Exercise 1

a) For weight, create a cross-tabulation of the mean, standard deviation and counts by species and width group. Comment on any interesting features (e.g. apparent differences between species and/or width groups).

		Weight		
		Mean	Std	N
Species	**widthgroup**			
Bream	**thinner**	344.00	137.22	3
	wider	653.29	192.29	31
Perch	**thinner**	133.19	64.58	33
	wider	739.57	263.64	23

The weight sample means for each species and width group suggest that there may be some weight differences between species and width group. The counts indicate that there is not equal number of observations for each species and width group. It can be noted that the data is not balanced with respect to the species and width group variables.

Moving on to the cross-classified table for all two explanatory variables. There is some variation in weight means. While it is easy to see which combinations seem to have higher and lower sample means, it is not as easy to see where the differences are with respect to the original variables. The mean of weight for thinner width group do appear to generally be lower than wider width group ignoring the difference of species. For Perch wider width group, we have the highest mean-739.57 and standard deviation-263.64 and the sample size-23. For Bream wider width group, we have the second highest mean-653.29 and standard deviation-192.29 and the second highest sample size-31. For Bream thinner width group, we have the mean-344.00 and standard deviation-137.22 and the lowest sample size-3. For Perch thinner width group, we have the lowest mean-133.19 and the lowest standard deviation-64.58 but the highest sample size-33.

b) Obtain your best ANOVA model for weight as a function of species and width group and possibly the interaction between them. Show and explain your selection process for the model and be sure to use the correct SAS procedure. Comment on significance of the model and the individual terms in the model.

I choose this model with two predictor variabls and the intersection of these two variables. Notice that I used proc GLM instead of proc ANOVA in this question because while I used proc ANOVA, SAS gave me a warning that PROC ANOVA has determined that the number of observations in each cell is not equal and PROC GLM may be more appropriate.

Class Level Information		
Class	**Levels**	**Values**
Species	2	Bream Perch
widthgroup	2	thinner wider

Number of Observations Read	91
Number of Observations Used	90

Some basic information about the data is provided first, namely the classification variable(s), levels and values are 2 for the classification variable(s) and the amount of data read and used which is 91. Then analysis of variance tables and diagnostics follow.

Dependent Variable: Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6502167.798	2167389.266	66.34	<.0001
Error	86	2809594.458	32669.703		
Corrected Total	89	9311762.256			

In this example, the p-value is less than 0.0001, so the null model is rejected in favor of the group model. The variation described by differences across weight variable is significantly greater (from a statistical perspective) than expected due to chance.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
widthgroup	1	6281680.125	6281680.125	192.28	<.0001
Species	1	19616.986	19616.986	0.60	0.4405
Species*widthgroup	1	200870.687	200870.687	6.15	0.0151

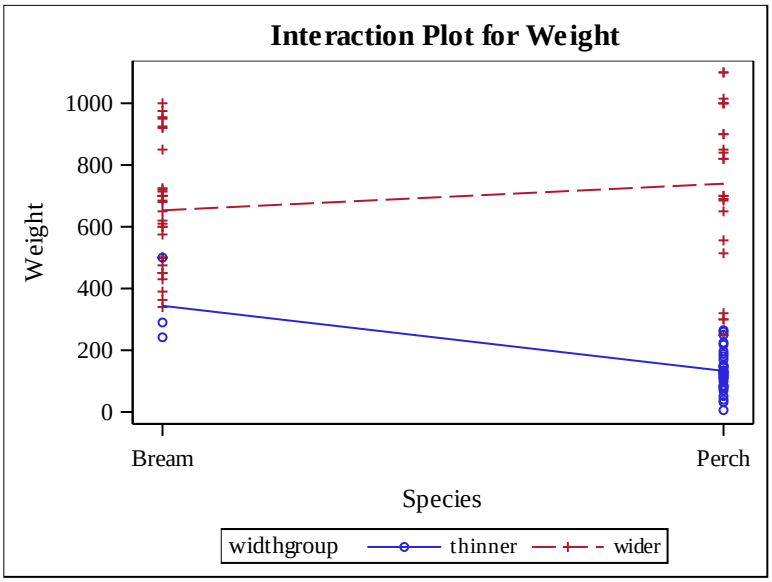
While the two terms together and the intersection are better than error only, the individual terms must be checked to see if their individual contributions are significant enough to warrant retention in the model. The type I sums of squares for the model follows. Remember that the type I sums of squares are sequential. Given that only a mean is in the model already, the model sum of squares would increase by 6281680.125 if origin is added. Given that a mean and width group are already included in the model, adding intersection to the model increases the model sum of squares by 200870.687. As before, the mean square for each source is divided by the error mean square to obtain the F statistics for that source. Note that, the species variable is not statistically significant in this case, but we still let it retain it in the model because we can't remove the intersection of the two variables since it is significantly significant.

