|  | | **Height** | | |
| --- | --- | --- | --- | --- |
| **Mean** | **Std** | **N** |
| **Species** | **widthgroup** | 12.59 | 1.12 | 3 |
| **Bream** | **thinner** |
| **wider** | 15.43 | 1.85 | 32 |
| **Perch** | **thinner** | 5.77 | 1.16 | 33 |
| **wider** | 10.86 | 1.68 | 23 |
| **Pike** | **thinner** | 6.20 | 0.62 | 6 |
| **wider** | 8.54 | 1.45 | 11 |
| **Roach** | **thinner** | 6.29 | 0.84 | 17 |
| **wider** | 8.98 | 0.47 | 3 |

**STAT 448 Homework #3 Solutions**

1)

Cross Tabulation of Height by Species and Widthgroup

a) Based on the counts, Perch is the most common fish in this dataset, while Pike and Roach have the smallest number of observations with counts of 17 and 20 respectively. Bream is the largest fish, having the largest average height and the highest percentage of wider widths. Perch have the largest difference in height between its wider and thinner width groups. Roach tend to be thinner, while Pike tend to be wider than average. For every species, wider fish tend to be taller. Additionally, since the data is unbalanced, the glm procedure should be used for ANOVA models.

ANOVA Model: height = species \* widthgroup

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 7 | 1900.846751 | 271.549536 | 130.58 | <.0001 |
| **Error** | 120 | 249.555985 | 2.079633 |  |  |
| **Corrected Total** | 127 | 2150.402736 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **Height Mean** |
| --- | --- | --- | --- |
| 0.883949 | 14.92576 | 1.442093 | 9.661774 |

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Species** | 3 | 1489.009695 | 496.336565 | 238.67 | <.0001 |
| **widthgroup** | 1 | 377.381503 | 377.381503 | 181.47 | <.0001 |
| **Species\*widthgroup** | 3 | 34.455553 | 11.485184 | 5.52 | 0.0014 |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Species** | 3 | 355.6852405 | 118.5617468 | 57.01 | <.0001 |
| **widthgroup** | 1 | 154.1657747 | 154.1657747 | 74.13 | <.0001 |
| **Species\*widthgroup** | 3 | 34.4555533 | 11.4851844 | 5.52 | 0.0014 |

b) The two-way ANOVA model with interaction term is the best model for height as a function of species and width groups. Based on the type I and type III error, the interaction term and the two main effect terms are all very statistically significant in the model at a 5% level. This indicates all three terms should be kept in the final ANOVA model.

c) This ANOVA model is very statistically significant at a 5% level, along with the individual terms in the model. Based on the R-square value, 88.4% of the variance in height is explained by species, the width group and their interaction term.

| **Species** | **Height LSMEAN** | **LSMEAN Number** |
| --- | --- | --- |
| **Bream** | 14.0065313 | 1 |
| **Perch** | 8.3155752 | 2 |
| **Pike** | 7.3688455 | 3 |
| **Roach** | 7.6343618 | 4 |

| **Least Squares Means for Effect Species t for H0: LSMean(i)=LSMean(j) / Pr > |t|  Dependent Variable: Height** | | | | |
| --- | --- | --- | --- | --- |
| **i/j** | **1** | **2** | **3** | **4** |
| **1** |  | 11.92076 <.0001 | 11.67085 <.0001 | 10.15899 <.0001 |
| **2** | -11.9208 <.0001 |  | 2.280943 0.1083 | 1.384062 0.5118 |
| **3** | -11.6709 <.0001 | -2.28094 0.1083 |  | -0.45684 0.9682 |
| **4** | -10.159 <.0001 | -1.38406 0.5118 | 0.456836 0.9682 |  |

Least Squares Difference in Height Group Means for Species

Based on the results, Bream has a statistically significant greater height than every other species. There are no other statistically significant results, although Perch is almost significantly greater than Pike at a 10% level of significance.

