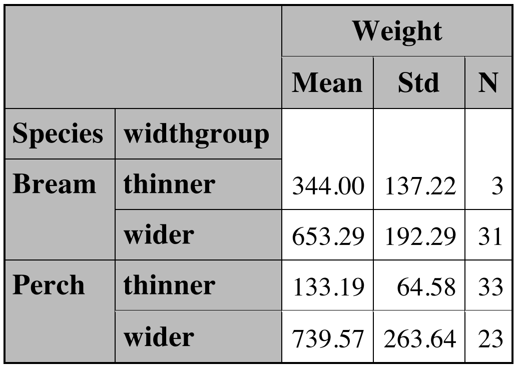
Homework 3

Jiangyan Feng (NetID: jf8)

Exercise 1

a)



		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 cross-tabulation table is shown as above. The interesting features are:

1. For Bream species, thinner fish is much less compared with the wider fish. And the weight difference between thinner and wider group is ~300, relatively large.
2. For Perch species, in contrast to Bream species, wider fish is more than the thinner fish. And the weight difference between thinner and wider group is ~600, which is much larger than the difference observed within Bream species.
3. The count for each cell is different, which indicates the unbalanced data.
4. Comparing the Bream and Perch species, the mean weight for thinner fish in Bream is larger than that of the Perch. By contrast, the mean weight for wider fish in Bream is smaller than that of the Perch.

b) According to part a, the data is unbalanced. Therefore, glm procedure is used. The selection process is:

1. Build main effects of species and widthgroup to build model.

For the model, the F value and P value is 91.05 and <0.0001, respectively. This indicates the model is statistically significant.

The resulting R^2 is 0.676703. According to the Type 3 SS, the P value for species and widthgroup is 0.4535 and <0.0001, respectively. This indicates that species may not be needed.

Dependent Variable: Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6301297.111	3150648.555	91.05	<.0001
Error	87	3010465.145	34603.048		
Corrected Total	89	9311762.256			

R-Square	Coeff Var	Root MSE	Weight Mean
0.676703	39.21748	186.0189	474.3267

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Species	1	1257048.002	1257048.002	36.33	<.0001
widthgroup	1	5044249.108	5044249.108	145.77	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Species	1	19616.986	19616.986	0.57	0.4535
widthgroup	1	5044249.108	5044249.108	145.77	<.0001

2. Switch the order of species and widthgroup and use the main effects to build model.
For the model, the F value and P value is 181.54 and <0.0001, respectively. This indicates the model is statistically significant.
The resulting R^2 is 0.676703. According to the Type 3 SS, the P value for widthgroup and species is <0.0001 and 0.4535, respectively. This indicates that species may not be needed.

Comparing this model with the model built in step 1: although F value is larger than model 1, R^2 value is exactly the same. Type 3 SS values in both models indicate that species may not be needed for the model construction.

Dependent Variable: Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6301297.111	3150648.555	91.05	<.0001
Error	87	3010465.145	34603.048		
Corrected Total	89	9311762.256			

R-Square	Coeff Var	Root MSE	Weight Mean
0.676703	39.21748	186.0189	474.3267

Source	DF	Type I SS	Mean Square	F Value	Pr > F
widthgroup	1	6281680.125	6281680.125	181.54	<.0001
Species	1	19616.986	19616.986	0.57	0.4535

Source	DF	Type III SS	Mean Square	F Value	Pr > F
widthgroup	1	5044249.108	5044249.108	145.77	<.0001
Species	1	19616.986	19616.986	0.57	0.4535

3. Add interactions and build the model.
For the model, the F value and P value is 66.34 and <0.0001, respectively. This indicates the model is statistically significant.
The resulting R^2 is 0.698275. According to the Type 3 SS, the P value for species, widthgroup, species*widthgroup is 0.3015, <0.0001, 0.0151, respectively. This indicates that species may not be needed.

Comparing this model with the model built in step 1 and step 2:
R^2 value increases from 0.676703 to 0.698275. Type 3 SS values in all three models indicate that species may not be needed for the model construction. However, Type 3 SS for species*widthgroup shows the interaction between species and widthgroup is significant. Therefore, we decide to keep species term. And use the main effects and interactions of species and widthgroup to build the best model.

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			

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

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Species	1	1257048.002	1257048.002	38.48	<.0001
widthgroup	1	5044249.108	5044249.108	154.40	<.0001
Species*widthgroup	1	200870.687	200870.687	6.15	0.0151

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

c)

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			

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

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Species	1	1257048.002	1257048.002	38.48	<.0001
widthgroup	1	5044249.108	5044249.108	154.40	<.0001
Species*widthgroup	1	200870.687	200870.687	6.15	0.0151

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

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	

Species	Weight LSMEAN	95% Confidence Limits	
Bream	498.645161	390.016960	607.273363
Perch	436.379578	387.579813	485.179343

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

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

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

widthgroup	Weight LSMEAN	95% Confidence Limits	
thinner	238.596970	130.259607	346.934332
wider	696.427770	646.985692	745.869848

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

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
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: Weight

i/j	1	2	3	4
1		0.0290	0.2216	0.0033
2	0.0290		<.0001	0.3124
3	0.2216	<.0001		<.0001
4	0.0033	0.3124	<.0001	

Species	widthgroup	Weight LSMEAN	95% Confidence Limits	
Bream	thinner	344.000000	136.549745	551.450255
Bream	wider	653.290323	588.755555	717.825091
Perch	thinner	133.193939	70.645334	195.742545
Perch	wider	739.565217	664.642990	814.487445

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

The analysis results for the model chosen is shown as above.
R^2 is 0.698275, indicating that ~69.83% of variations are explained by this model.
According to the Tukey-Kramer test for the least squares means:

1. Pvalue for comparing Bream and Perch species is: 0.3015. 0 is within the 95% confidence ranges [-56.820543, 181.351709]. These results indicate there is no significant difference between these two species.

2. Pvalue for comparing thinner and wider widthgroup is: <0.0001. 0 not inside the 95% confidence ranges [-576.916926, -338.744674]. Differences between Means is -457.830800.

These results indicate there is significant difference between these two widthgroups. And considering the sign and magnitude of the difference, the weight for the thinner group is smaller than that for the wider group by ~458 on average.

3. For the interaction term, P value for the comparing four different groups are organized in the above table.

There are four comparisons are showing significant group differences:

Bream and thinner vs. Bream and wider: Pvalue = 0.0290

Bream and thinner vs. Perch and wider: Pvalue = 0.0033

Bream and wider vs. Perch and thinner: Pvalue = <0.0001

Perch and thinner vs. Perch and wider: Pvalue = <0.0001

All the other comparisons are not significantly different.

Exercise 2

a) The process of building the linear regression model is:

1. We first fit the linear regression model using all the data. The results are shown as below.

F value and P value for the model is 907.19 and <0.0001, respectively. This shows the model is statistically significant.

R-Square is 0.9116, showing that the model is also practically significant.

The parameters for intercept and length1 is -667.14554 and 41.55173. The P values for these two are <0.0001 and <0.0001. This shows statistical significance. Moreover, it shows that with 1 unit increase for length1, weight is expected to increase 41.55173.

The residual vs. predicted value distribution looks flat around the 0. And only three points out of the 2 standard deviations. These findings suggest our assumption of normal distribution is satisfied.

Residual vs. Quantile line fits well to the straight line, showing the normal distribution.

The cook's D shows there is a single unduly influential point, which should be removed from the data. The detailed information about this data point is also shown as blow.

Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	91
Number of Observations Used	90
Number of Observations with Missing Values	1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	8488366	8488366	907.19	<.0001
Error	88	823396	9356.77420		
Corrected Total	89	9311762			

Root MSE	96.73042	**R-Square**	0.9116
Dependent Mean	474.32667	**Adj R-Sq**	0.9106
Coeff Var	20.39321		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-667.14554	39.24568	-17.00	<.0001
Length1	1	41.55173	1.37956	30.12	<.0001

Obs	widthgroup	Species	Weight	Length1	Length2	Length3	Height	Width	cd
36	thinner	Perch	5.9	7.5	8.4	8.8	2.112	1.408	0.78126

2. We then find and remove the single point with large cook's D. And refit the model using the new data. The results are shown as below.

F value and P value for the model is 1076.81 and <0.0001, respectively. This shows the model is statistically significant.

R-Square is 0.9252, showing that the model is also practically significant.

The parameters for intercept and length1 is -715.99725 and 43.16899. The P values for these two are <0.0001 and <0.0001. This shows statistical significance. Moreover, it shows that with 1 unit increase for length1, weight is expected to increase 43.16899.