The F statistics for width group are statistically significant. Adding width group to an error only model adds significantly more explained variation than expected due to chance and so width group should be retained if added first. Given that width group and species is already in the model, the amount of additional variation described by then adding the intersection is significantly greater than expected due to chance and so grade would be retained as well.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
widthgroup	1	1908257.358	1908257.358	58.41	<.0001
Species	1	35295.802	35295.802	1.08	0.3015
Species*widthgroup	1	200870.687	200870.687	6.15	0.0151

The type III sums of squares tell us about the amount of explained variation lost if a term is removed. If width group is removed from a model containing species and intersection, the amount of explained variation lost is only 1908257.358 as compared to the 6281680.125 shown in the type I sum of squares table. The type III sum of squares for the intersection is the same as type I in this case because in both types of sums of squares grade is the last source in, so adding it last or removing it first would have the same magnitude of change in the overall sum of squares.

As can be seen, the width group and intersection of type III sums of squares are both statistically significant. This indicates that removing either term for the model would result in a loss of explained variation significantly greater than expected due to chance, so neither term should be removed.



Additionally, by default proc glm provides an interaction plot here. The interaction plot shows the interaction in the model. In this case, there is no interaction in the model, so the lines are parallel. The colored symbols on the left and right are actual weight for the Bream and Perch breeds from the data. The legend at the bottom indicates the shapes and colors for grade levels.

The lines just help to guide the eye to make comparison. The expected number of days missed for a given origin and grade level is at one end of a line. The expected number of weights for the other species and the same width group level is at the other end of the line. When there is no interaction, the lines will be parallel.

c) For the model chosen in part b, comment on variation explained by the model, and any significant group differences (main effects and interactions if the interaction term is still in your model). What does this tell us about weight differences between species and/or wider or thinner fish?

R-Square	Coeff Var	Root MSE	Weight Mean
0.698275	38.10615	180.7476	474.3267

Looking at the R-Square value in the table that follows would provide some indication of practical significance. The R-Square value, which can be obtained by dividing the sum of squares for the model by the total sum of squares, tells the percentage of overall variation in the response described by the model. In this example, about 69.8% of the variation in weight could be described by two variables and the intersection. This is a moderate percentage, and it is a good start, adding additional explanatory variables to the model will help to describe more of the variation in whole weight values.

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

widthgroup	Weight LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
thinner	238.596970	<.0001
wider	696.427770	

Starting with the width group effect, a simple hypothesis test table can be obtained here as there are only two levels. Under the null, the expected number of widths is the same for thinner and wider group. The small p-value indicates that is unlikely to be true and suggests there is a significant difference between the expected number of weights of these two width groups.

Least Squares Means for Effect widthgroup				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-457.830800	-576.916926	-338.744674

The associated confidence interval for the difference of means indicates that the difference is statistically significant as 0 is not in the interval. The tables generated for the confidence intervals do not include significance indicators (***), so, the interval needs to be examined directly. The result indicates that tinner group were expected to be lighter of 457.83 on average with a 95% confidence interval of (-576.92, -338.74).

Least Squares Means

Adjustment for Multiple Comparisons: Tukey-Kramer

Species	Weight LSMEAN	H0:LSMean1=LSMean2
		Pr > |t|
Bream	498.645161	0.3015
Perch	436.379578	

Starting with the species effect, a simple hypothesis test table can be obtained here as there are only two levels. Under the null, the expected number of weights is the same for species group. The small p-value indicates that is unlikely to be true and suggests there is a significant difference between the expected number of weights of these two species groups.

Least Squares Means for Effect Species				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	62.265583	-56.820543	181.351709

The associated confidence interval for the difference of means indicates that the difference is not statistically significant as 0 is in the interval. The tables generated for the confidence intervals do not include significance indicators (***), so, the interval needs to be examined directly.