Least Squares Difference in Height Group Means for Widthgroup

| **widthgroup** | **Height LSMEAN** | **H0:LSMean1=LSMean2** | |
| --- | --- | --- | --- |
| **t Value** | **Pr > |t|** |
| **thinner** | 7.7123619 | -8.61 | <.0001 |
| **wider** | 10.9502949 |  |  |

Based on the results, the height of wider fish is significantly greater than the height of thinner fish.

| **Species** | **widthgroup** | **Height LSMEAN** | **LSMEAN Number** |
| --- | --- | --- | --- |
| **Bream** | **thinner** | 12.5864000 | 1 |
| **Bream** | **wider** | 15.4266625 | 2 |
| **Perch** | **thinner** | 5.7748242 | 3 |
| **Perch** | **wider** | 10.8563261 | 4 |
| **Pike** | **thinner** | 6.1961000 | 5 |
| **Pike** | **wider** | 8.5415909 | 6 |
| **Roach** | **thinner** | 6.2921235 | 7 |
| **Roach** | **wider** | 8.9766000 | 8 |

| **Least Squares Means for Effect Species\*widthgroup t for H0: LSMean(i)=LSMean(j) / Pr > |t|  Dependent Variable: Height** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **i/j** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| **1** |  | -3.26187 0.0303 | 7.832864 <.0001 | 1.954381 0.5169 | 6.266757 <.0001 | 4.306234 0.0009 | 6.969836 <.0001 | 3.065741 0.0527 |
| **2** | 3.26187 0.0303 |  | 26.97689 <.0001 | 11.59344 <.0001 | 14.38777 <.0001 | 13.66005 <.0001 | 21.10545 <.0001 | 7.407508 <.0001 |
| **3** | -7.83286 <.0001 | -26.9769 <.0001 |  | -12.9726 <.0001 | -0.65822 0.9979 | -5.51069 <.0001 | -1.20156 0.9300 | -3.68183 0.0081 |
| **4** | -1.95438 0.5169 | -11.5934 <.0001 | 12.97256 <.0001 |  | 7.049433 <.0001 | 4.378534 0.0007 | 9.895322 <.0001 | 2.123435 0.4061 |
| **5** | -6.26676 <.0001 | -14.3878 <.0001 | 0.658224 0.9979 | -7.04943 <.0001 |  | -3.2047 0.0357 | -0.14022 1.0000 | -2.72674 0.1247 |
| **6** | -4.30623 0.0009 | -13.66 <.0001 | 5.510692 <.0001 | -4.37853 0.0007 | 3.204704 0.0357 |  | 4.031142 0.0024 | -0.46312 0.9998 |
| **7** | -6.96984 <.0001 | -21.1055 <.0001 | 1.201559 0.9300 | -9.89532 <.0001 | 0.140223 1.0000 | -4.03114 0.0024 |  | -2.9726 0.0677 |
| **8** | -3.06574 0.0527 | -7.40751 <.0001 | 3.681831 0.0081 | -2.12344 0.4061 | 2.726745 0.1247 | 0.463125 0.9998 | 2.972599 0.0677 |  |

Least Squares Difference in Height Group Means for Species\*Widthgroup

Bream in the thinner width group have significantly greater height than thinner Perch, thinner and wider Pike, thinner Roach, and nearly wider Roach although the difference is not statistically significant at a 5% level. Bream in the wider width group is significantly larger than every other group, including its thinner width group counterpart. Perch in the thinner width group is significantly shorter than wider Perch, wider Pike and wider Roach. Perch in the wider width group is significantly taller than thinner or wider Pike, and thinner Roach. Wider Pike is significantly taller than its thinner counterpart and thinner Roach. There is no significant difference between wider and thinner Roach. Overall, these results emphasize that wider fish tend to be taller and Bream is the tallest species of fish.