The residual vs. predicted value distribution looks flat around the 0. And only three points out of the 2 standard deviations. These findings suggest our assumption of normal distribution is satisfied.

Residual vs. Quantile line fits well to the straight line, showing the normal distribution.

The cook's D shows there are few points beyond the line. However, the difference between these points is small. Therefore, we stop removing any data points.

Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	90
Number of Observations Used	89
Number of Observations with Missing Values	1

Analysis of Variance

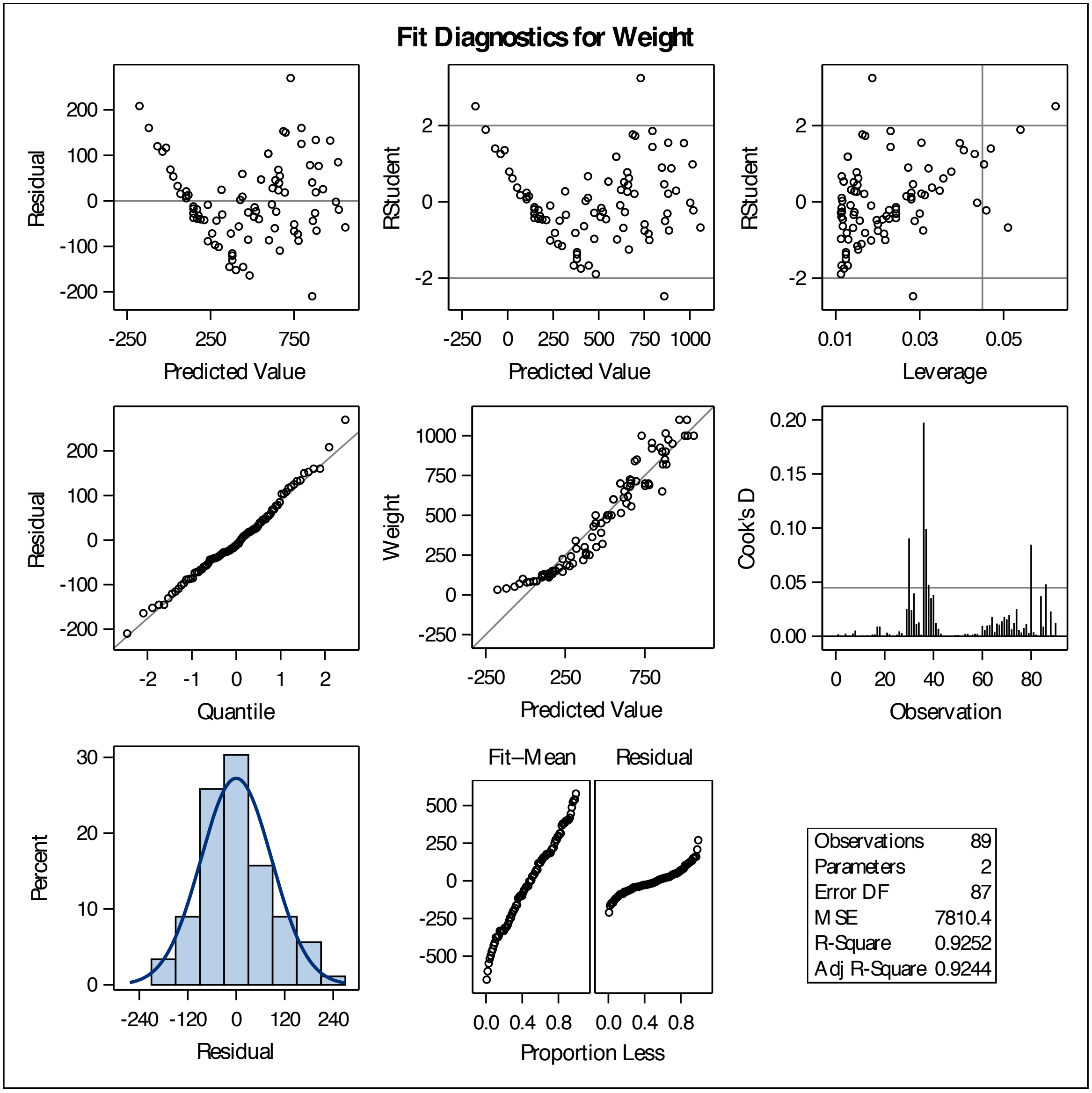
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8410364	8410364	1076.81	<.0001
Error	87	679509	7810.44845		
Corrected Total	88	9089873			

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

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

Model: MODEL1
Dependent Variable: Weight



b)

Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	90
Number of Observations Used	89
Number of Observations with Missing Values	1

Analysis of Variance

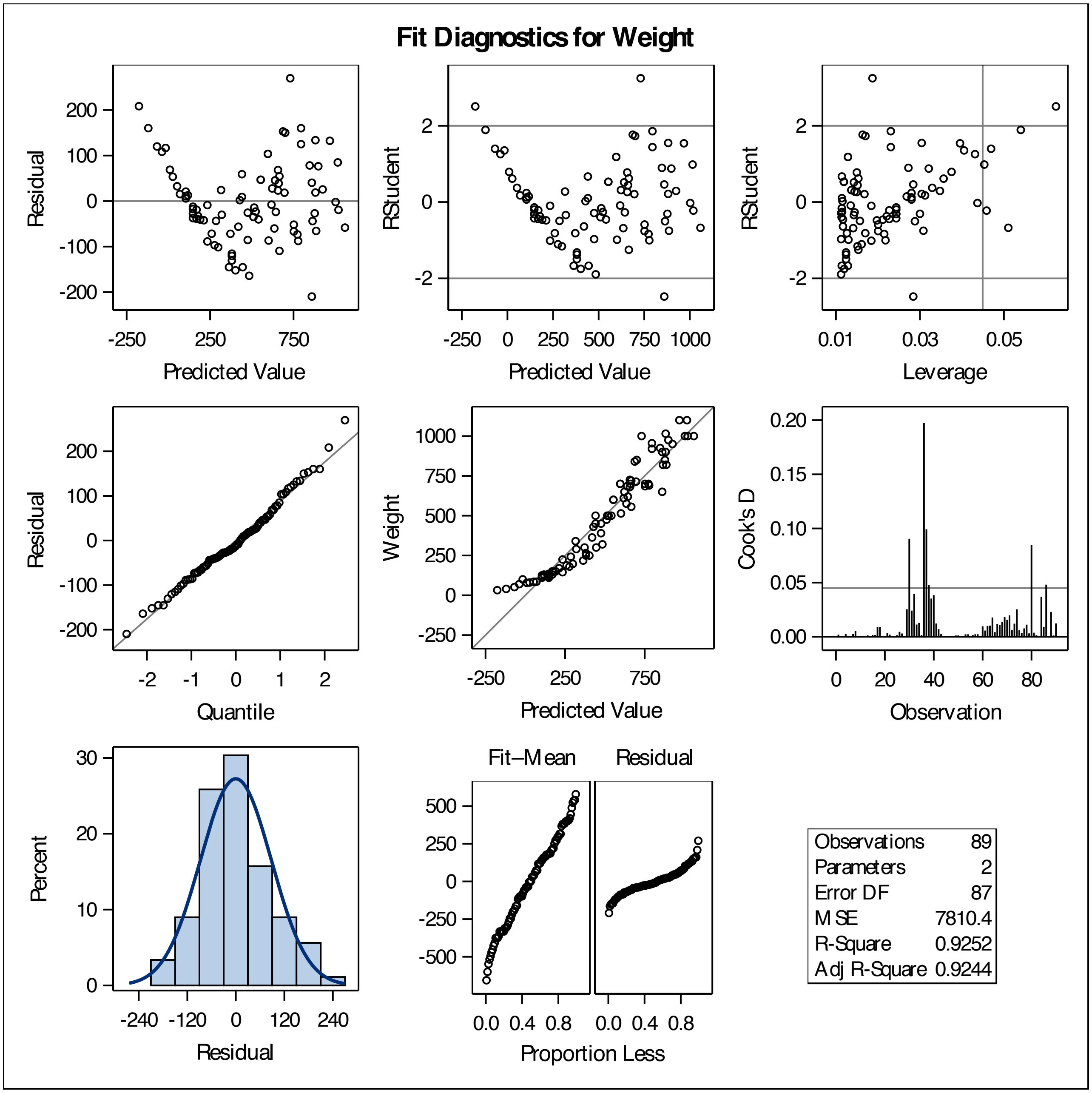
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8410364	8410364	1076.81	<.0001
Error	87	679509	7810.44845		
Corrected Total	88	9089873			