Least Squares Means

Adjustment for Multiple Comparisons: Tukey-Kramer

Species	widthgroup	Weight LSMEAN	LSMEAN Number
Bream	thinner	344.000000	1
Bream	wider	653.290323	2
Perch	thinner	133.193939	3
Perch	wider	739.565217	4

Least Squares Means for Effect Species*widthgroup				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-309.290323	-595.621779	-22.958866
1	3	210.806061	-74.758777	496.370898
1	4	-395.565217	-686.257302	-104.873133
2	3	520.096383	401.649606	638.543160
2	4	-86.274895	-216.598532	44.048743
3	4	-606.371278	-735.001851	-477.740705

With more levels for the categorical predictor or interaction a table with expected values and an index for referencing further comparisons is provided. In the confidence interval table, the species values and the width group will be indexed by the LSMEAN Number value from the LSMEAN table.

The bream and thinner group were expected to have weight of 344 on average, the average weight of Bream and wider group is around 653.29, the weight of Perch and thinner group is around 133.19 and the average weight of Perch and wider group is around 739.56 .The observed significant difference is for Bream thinner and Bream wider. Bream thinner was expected to be lighter than Bream wider and the associated confidence interval does not include 0, so the difference is statistically significant. Bream thinner was expected to be lighter than Perch wider and the associated confidence interval does not include 0, so the difference is statistically significant. Bream wider was expected to be heavier than Perch thinner and the associated confidence interval does not include 0, so the difference is statistically significant. Perch thinner was expected to be lighter than Perch wider and the associated confidence interval does not include 0, so the difference is statistically significant.

Exercise 2

We might expect that fish would have a certain amount of symmetry. Consider modeling the weight as a function of the length measurement length1.

a) Fit a linear regression model for weight as a function of length1 ignoring species. If any points are unduly influential, note and remove points until undue influence is removed, and refit the model.

There is a point there is unduly influential while we first fit in the model which we can look at the Cook's Distance graph on Diagnostics panel. I created the new data set which the value of cook distance attached, while I filter out the unduly influential point and refit the model again, below is the new model.

b) Comment on the quality of the final model, significance of the parameters, and any remaining issues noted in the diagnostics. What does this model tell us about the relationship between weight and length for the types of fish included in this data?

Model: MODEL1
Dependent Variable: Weight

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	8410364	8410364	1076.81	<.0001
Error	87	679509	7810.44845		
Corrected Total	88	9089873			

In this example, the p-value is less than 0.0001, so the null model is rejected in favor of the one-way length1 group model. The variation described by differences across weight variable is significantly greater (from a statistical perspective) than expected due to chance.

Root MSE	88.37674	**R-Square**	0.9252
Dependent Mean	479.58989	**Adj R-Sq**	0.9244
Coeff Var	18.42757		

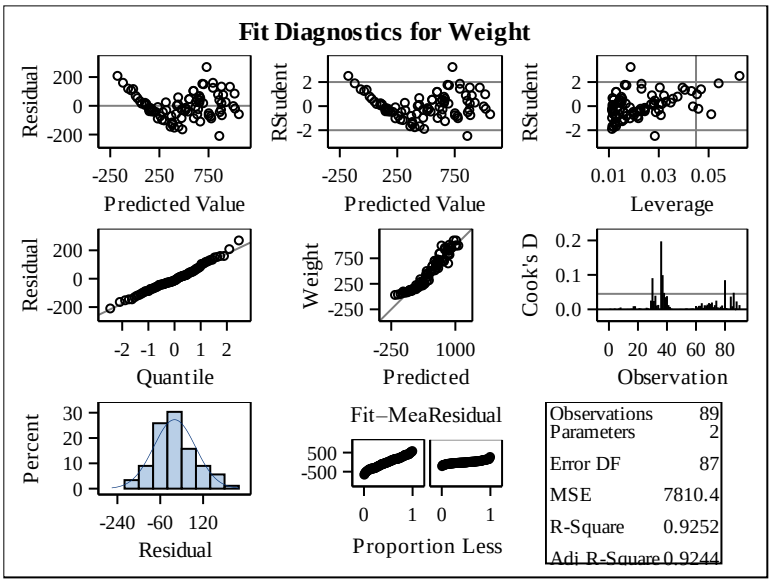
The results for fitting a simple linear regression model of weight as a function of length1 follow. The model is statistically significant and describes just over 92.52% of the variation in length1 related weights.

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-715.99725	37.61947	-19.03	<.0001
Length1	1	43.16899	1.31554	32.81	<.0001

The parameter estimates are included in the following table, and the t statistics provided are the basis for determining statistical significance. The intercept $\beta 0$ is estimated to be about -715.99725 and the slope $\beta 1$ is estimated to be 43.16889, and both are statistically significant as indicated by the p-values less than 0.05.