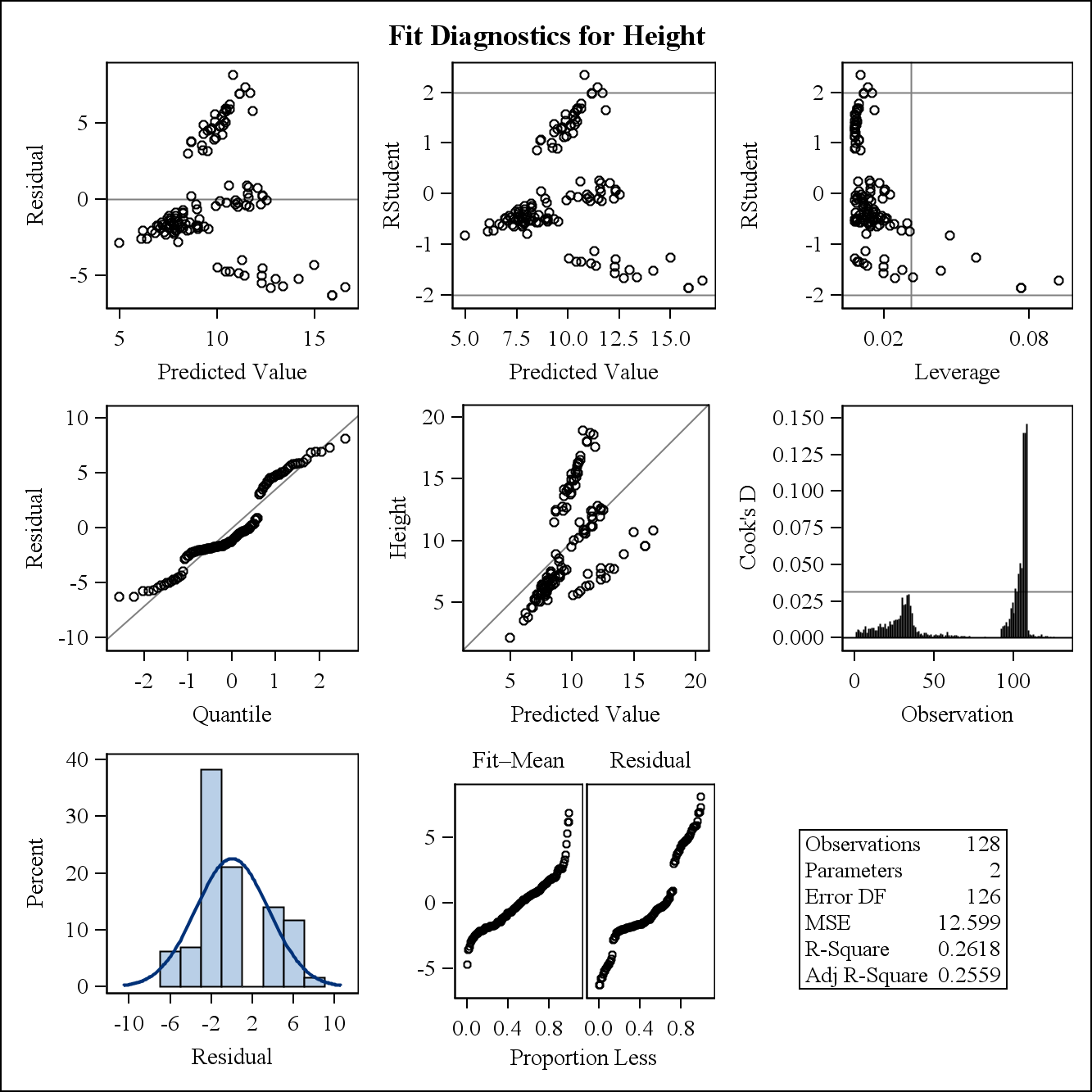
2)

Linear Regression Model: height = length1

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 562.90023 | 562.90023 | 44.68 | <.0001 |
| **Error** | 126 | 1587.50250 | 12.59923 |  |  |
| **Corrected Total** | 127 | 2150.40274 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 3.54954 | **R-Square** | 0.2618 |
| **Dependent Mean** | 9.66177 | **Adj R-Sq** | 0.2559 |
| **Coeff Var** | 36.73796 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 3.26284 | 1.00743 | 3.24 | 0.0015 |
| **Length1** | 1 | 0.22521 | 0.03369 | 6.68 | <.0001 |



a) A few points appear to be unduly influential; however, there are many observations in this model and none of their Cook’s Distances are greater than 1, so there is no need to remove any observations.

b) The linear regression model is very statistically significant and tells us that as length increases, so does height. This indicates that in general, longer fish tend to be taller fish. More specifically, for every one cm increase in length, height increases by .225 cm. About 26% of the variation in height is explained by length in this model.