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

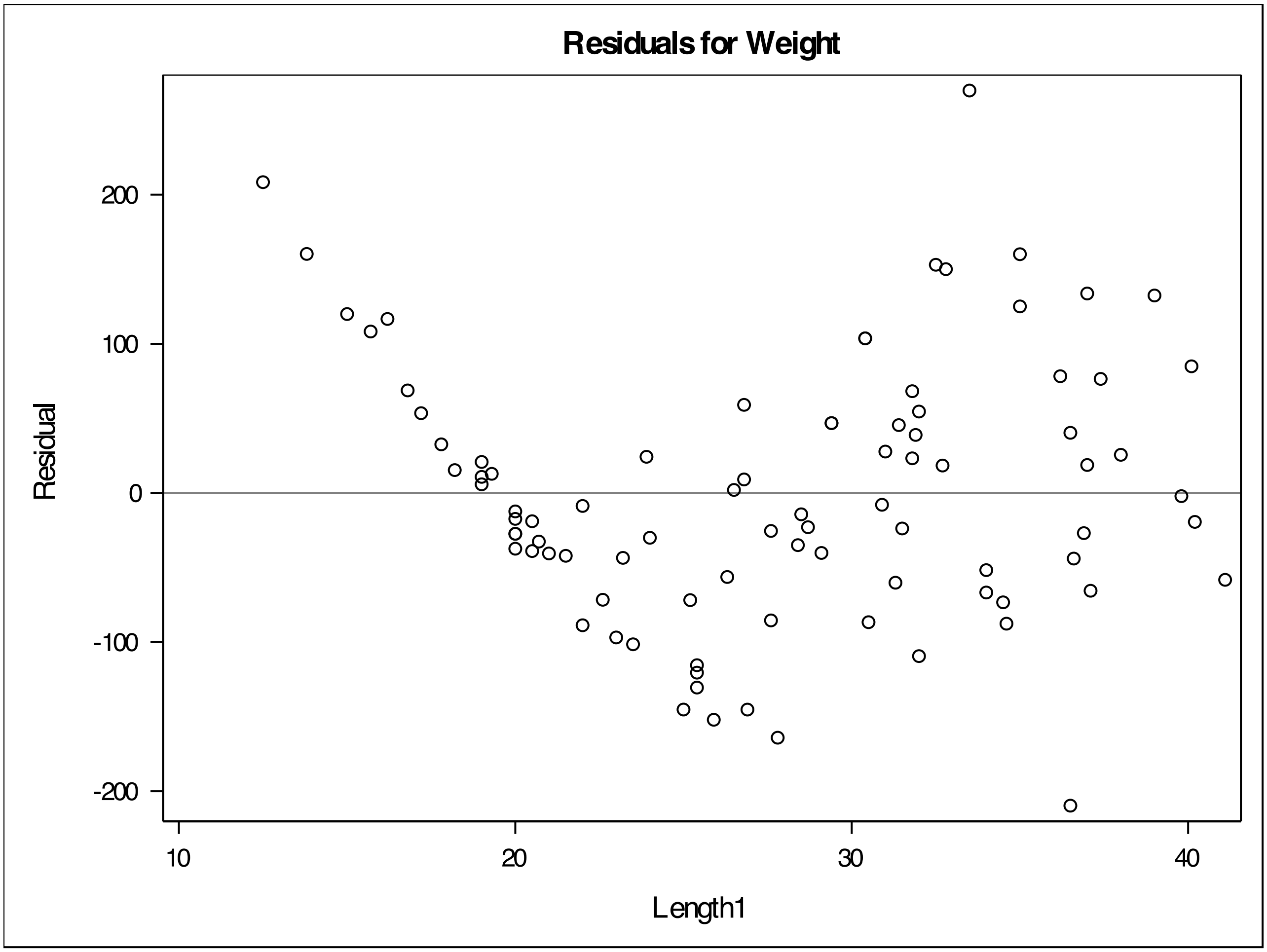
Parameter Estimates

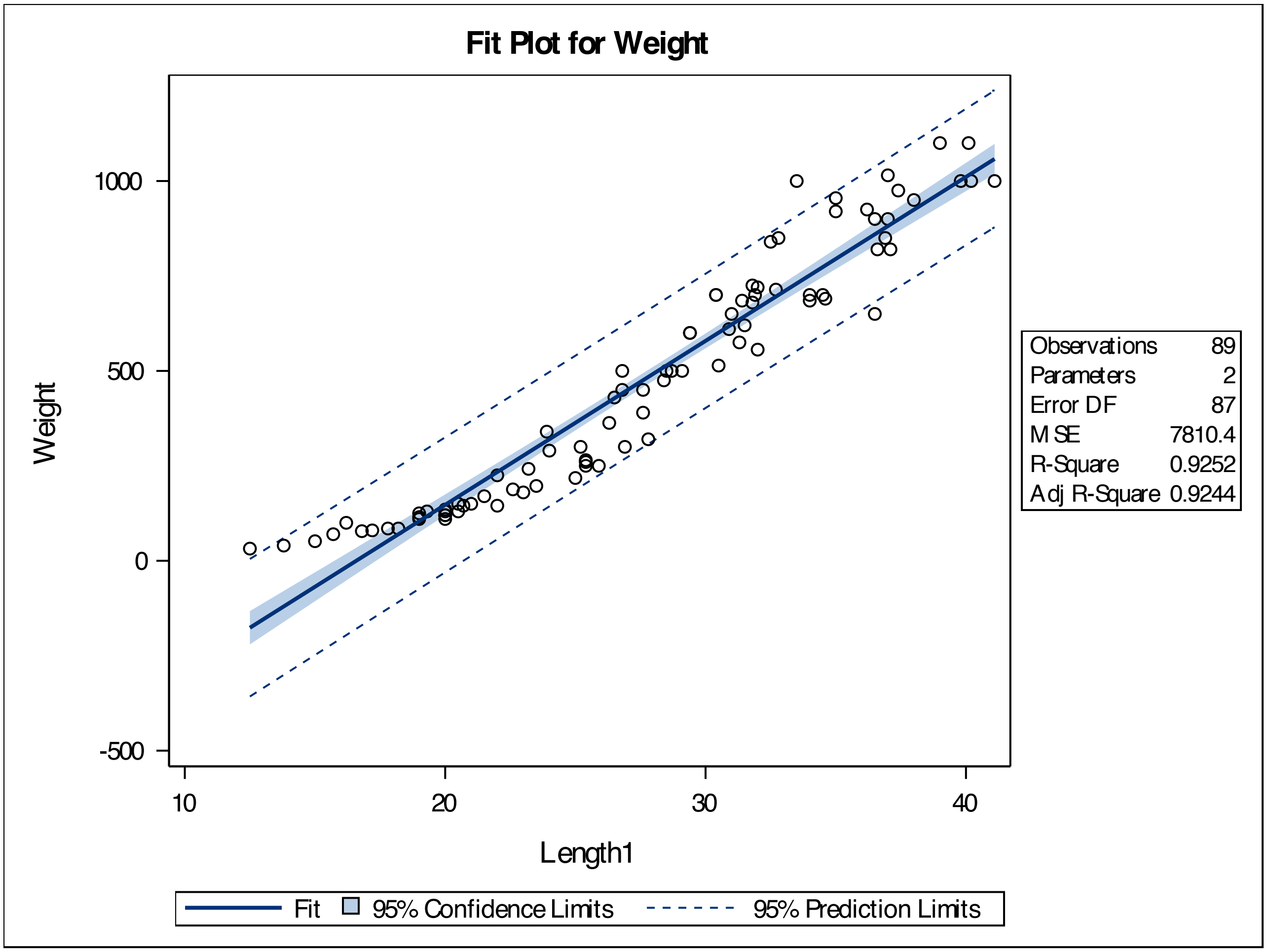
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-715.99725	37.61947	-19.03	<.0001
Length1	1	43.16899	1.31554	32.81	<.0001

