcov;
CLASS sex;
VAR logRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight ShellWeight;
PRIORS EQUAL;
run;
quit;

data abaloneMFonly;
set abalone2;
if Sex in ('M','F'); run;
proc print data = abaloneMFonly; run;

PROC DISCRIM data = abaloneMFonly CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=test wcov pcov;
CLASS sex;
VAR logRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight ShellWeight;
PRIORS EQUAL;
run;
quit;

PROC DISCRIM data = abalone2 CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=test;
CLASS sex;
VAR logRings logLength logDiameter logHeight logVolume logWholeWeight logShuckedWeight logVisceraWeight logShellWeight;
PRIORS EQUAL;
run;
quit;


PROC DISCRIM data = pcaCombined2 CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=YES;
CLASS sex;
VAR Prin1C Prin2C;
PRIORS EQUAL;
run;
quit;
proc print data=dis_func; run;

*automatic variable selection using stepdisc for LDA variable selection;
proc stepdisc data=abaloneMFonly bsscp tsscp;
  class Sex;
  var LogRings Length Diameter Height Volume WholeWeight ShuckedWeight VisceraWeight ShellWeight;
run;
```

```
*rerun LDA using auto-selected variables;
PROC DISCRIM data = abalone CROSSVALIDATE OUTSTAT=DIS_FUNC POOL=test wcov pcov;
CLASS sex;
VAR Diameter Height ShuckedWeight VisceraWeight;
PRIORS EQUAL;
run;
quit;
```