There are some issues with diagnostics. For example, the residuals seem to increase as predicted values increase and they appear to be bunched together in four distinct groups. Additionally, there are unusual trends in the normal quantile plot that are likely caused by these clumps in the data. This indicates we should proceed with caution when using this model for inference. Additionally, after inspecting the plot of predicted values and actual values, it appears there are 3 or 4 distinct groups of observations with similar slopes. Perhaps, these groups are the different species of fish as the next two questions investigate. Overall, although the model is statistically significant, the percentage of variation explained is not very large and there are some obvious diagnostic issues. This indicates it may be more appropriate to split the linear regression model up by species of fish.

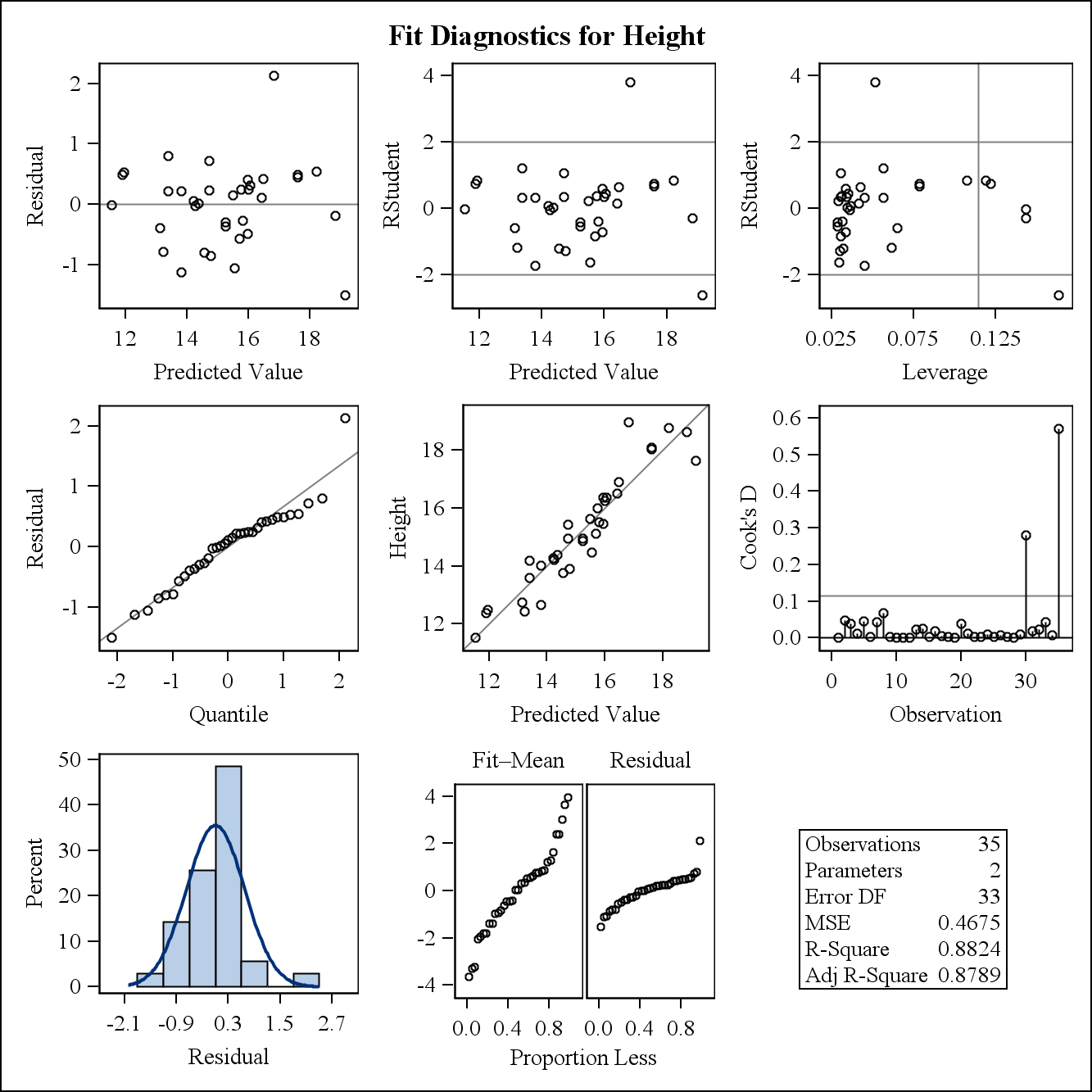
3)