Model: MODEL1
Dependent Variable: Weight



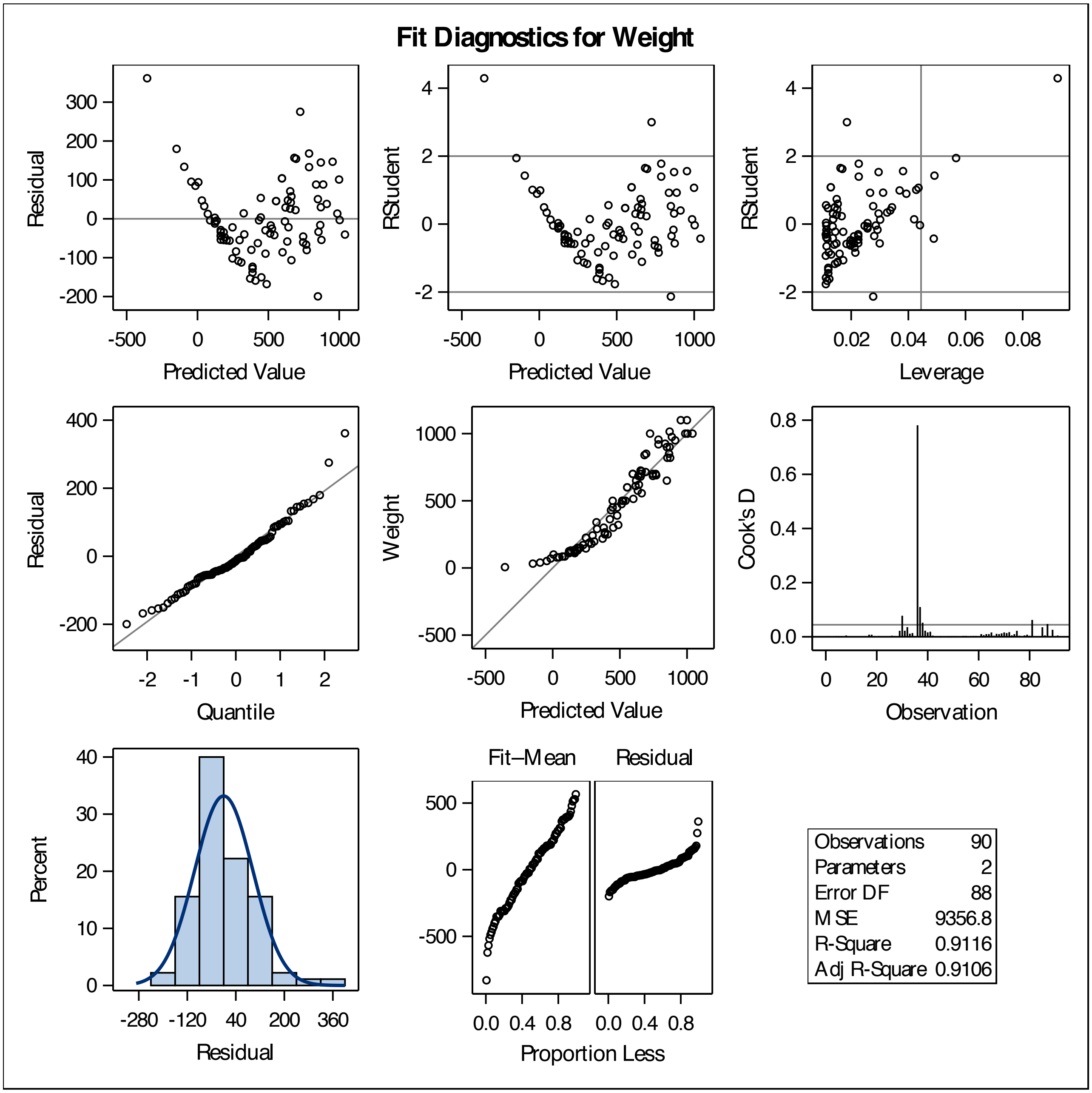
Observations	89
Parameters	2
Error DF	87
MSE	7810.4
R-Square	0.9252
Adj R-Square	0.9244


Residuals for Weight
Residual
200
100
0
-100
-200
10
20
30
40
Length1



Observations	89
Parameters	2
Error DF	87
MSE	7810.4
R-Square	0.9252
Adj R-Square	0.9244

Model: MODEL1
Dependent Variable: Weight



Fit Diagnostics for Weight

Observations	90
Parameters	2
Error DF	88
MSE	9356.8
R-Square	0.9116
Adj R-Square	0.9106

The detailed results for the final model are shown as above.

F value and P value for the model is 1076.81 and <0.0001, respectively. This shows the model is statistically significant.

R-Square is 0.9252, showing that the model is also practically significant.

The parameters for intercept and length1 is -715.99725 and 43.16899. The P values for these two are <0.0001 and <0.0001. This shows statistical significance. Moreover, it shows that with 1 unit increase for length1, weight is expected to increase 43.16899.

The residual vs. predicted value distribution looks flat around the 0. And only three points out of the 2 standard deviations. These findings suggest our assumption of normal distribution is satisfied.

Residual vs. Quantile line fits well to the straight line, showing the normal distribution.

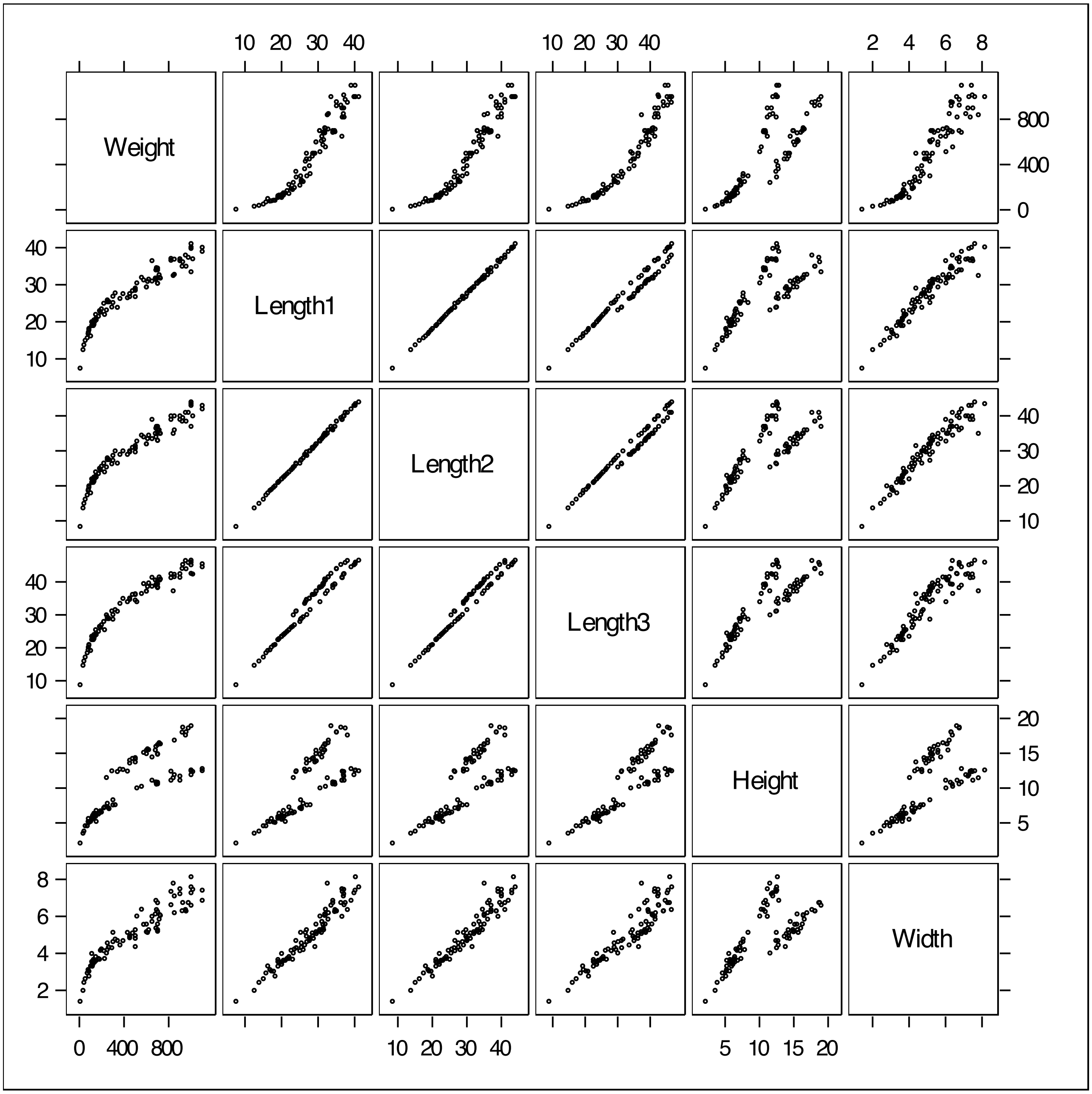
The cook's D shows there are few points beyond the line. However, the difference between these points is small.

Taken together, the final model is: weight = 43.16899 * length1-715.99725. There is positive relationship between weight and lenght1: the larger the lenght1, the larger the weight. In particular, with 1 unit increase for length1, weight is expected to increase 43.16899. As length1 approaches 0, weight is close to -715.99725, which may not make sense. It indicates the predicted weight for length1 closer to 0 may not be accurate.