The slope indicates an expected increase of about 43.17 weight for each increase of one length 1.The statistically significant negative intercept term might seem to suggest that in the absence of length1 and has a negative weight. This of course is not true and is also not a problem with the model. The intercept again, only provides an anchor point.

Model: MODEL1
Dependent Variable: Weight



Further, some diagnostics should be checked to see if there are any issues with assumptions of the model or undue influence from observations. The following diagnostic panel is generated by default and is very useful for assessing issues with normality, non-constant variance, and undue influence. We will focus on the Residual and RStudent (Studentized residual) vs. Predicted plots in the first row, the quantile plot in the second row and histogram in the third row, and the Cook's D plot in the second row.

The residuals vs. predicted and Studentized residuals vs. predicted value plots can be used to assess the equal variance assumption. We can see that there is a curve in the first and second plot, that indicates that there is no linearity. It is also spread out in this case meaning that there are no constant variances. The Studentized residuals are scaled to behave like a standard normal, so roughly 95% of the residuals should fall between the two lines drawn.

The histogram and quantile plot look reasonable in this case given the sample size. There does appear to be one point that is a bit farther off at the top end and at the bottom end.

The Cook's D plot in this case indicates that all the Cook distance of the data retained in the model looks fine since we already removed the unduly high influential point at the first place while we fit the model. The plot shows the value of Cook's distance for each of the observations in order. A line is also drawn for a very aggressive threshold for considering removal. In this case the horizontal line is around 0.05. If points have Cook's distances above that line but below 1, the value should be compared to the other Cook's distances. If the value is not too far about the line and there are other observations with Cook's distances that are somewhat close in magnitude, the point could be retained. If there is a big difference, removal should be considered. With Cook's distances, removing a point will cause the Cook's distances for all the remaining points to change because the measure is a leave-one-out diagnostic. Once a point is removed, the remaining points are effectively anew smaller data set. For this reason, removal based on Cook's distance should be done one point at a time.

Exercise 3
Now consider modeling with the other possible continuous predictors for weight as well.
a) Perform and explain model selection for a linear regression model of weight as a function of the other continuous variables in the data set. Consider the additional contribution to the model $R2$ when determining whether to keep significant terms with high variance inflation. If any points are unduly influential, note and remove points until undue influence is removed, and refit the model

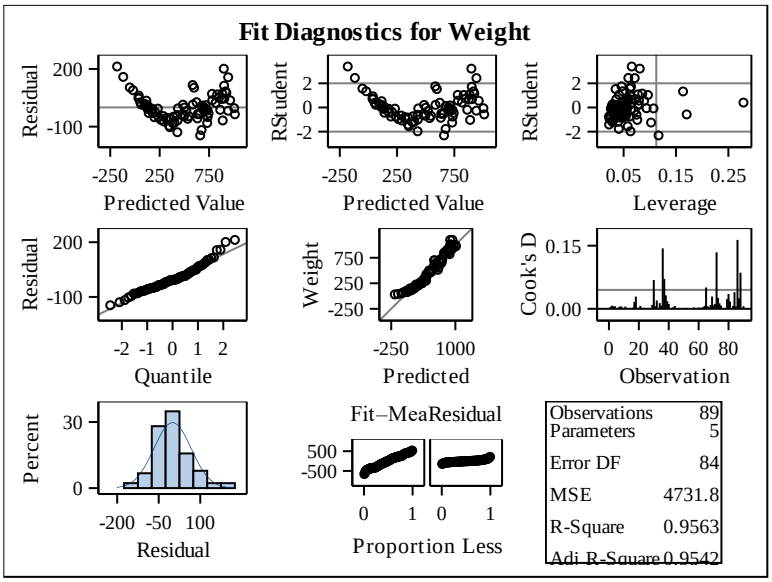
Model: MODEL1
Dependent Variable: Weight

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Length2		1	0.9334	0.9334	42.3533	1219.43	<.0001
2	Height		2	0.0089	0.9423	27.3022	13.29	0.0005
3	Width		3	0.0095	0.9518	11.2289	16.66	0.0001
4	Length3		4	0.0045	0.9563	4.6222	8.65	0.0042

Model: MODEL1
Dependent Variable: Weight

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-621.85740	36.15476	-17.20	<.0001	0
Length2	1	72.56429	18.77869	3.86	0.0002	380.96563
Length3	1	-57.89679	19.69055	-2.94	0.0042	543.85791
Height	1	45.98252	11.03597	4.17	<.0001	42.65531
Width	1	67.92013	20.27507	3.35	0.0012	16.31718

Model: MODEL1
Dependent Variable: Weight



The first step I did was doing a stepwise selection and output a Variance inflation table. Not only that, but I also remove the unduly influential point while doing the operation above. We can tell from the Cook's D graph in the Diagnostics Panel that there is not unduly high influential point. While the pairwise selection help us to eliminate the Length 1 variable already, the remain variables still have high VIF values, therefore we need to keep remove the variables from the model. We first notice that Length3 has the highest VIF values and the lowest partial R-Square therefore we remove it from the model.