Linear Regression Model: height = length1 for Bream

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 115.81379 | 115.81379 | 247.71 | <.0001 |
| **Error** | 33 | 15.42867 | 0.46754 |  |  |
| **Corrected Total** | 34 | 131.24247 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.68377 | **R-Square** | 0.8824 |
| **Dependent Mean** | 15.18321 | **Adj R-Sq** | 0.8789 |
| **Coeff Var** | 4.50343 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | -0.38087 | 0.99563 | -0.38 | 0.7045 |
| **Length1** | 1 | 0.51357 | 0.03263 | 15.74 | <.0001 |



a) There does appear to be one observation that may be unduly influential; however, its Cook’s Distance in less than 1 indicating it is likely fine to keep this observation in the model, though it would be reasonable to consider removing it.

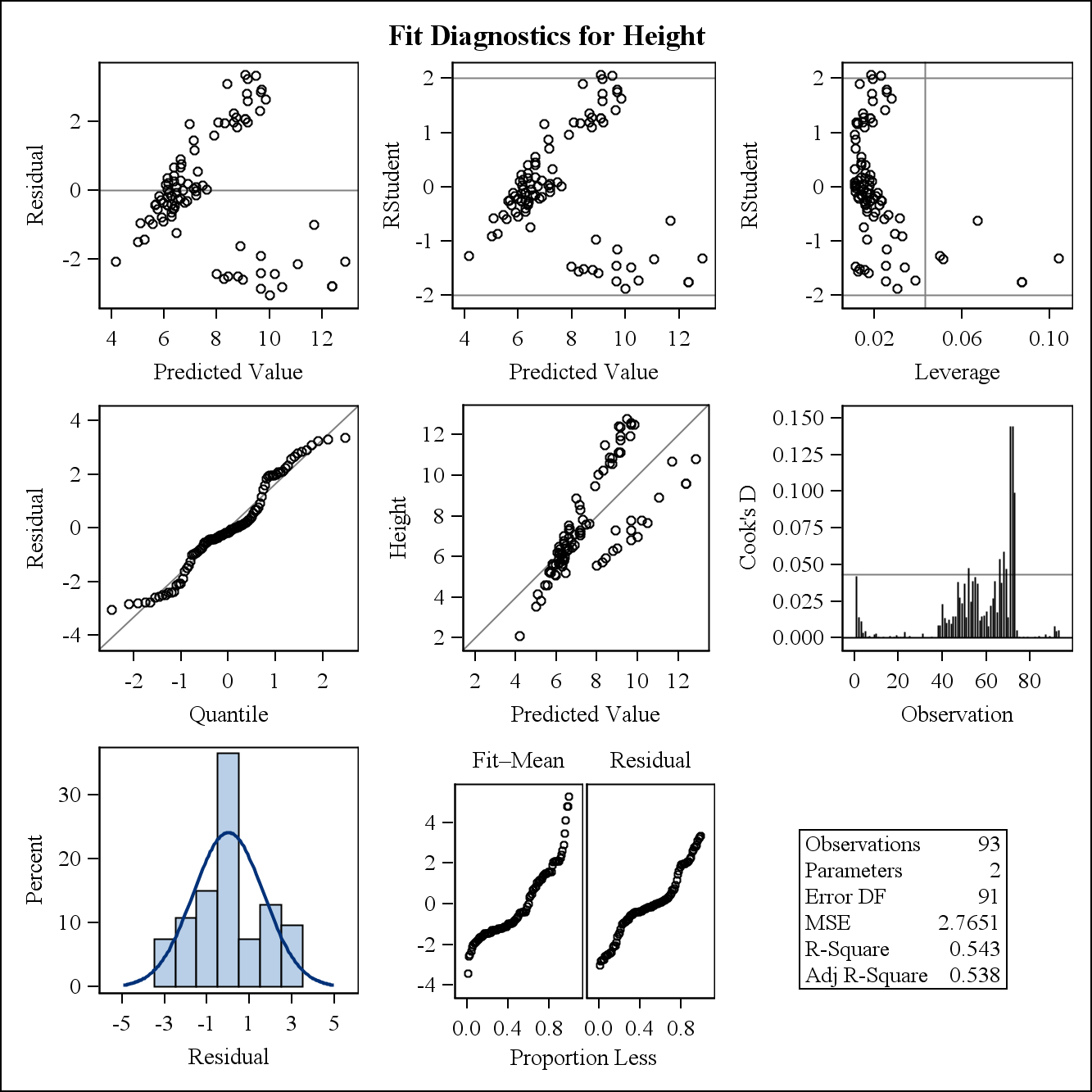
b) The model is very statistically significant and explains about 88% of the variation in height, much larger than the model including every species of fish. Although the intercept is negative, it is statistically insignificant so we do not have to worry about the real world consequences of a negative intercept. The model tells us that longer bream fish tend to be taller, the same relationship as the model including every species of fish. More specifically, for every cm increase in length, Bream height increases by .51 cm. The beta value for length is higher in this model compared to the model from question 2, indicating height increases more rapidly for increases in length compared to other species. There do not appear to be any problems with the diagnostics of the model. In fact, looking at the graph of the predicted height values and actual values, the model appears to be very successful.