Exercise 3

a) The process of model construction is shown as below:

1. We first plotted the scatter plots to see the correlations as shown below. It seems there is correlation between weight and all the five continuous predictors. It also shows there is redundant predictors as there is clear correlation between many predictors, such as length1 vs. length2.



2. We then use all the predictors to build the model and check the vif values. The detailed results are shown as below. Although P value = <0.0001 indicates the model is statistically significant, all the predictors have vif > 10. This suggests that there are many redundant predictors to remove.

Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	91
Number of Observations Used	90
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	8815944	1763189	298.71	<.0001
Error	84	495819	5902.60140		
Corrected Total	89	9311762			

Root MSE	76.82839	**R-Square**	0.9468
Dependent Mean	474.32667	**Adj R-Sq**	0.9436
Coeff Var	16.19736		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-567.52745	47.97637	-11.83	<.0001	0
Length1	1	-3.80093	37.13383	-0.10	0.9187	1148.52869
Length2	1	83.81404	39.83550	2.10	0.0384	1499.66701
Length3	1	-69.76502	21.92633	-3.18	0.0021	588.45091
Height	1	53.04414	12.79668	4.15	<.0001	48.01682
Width	1	74.51949	22.59148	3.30	0.0014	17.37290

3. We then use 3 different selections procedures to select the significant terms: stepwise, forward, backward selections. The results are shown as below.

All three selection procedures suggest the following significant terms: length2, height, width, lenght3.

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	0.9185	0.9185	42.4949	992.40	<.0001
2	Height		2	0.0097	0.9282	29.1958	11.76	0.0009
3	Width		3	0.0120	0.9402	12.3337	17.20	<.0001
4	Length3		4	0.0065	0.9467	4.0105	10.44	0.0018

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Length2	1	0.9185	0.9185	42.4949	992.40	<.0001
2	Height	2	0.0097	0.9282	29.1958	11.76	0.0009
3	Width	3	0.0120	0.9402	12.3337	17.20	<.0001
4	Length3	4	0.0065	0.9467	4.0105	10.44	0.0018

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Length1	4	0.0000	0.9467	4.0105	0.01	0.9187

4. Based on step3, we build new model using the parameters: length2, height, width, lenght3. The results are shown as below.
Although the model is significant, the cook’s d suggests there is an unduly influential point. The detailed information about this point is also shown as below.

Model: MODEL1
Dependent Variable: Weight

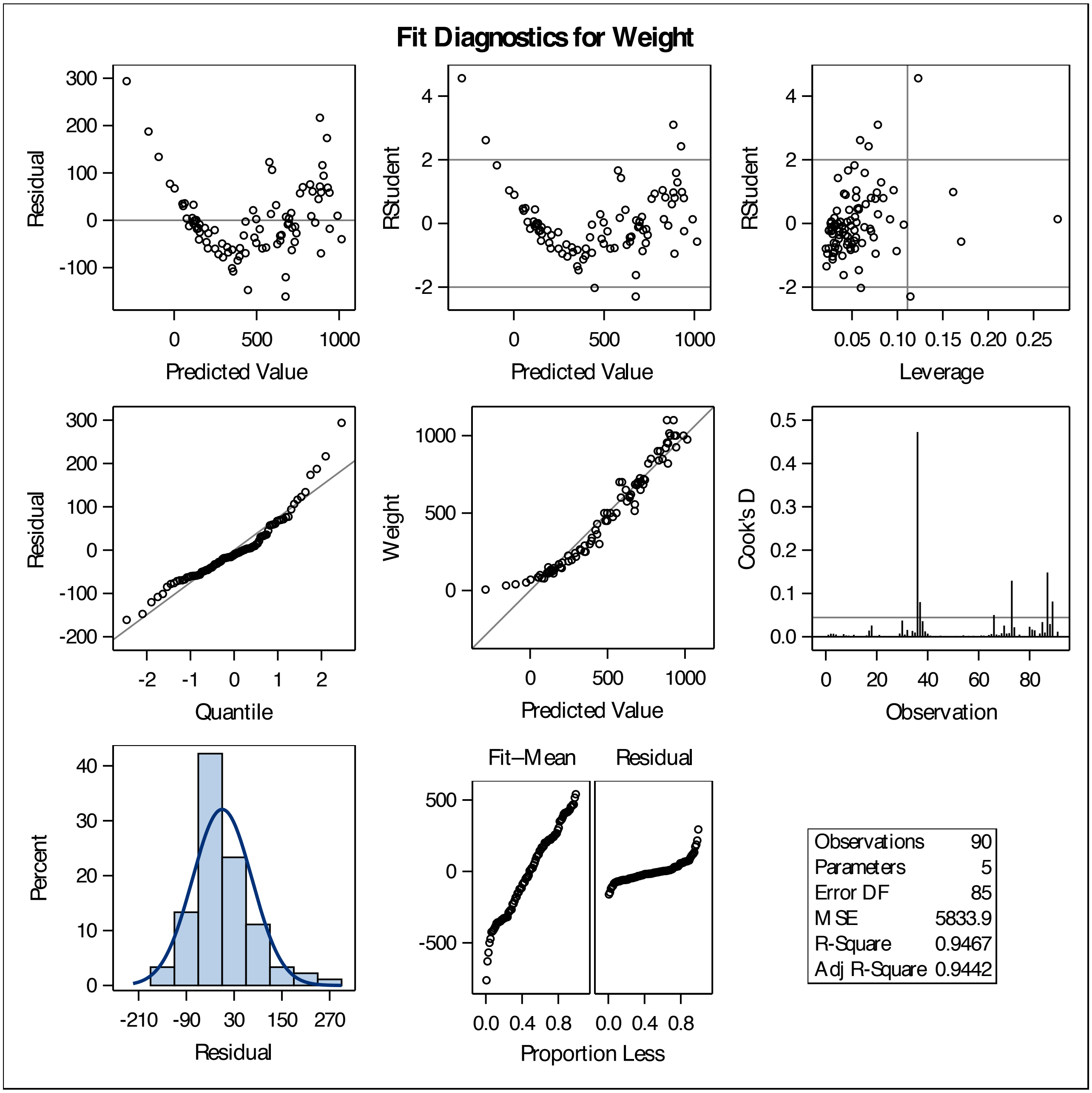
Number of Observations Read	91
Number of Observations Used	90
Number of Observations with Missing Values	1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	8815882	2203970	377.79	<.0001
Error	85	495880	5833.88658		
Corrected Total	89	9311762			

Root MSE	76.37988	**R-Square**	0.9467
Dependent Mean	474.32667	**Adj R-Sq**	0.9442
Coeff Var	16.10280		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-564.51116	37.63877	-15.00	<.0001
Length2	1	80.34202	20.76509	3.87	0.0002
Height	1	53.44273	12.11862	4.41	<.0001
Width	1	74.56867	22.45451	3.32	0.0013
Length3	1	-70.01399	21.66379	-3.23	0.0018

Model: MODEL1
Dependent Variable: Weight



Fit Diagnostics for Weight

Observations	90
Parameters	5
Error DF	85
MSE	5833.9
R-Square	0.9467
Adj R-Square	0.9442

Obs	widthgroup	Species	Weight	Length1	Length2	Length3	Height	Width	cd
36	thinner	Perch	5.9	7.5	8.4	8.8	2.112	1.408	0.47271

4. Based on step4, we remove the unduly influential point and rebuild the model. The results are shown as below.

The model is significant with F value and P value: 459.26 and <0.0001. This time, the cook's D shows there is no unduly influential point. Therefore, we use this model as our final model.

Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	90
Number of Observations Used	89
Number of Observations with Missing Values	1

Analysis of Variance

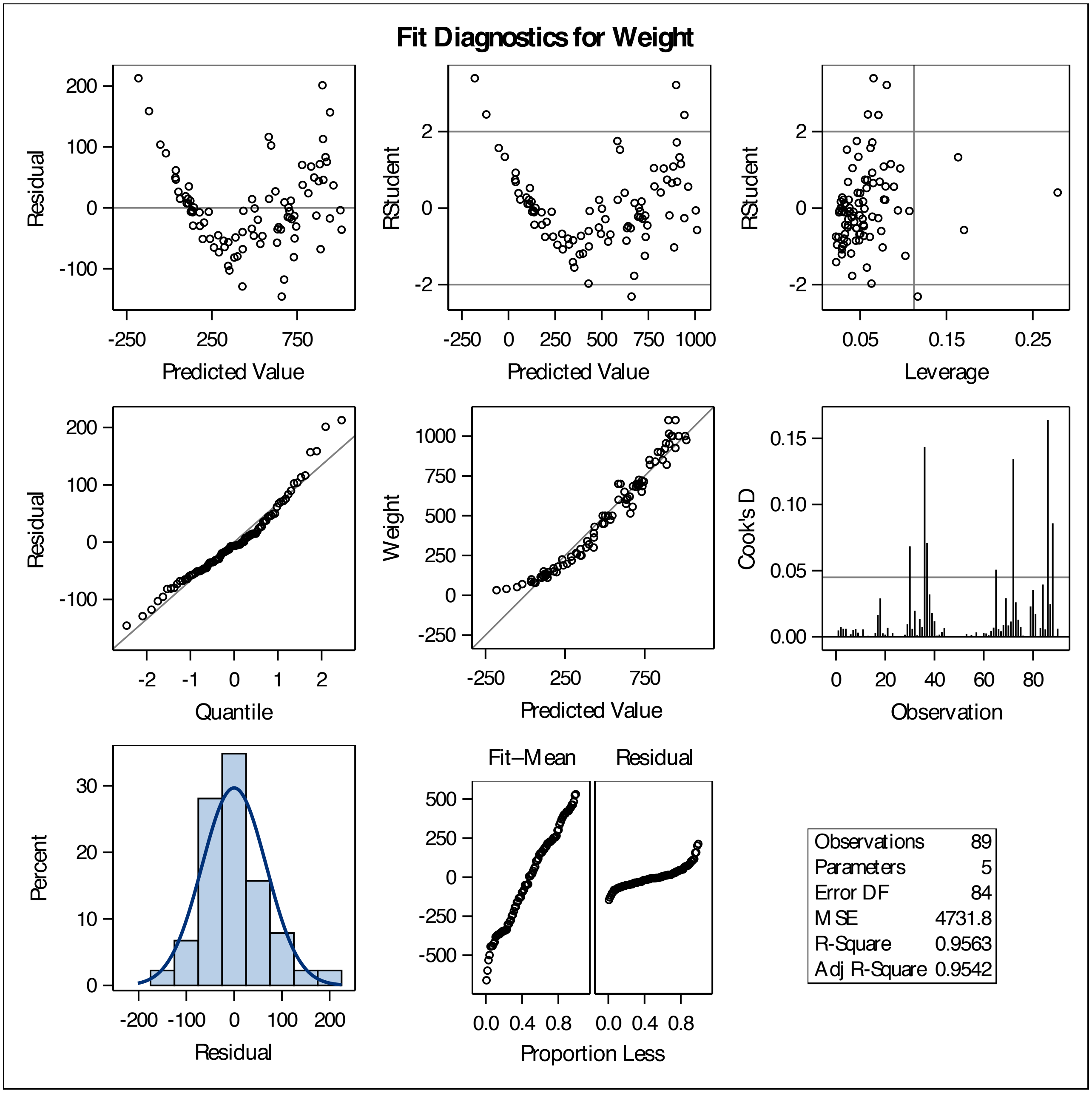
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	8692404	2173101	459.26	<.0001
Error	84	397469	4731.77618		
Corrected Total	88	9089873			

Root MSE	68.78791	**R-Square**	0.9563
Dependent Mean	479.58989	**Adj R-Sq**	0.9542
Coeff Var	14.34307		

Parameter Estimates

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

Model: MODEL1
Dependent Variable: Weight



Observations	89
Parameters	5
Error DF	84
MSE	4731.8
R-Square	0.9563
Adj R-Square	0.9542

b)

Model: MODEL1
Dependent Variable: Weight

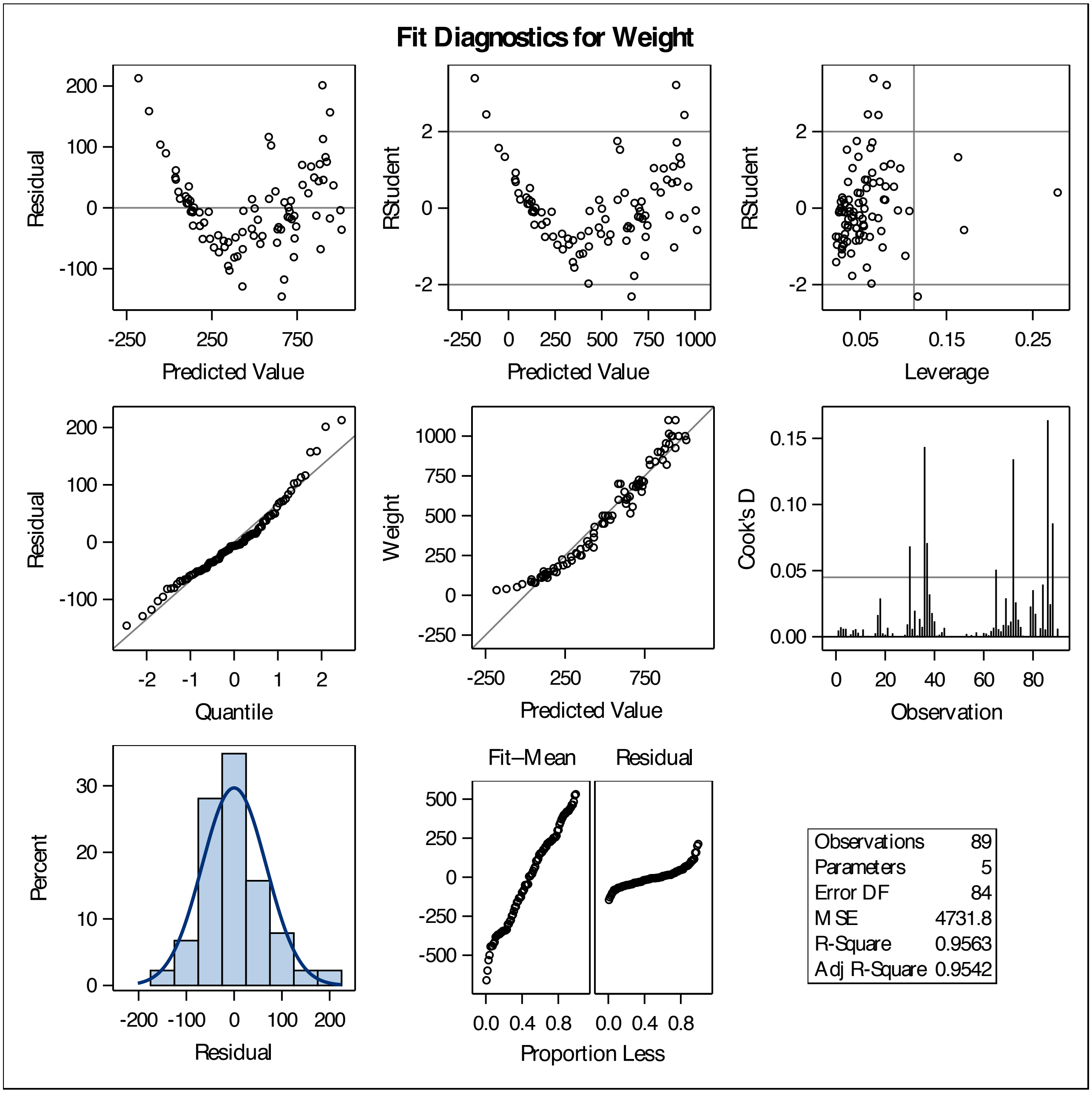
Number of Observations Read	90
Number of Observations Used	89
Number of Observations with Missing Values	1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	8692404	2173101	459.26	<.0001
Error	84	397469	4731.77618		
Corrected Total	88	9089873			

Root MSE	68.78791	**R-Square**	0.9563
Dependent Mean	479.58989	**Adj R-Sq**	0.9542
Coeff Var	14.34307		

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

Model: MODEL1
Dependent Variable: Weight



Fit Diagnostics for Weight

Observations	89
Parameters	5
Error DF	84
MSE	4731.8
R-Square	0.9563
Adj R-Square	0.9542

The detailed results for the final model are shown as above.

F value and P value for the model is 459.26 and <0.0001, respectively. This shows the model is statistically significant.

R-Square is 0.9563, showing that the model is also practically significant.

The parameters for intercept, length2, height, width, length3 are -621.85740, 72.56429, 45.98252, 67.92013, -57.89679. The P are <0.0001, 0.0002, <0.0001, 0.0012, 0.0042. This shows statistical significance. Moreover, it shows that with 1unit increase for length2, height, width, lenght3, the weight is expected to increase by 72.56429, 45.98252, 67.92013, -57.89679, individually.

The residual vs. predicted value distribution looks flat around the 0. And only five points out of the 2 standard deviations. These findings suggest our assumption of normal distribution is satisfied.

Residual vs. Quantile line fits well to the straight line, showing the normal distribution.