Model: MODEL1
Dependent Variable: Weight

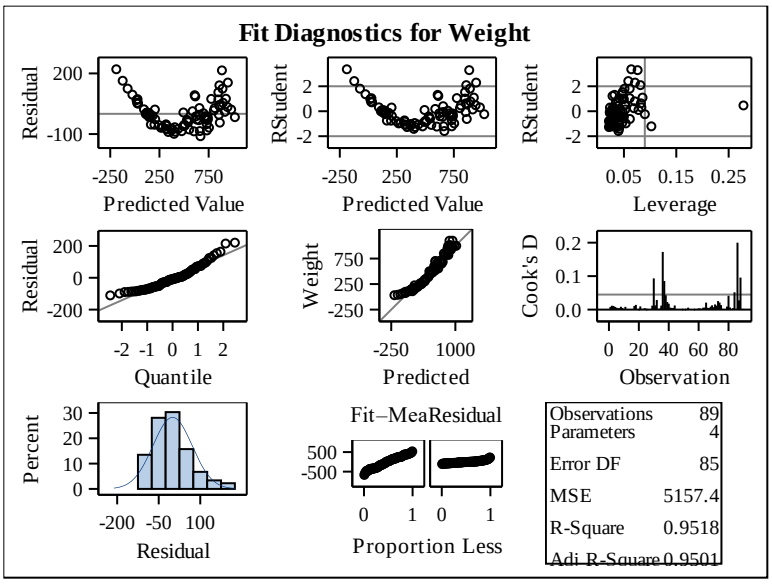
Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Length2		1	0.9334	0.9334	32.3714	1219.43	<.0001
2	Height		2	0.0089	0.9423	18.6567	13.29	0.0005
3	Width		3	0.0095	0.9518	4.0000	16.66	0.0001

Model: MODEL1
Dependent Variable: Weight

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-664.36705	34.59661	-19.20	<.0001	0
Length2	1	18.78804	4.44723	4.22	<.0001	19.60325
Height	1	14.62578	2.96475	4.93	<.0001	2.82437
Width	1	83.41873	20.43945	4.08	0.0001	15.21434

Model: MODEL1

Dependent Variable: Weight



The next step I did was doing a stepwise selection and output a Variance inflation table. Not only that, but I also remove the unduly influential point while doing the operation above. We can tell from the Cook's D graph in the Diagnostics Panel that there is not unduly high influential point. While the pairwise selection help us to elimate the Length 3 variable already, the remain variables still have high VIF values, therefore we need to keep remove the variables from the model. We first notice that Width has the second highest VIF values and the second lowest partial R-Square therefore we remove it from the model.

b) Comment on the quality of the final model, significance of the parameters, and any remaining issues noted in the diagnostics. What does this model tell us about the relationship between weight and the dimensions for the types of fish included in this data?

Model: MODEL1
Dependent Variable: Weight

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Length2		1	0.9334	0.9334	14.2944	1219.43	<.0001
2	Height		2	0.0089	0.9423	3.0000	13.29	0.0005

Model: MODEL1
Dependent Variable: Weight

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-716.50991	34.95546	-20.50	<.0001	0
Length2	1	35.68641	1.76461	20.22	<.0001	2.61102
Height	1	11.30016	3.09921	3.65	0.0005	2.61102

The final model looks good because the VIF value of variables are not over 10 therefore this is my final model. Not only that, but I also remove the unduly influential point while doing the operation above. We can tell from the Cook's D graph in the Diagnostics Panel that there is not unduly high influential point.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	8565590	4282795	702.52	<.0001
Error	86	524283	6096.31523		
Corrected Total	88	9089873			

In this example, the p-value is less than 0.0001, so the null model is rejected in favor of the length2 and height model. The variation described by differences across weight variable is significantly greater (from a statistical perspective) than expected due to chance.