Linear Regression Model: height = length1 for Perch, Pike and Roach species

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 298.95091 | 298.95091 | 108.12 | <.0001 |
| **Error** | 91 | 251.62351 | 2.76509 |  |  |
| **Corrected Total** | 92 | 550.57442 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1.66286 | **R-Square** | 0.5430 |
| **Dependent Mean** | 7.58381 | **Adj R-Sq** | 0.5380 |
| **Coeff Var** | 21.92640 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 2.90672 | 0.48173 | 6.03 | <.0001 |
| **Length1** | 1 | 0.16884 | 0.01624 | 10.40 | <.0001 |



There appear to be a few highly influential points, however; their Cook’s Distances are much less than 1 and there are quite a few observations, so no observations need to be removed from the model.

The model is very statistically significant and explains about 54% of the variation in height. Similar to previous models, as the length increases so does the height, meaning taller fish tend to be longer fish. Specifically, for every cm increase in length, height increases by .17 cm.

There are some issues with the diagnostics. Similar to the model in question 2, the residuals seem to increase as the predicted value increase and there seems to be two or three distinct groups. Looking at the graph of predicted values and actual values, it seems there is a split in the data, with one group requiring a higher beta value for length and one requiring a lower beta value for length. It will be interesting to see whether these differences are caused by differences in species as every species will be fit to its own model in question 4.