The cook's D shows there are few points beyond the line. However, the difference between these points is small.

Taken together, the final model is: weight = 72.56429*length2 + 45.98252 *height + 67.92013 * width -57.89679*lenght3-621.85740.

As all predictors approaches 0, weight is close to -621.85740, which may not make sense. It indicates the predicted weight when all predictors go closer to 0 may not be accurate. Overall, it shows the larger the length2, the larger the height, the larger the width, the smaller the lenght3, the larger the weight for the fish. In other words, all predictors show positive correlation with weight except length3, which indicates negative correlation.

Exercise 4

a) The process of building the model is:

1. We first generate new variables: log terms for weight and all the predictors.
2. We then use 3 different selections procedures to select the significant terms: stepwise, forward, backward selections. The results are shown as below.

Both stepwise and backward show the significant terms are: log(width), log(height), log(length1).

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	logl3		1	0.9881	0.9881	57.6913	7306.13	<.0001
2	logw		2	0.0033	0.9914	19.7313	33.51	<.0001
3	logh		3	0.0010	0.9925	9.1470	11.87	0.0009
4	logl1		4	0.0006	0.9930	4.0186	7.21	0.0087
5		logl3	3	0.0000	0.9930	2.1328	0.12	0.7348

Model: MODEL1
Dependent Variable: logwe

Summary of Forward Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	logl3	1	0.9881	0.9881	57.6913	7306.13	<.0001
2	logw	2	0.0033	0.9914	19.7313	33.51	<.0001
3	logh	3	0.0010	0.9925	9.1470	11.87	0.0009
4	logl1	4	0.0006	0.9930	4.0186	7.21	0.0087

Model: MODEL1
Dependent Variable: logwe

Summary of Backward Elimination							
Step	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	logl2	4	0.0000	0.9930	4.0186	0.02	0.8919
2	logl3	3	0.0000	0.9930	2.1328	0.12	0.7348

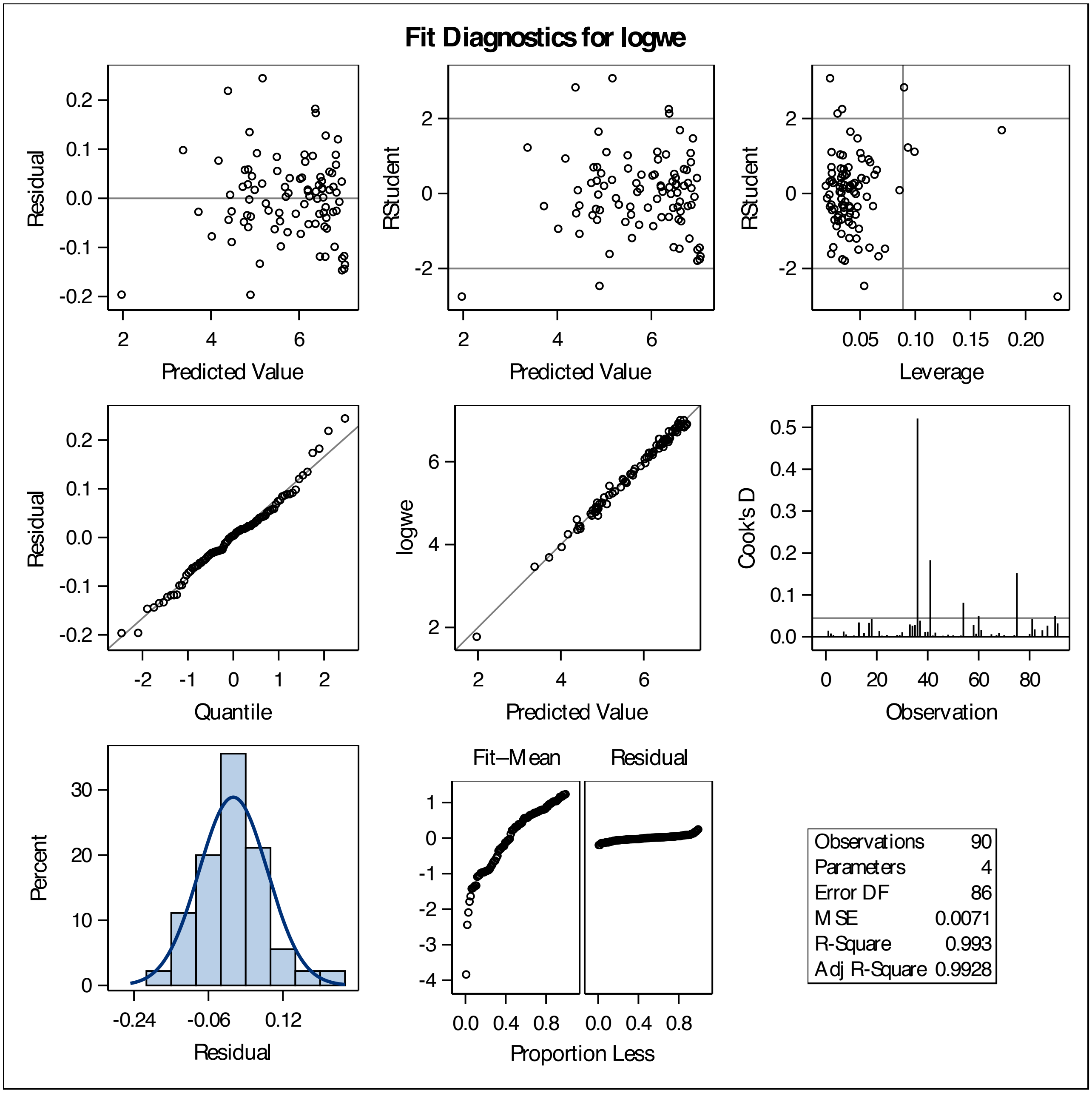
3. Based on step2, we build new model using the parameters: log(width), log(height), log(length1).
The results are shown as below.
Although the model is significant, the cook’s d suggests there is an unduly influential point.
The detailed information about this point is also shown as below.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	87.17324	29.05775	4085.17	<.0001
Error	86	0.61172	0.00711		
Corrected Total	89	87.78496			

Root MSE	0.08434	**R-Square**	0.9930
Dependent Mean	5.80801	**Adj R-Sq**	0.9928
Coeff Var	1.45211		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-2.04424	0.27192	-7.52	<.0001
logw	1	0.70019	0.12524	5.59	<.0001
logh	1	0.59888	0.03866	15.49	<.0001
logl1	1	1.65169	0.14789	11.17	<.0001

Model: MODEL1
Dependent Variable: logwe



Fit Diagnostics for logwe

Observations	90
Parameters	4
Error DF	86
MSE	0.0071
R-Square	0.993
Adj R-Square	0.9928

Obs	widthgroup	Species	Weight	Length1	Length2	Length3	Height	Width	logwe
36	thinner	Perch	5.9	7.5	8.4	8.8	2.112	1.408	1.77495

Obs	logl1	logl2	logl3	logh	logw	cd
36	2.01490	2.12823	2.17475	0.74764	0.34217	0.52123

4. Based on step3, we remove the unduly influential point and rebuild the model. The results are shown as below.

This time, the cook's D shows there is no unduly influential point. Therefore, we use this model as our final model.

The detailed results for the final model are shown as below.

F value and P value for the model is 3569.32 and <0.0001, respectively. This shows the model is statistically significant.

R-Square is 0.9921, showing that the model is also practically significant.

The parameters for intercept, log(width), log(height), log(length1) are -1.78499, 0.75908, 0.60896, 1.53822. The P are all <0.0001. This shows statistical significance. Moreover, it shows that with 1unit increase for log(width), log(height), log(length1), the log(weight) is expected to increase by 0.75908, 0.60896, 1.53822, individually.

The residual vs. predicted value distribution looks flat around the 0. And only five points out of the 2 standard deviations. These findings suggest our assumption of normal distribution is satisfied.

Residual vs. Quantile line fits well to the straight line, showing the normal distribution.

The cook's D shows there are few points beyond the line. However, the difference between these points is small.

Taken together, the final model is: weight = 0.75908* log(width)+ 0.60896* log(height)+ 1.53822* log(length1) -1.78499. As log terms of all predictors approaches 0, weight is close to -1.78499, which may not make sense. It indicates the predicted weight when log terms of all predictors approaches 0 may not be accurate. Overall, it shows the positive correlation with weight and the predictors.