Root MSE	78.07890	**R-Square**	0.9423
Dependent Mean	479.58989	**Adj R-Sq**	0.9410
Coeff Var	16.28035		

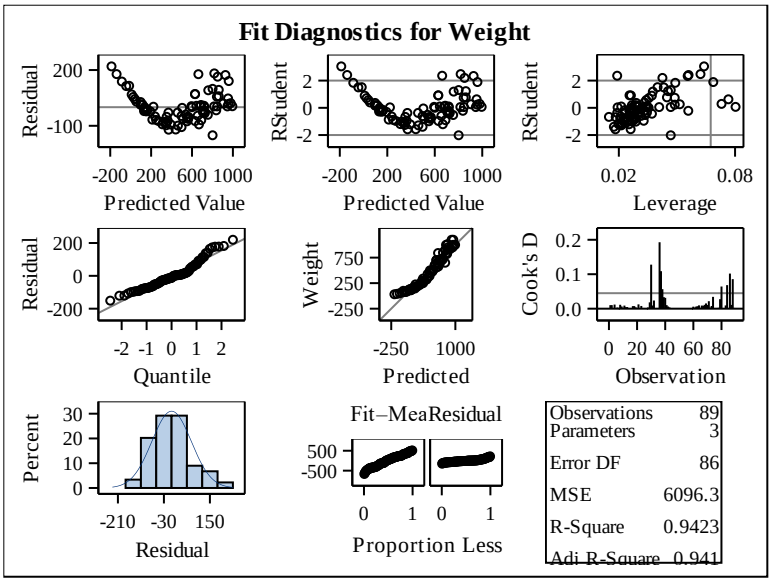
The results for fitting a simple linear regression model of weight as a function of length2 and height follow. The model is statistically significant and describes just over 94.23% of the variation in length2 and height related weights.

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-716.50991	34.95546	-20.50	<.0001	0
Length2	1	35.68641	1.76461	20.22	<.0001	2.61102
Height	1	11.30016	3.09921	3.65	0.0005	2.61102

The parameter estimates are included in the following table, and the t statistics provided are the basis for determining statistical significance. The intercept $\beta 0$ is estimated to be about -716.50991 and the slope $\beta 1$ is estimated to be 35.68641, the slope $\beta 2$ is estimated to be 11.30016, and all of them are statistically significant as indicated by the p-values less than 0.05.

The slope indicates an expected increase of about 35.68641 weight for each increase of one length 2. The slope indicates an expected increase of about 11.30016 weight for each increase of one height. The statistically significant negative intercept term might seem to suggest that in the absence of length2 and height and has a negative weight. This of course is not true and is also not a problem with the model. The intercept again, only provides an anchor point.

Model: MODEL1
Dependent Variable: Weight



Further, some diagnostics should be checked to see if there are any issues with assumptions of the model or undue influence from observations. The following diagnostic panel is generated by default and is very useful for assessing issues with normality, non-constant variance, and undue influence. We will focus on the Residual and RStudent (Studentized residual) vs. Predicted plots in the first row, the quantile plot in the second row and histogram in the third row, and the Cook's D plot in the second row.

The residuals vs. predicted and Studentized residuals vs. predicted value plots can be used to assess the equal variance assumption. We can see that there is a curve in the first and second plot, that indicates that there is no linearity. It is also spread out in this case meaning that there are no constant variances. The Studentized residuals are scaled to behave like a standard normal, so roughly 95% of the residuals should fall between the two lines drawn.

The histogram looks reasonable in this case given the sample size, but the quantile plot seems a little bit curvature in the middle, therefore we reject the normal assumption. There does appear to be one point that is a bit farther off at the top end and at the bottom end.

The Cook's D plot in this case indicates that all the Cook distance of the data retained in the model looks fine since we already removed the unduly high influential point at the first place while we fit the model. The plot shows the value of Cook's distance for each of the observations in order. A line is also drawn for a very aggressive threshold for considering removal. In this case the horizontal line is around 0.05. If points have Cook's distances above that line but below 1, the value should be compared to the other Cook's distances. If the value is not too far about the line and there are other observations with Cook's distances that are somewhat close in magnitude, the point could be retained. If there is a big difference, removal should be considered. With Cook's distances, removing a point will cause the Cook's distances for all the remaining points to change because the measure is a leave-one-out diagnostic. Once a point is removed, the remaining points are effectively anew smaller data set. For this reason, removal based on Cook's distance should be done one point at a time.

Exercise 4

It might be expected that volume (product of length, width, and height) might be more related to weight than a linear combination of the 3 dimensions. A log-linear model with log(weight) as the response, and logs of length, width and height measurements would be consistent with such a relationship. Add variables for these log measurements to the data set and repeat Exercise 3 using the log of weight as the response and the other log variables as possible predictors. Also comment on whether this model or the one obtained in Exercise 3 is a better model.

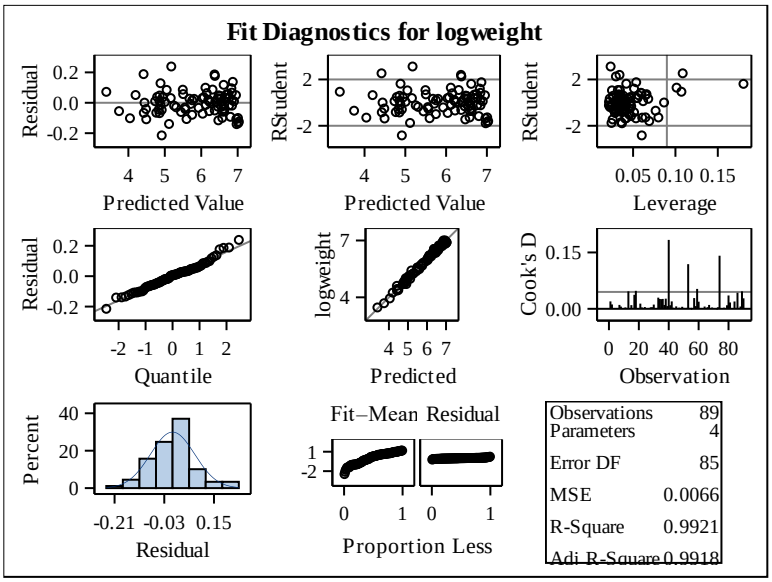
Model: MODEL1
Dependent Variable: logweight

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	loglength3		1	0.9854	0.9854	68.8280	5878.37	<.0001
2	logwidth		2	0.0042	0.9896	26.7967	34.49	<.0001
3	logheight		3	0.0016	0.9912	11.5220	15.87	0.0001
4	loglength1		4	0.0009	0.9921	4.0184	9.62	0.0026
5		loglength3	3	0.0000	0.9921	2.0676	0.05	0.8239

Model: MODEL1
Dependent Variable: logweight

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-1.78499	0.27858	-6.41	<.0001	0
loglength1	1	1.53822	0.14842	10.36	<.0001	22.64273
logheight	1	0.60896	0.03745	16.26	<.0001	3.71057
logwidth	1	0.75908	0.12262	6.19	<.0001	18.69628

Model: MODEL1
Dependent Variable: logweight



The first step I did was doing a stepwise selection and output a Variance inflation table. The table told us the variables that add to the model and the variable that get removed. Not only that, but I also remove the unduly influential point while doing the operation above. We can tell from the Cook's D graph in the Diagnostics Panel that there is not unduly high influential point. While the pairwise selection help us to eliminate the Length 1 variable already, the remain variables still have high VIF values, therefore we need to keep remove the variables from the model. We first notice that log length1 has the highest VIF values and the lowest partial R-Square therefore we remove it from the model.

Model: MODEL1
Dependent Variable: logweight

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	logwidth		1	0.9338	0.9338	234.228	1227.72	<.0001
2	logheight		2	0.0483	0.9822	3.0000	233.23	<.0001

Model: MODEL1
Dependent Variable: logweight

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	1.05999	0.07099	14.93	<.0001	0
logheight	1	0.77410	0.05069	15.27	<.0001	3.03871
logwidth	1	1.92202	0.07394	25.99	<.0001	3.03871

The final model looks good because the VIF value of variables are not over 10 therefore this is my final model. Not only that, but I also remove the unduly influential point while doing the operation above. We can tell from the Cook's D graph in the Diagnostics Panel that there is not unduly high influential point.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	70.06493	35.03247	2369.05	<.0001
Error	86	1.27173	0.01479		
Corrected Total	88	71.33666			

In this example, the p-value is less than 0.0001, so the null model is rejected in favor of the logheight and logweight model. The variation described by differences across weight variable is significantly greater (from a statistical perspective) than expected due to chance.