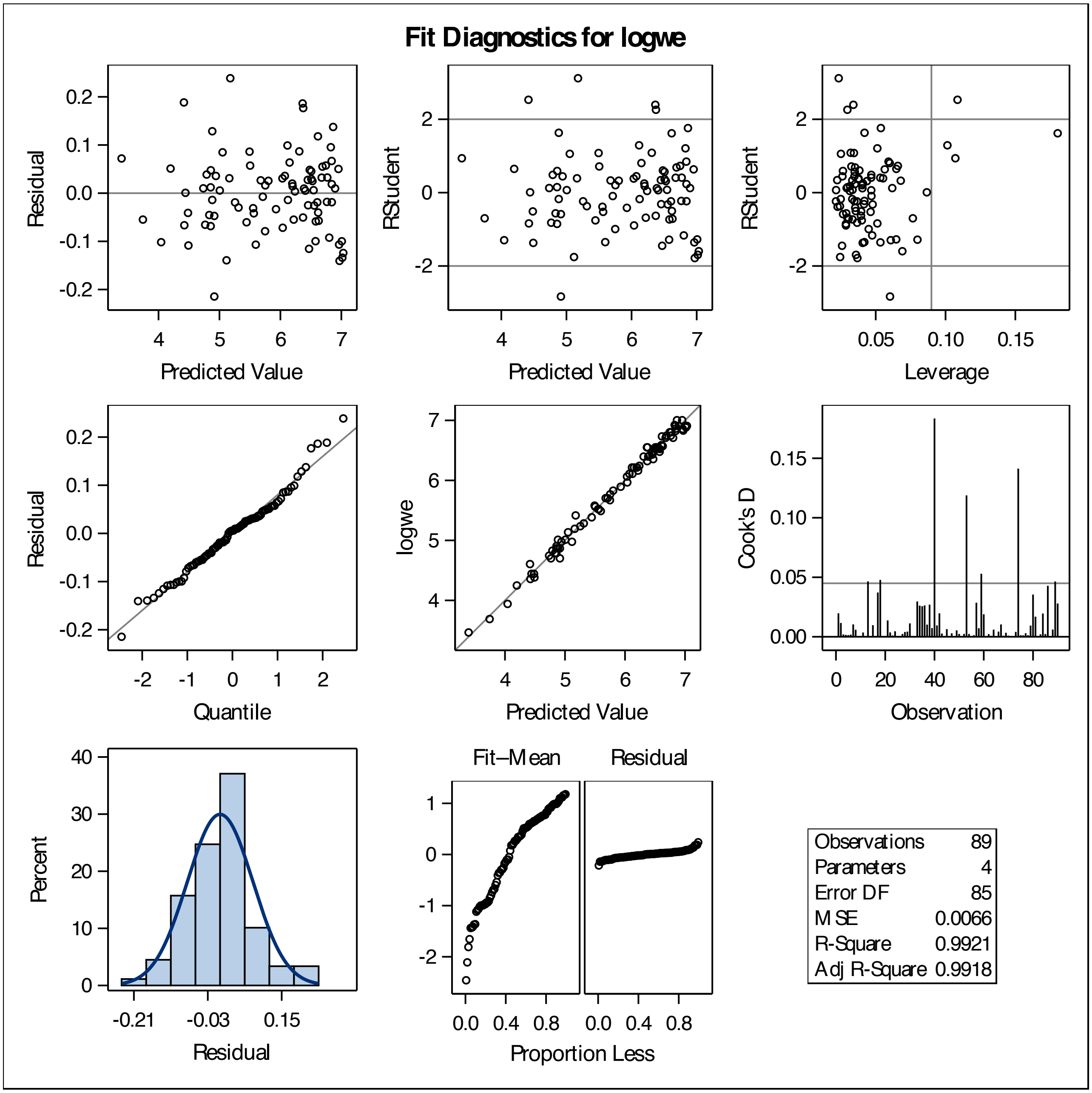
Comparing with the model in exercise 3, the R-Square increases from 0.9563 to 0.9921. This shows this model is better.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	70.77485	23.59162	3569.32	<.0001
Error	85	0.56181	0.00661		
Corrected Total	88	71.33666			

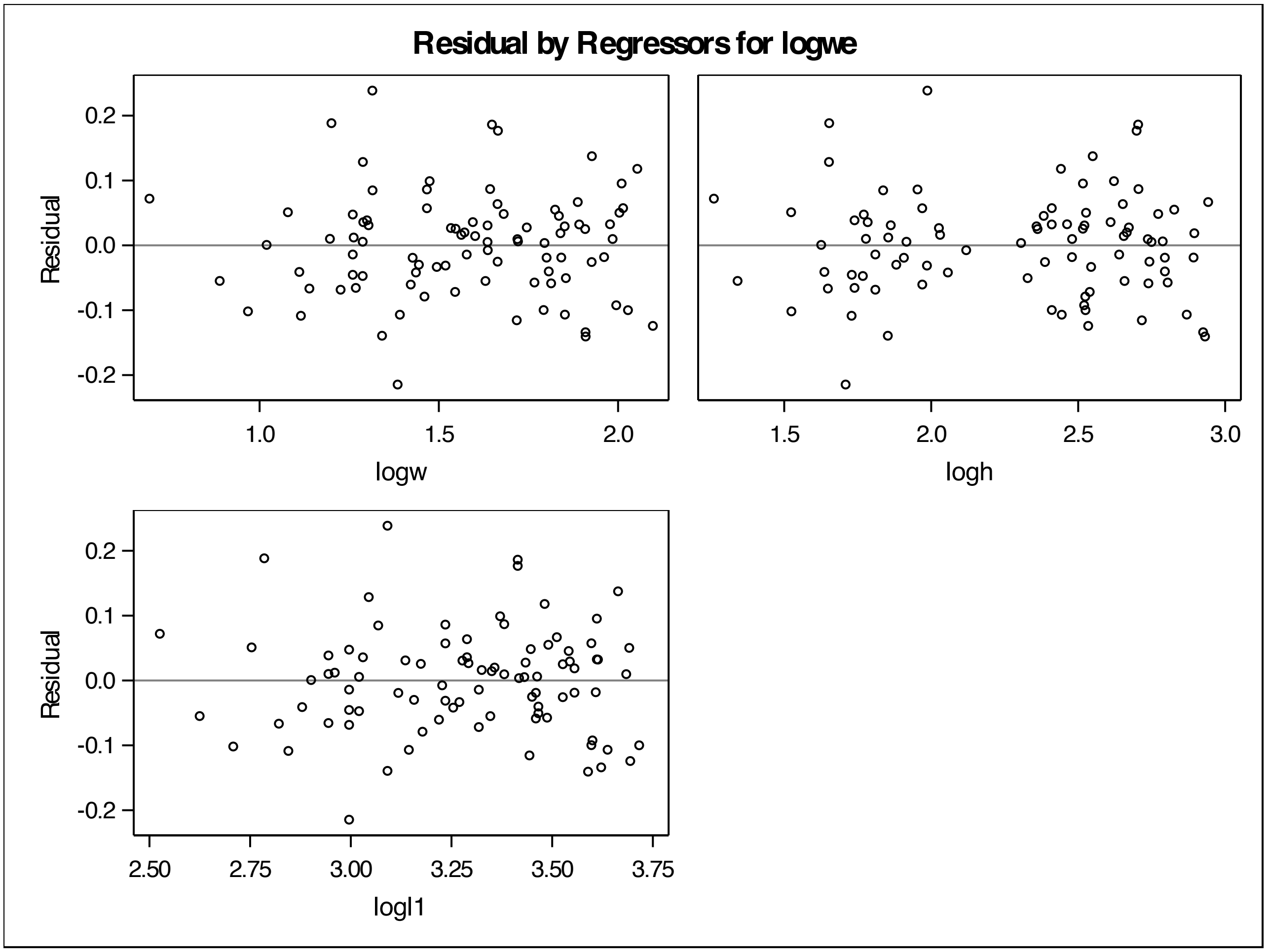
Root MSE	0.08130	**R-Square**	0.9921
Dependent Mean	5.85332	**Adj R-Sq**	0.9918
Coeff Var	1.38894		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-1.78499	0.27858	-6.41	<.0001
logw	1	0.75908	0.12262	6.19	<.0001
logh	1	0.60896	0.03745	16.26	<.0001
logl1	1	1.53822	0.14842	10.36	<.0001

Model: MODEL1
Dependent Variable: logwe



Observations	89
Parameters	4
Error DF	85
MSE	0.0066
R-Square	0.9921
Adj R-Square	0.9918


Residual by Regressors for logwe
Residual
0.2
0.1
0.0
-0.1
-0.2
1.0
1.5
2.0
logw
1.5
2.0
2.5
3.0
logh
Residual
0.2
0.1
0.0
-0.1
-0.2
2.50
2.75
3.00
3.25
3.50
3.75
logl1