Root MSE	0.12160	**R-Square**	0.9822
Dependent Mean	5.85332	**Adj R-Sq**	0.9818
Coeff Var	2.07752		

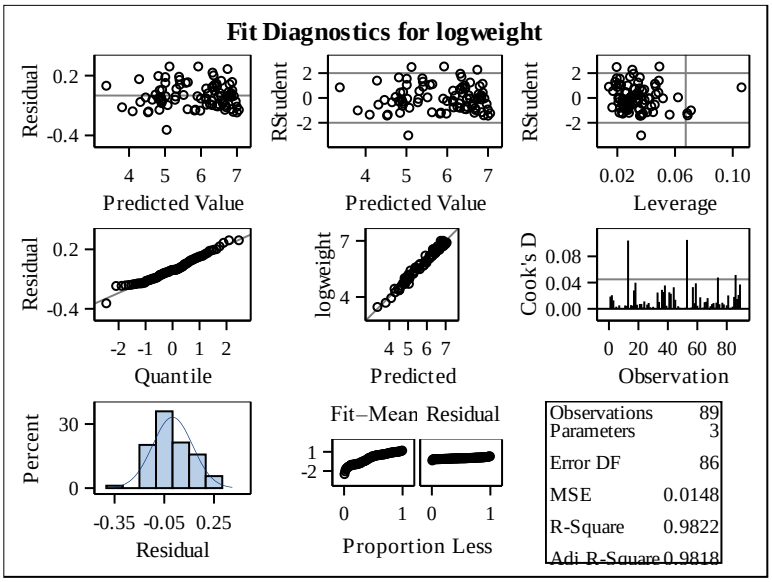
The results for fitting a simple linear regression model of weight as a function of logheight and logweight follow. The model is statistically significant and describes just over 98.22% of the variation in logheight and logweight related weights.

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	1.05999	0.07099	14.93	<.0001	0
logheight	1	0.77410	0.05069	15.27	<.0001	3.03871
logwidth	1	1.92202	0.07394	25.99	<.0001	3.03871

The parameter estimates are included in the following table, and the t statistics provided are the basis for determining statistical significance. The intercept $\beta 0$ is estimated to be about 1.0599 and the slope $\beta 1$ is estimated to be 0.77410, the slope $\beta 2$ is estimated to be 1.92202, and all of them are statistically significant as indicated by the p-values less than 0.05.

The slope indicates an expected increase of about 1.0599 weight for each increase of one logheight. The slope indicates an expected increase of about 1.92202 weight for each increase of one logwidth. The statistically significant postitive intercept term might seem to suggest that in the absence of logwidth and logwidth and has a positive weight. This of course is not true and is also not a problem with the model. The intercept again, only provides an anchor point.

Model: MODEL1
Dependent Variable: logweight



Further, some diagnostics should be checked to see if there are any issues with assumptions of the model or undue influence from observations. The following diagnostic panel is generated by default and is very useful for assessing issues with normality, non-constant variance, and undue influence. We will focus on the Residual and RStudent (Studentized residual) vs. Predicted plots in the first row, the quantile plot in the second row and histogram in the third row, and the Cook's D plot in the second row.

The residuals vs. predicted and Studentized residuals vs. predicted value plots can be used to assess the equal variance assumption. These plots should look relatively flat with no obvious increasing, decreasing or other trends as the predicted values increase. The Studentized residuals are scaled to behave like a standard normal, so roughly 95% of the residuals should fall between the two lines drawn.

The histogram looks reasonable in this case given the sample size, but the quantile plot seems a little bit curvature in the end, therefore we reject the normal assumption. There does appear to be one point that is a bit farther off at the top end and at the bottom end.

The Cook's D plot in this case indicates that all the Cook distance of the data retained in the model looks fine since we already removed the unduly high influential point at the first place while we fit the model. The plot shows the value of Cook's distance for each of the observations in order. A line is also drawn for a very aggressive threshold for considering removal. In this case the horizontal line is around 0.05. If points have Cook's distances above that line but below 1, the value should be compared to the other Cook's distances. If the value is not too far about the line and there are other observations with Cook's distances that are somewhat close in magnitude, the point could be retained. If there is a big difference, removal should be considered. With Cook's distances, removing a point will cause the Cook's distances for all the remaining points to change because the measure is a leave-one-out diagnostic. Once a point is removed, the remaining points are effectively anew smaller data set. For this reason, removal based on Cook's distance should be done one point at a time.

In conclusion, this model is better than model Exercise 3 because it describes just over 98.22% of the variation compared to 94.23% in the model Exercise 3. Not only that, the model in Exercise 4 suggest that they have linearity and constant variance compare to model in Exercise 3 which doesn't have linearity and constant variance.