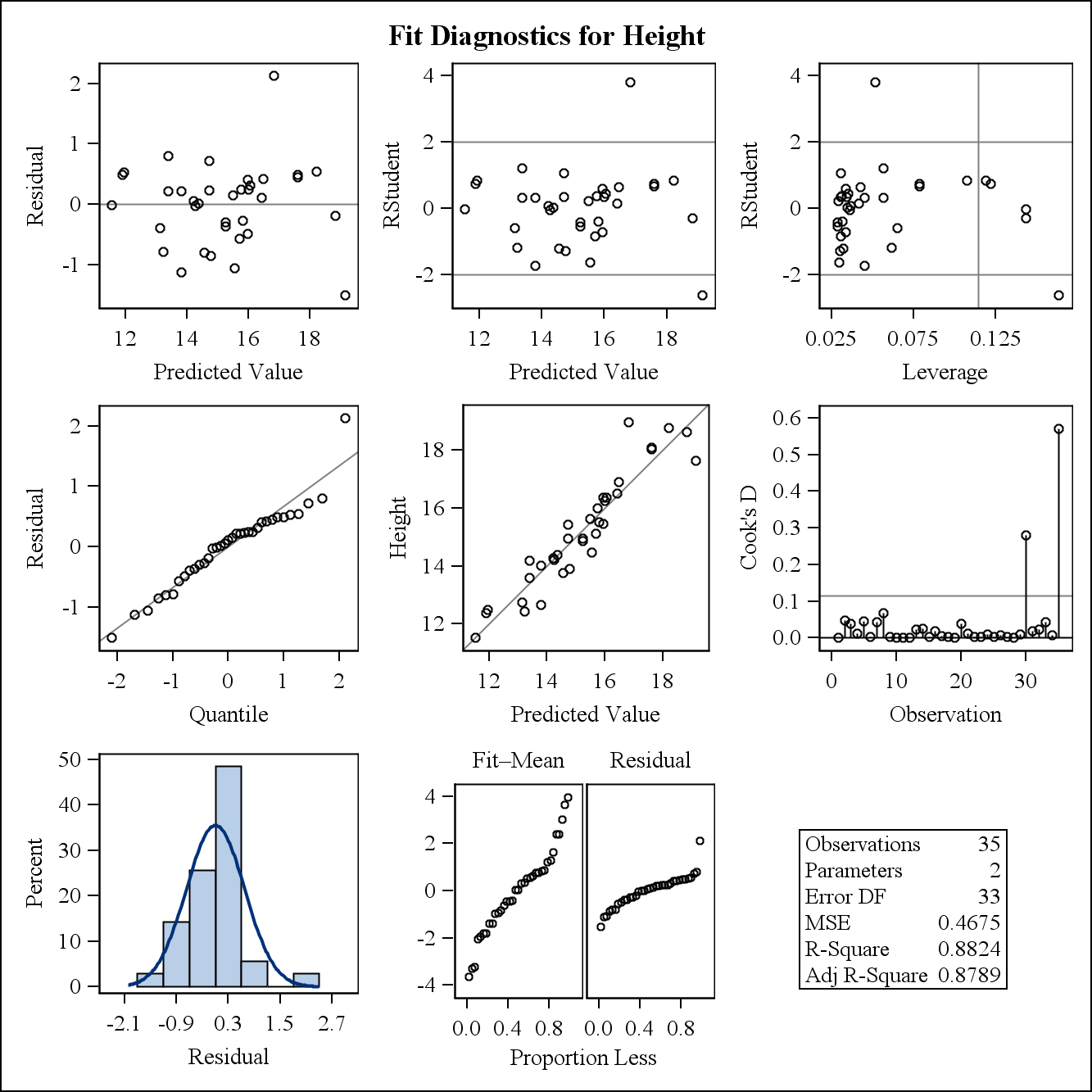
The amount of variation explained in this model increased as Bream were fitted to their own model. Overall, it seems these two models were more successful than the model from question 2 as they more accurately account for significant differences between species.

4)

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 115.81379 | 115.81379 | 247.71 | <.0001 |
| **Error** | 33 | 15.42867 | 0.46754 |  |  |
| **Corrected Total** | 34 | 131.24247 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.68377 | **R-Square** | 0.8824 |
| **Dependent Mean** | 15.18321 | **Adj R-Sq** | 0.8789 |
| **Coeff Var** | 4.50343 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | -0.38087 | 0.99563 | -0.38 | 0.7045 |
| **Length1** | 1 | 0.51357 | 0.03263 | 15.74 | <.0001 |

Linear Regression Model: height = length1 for Bream

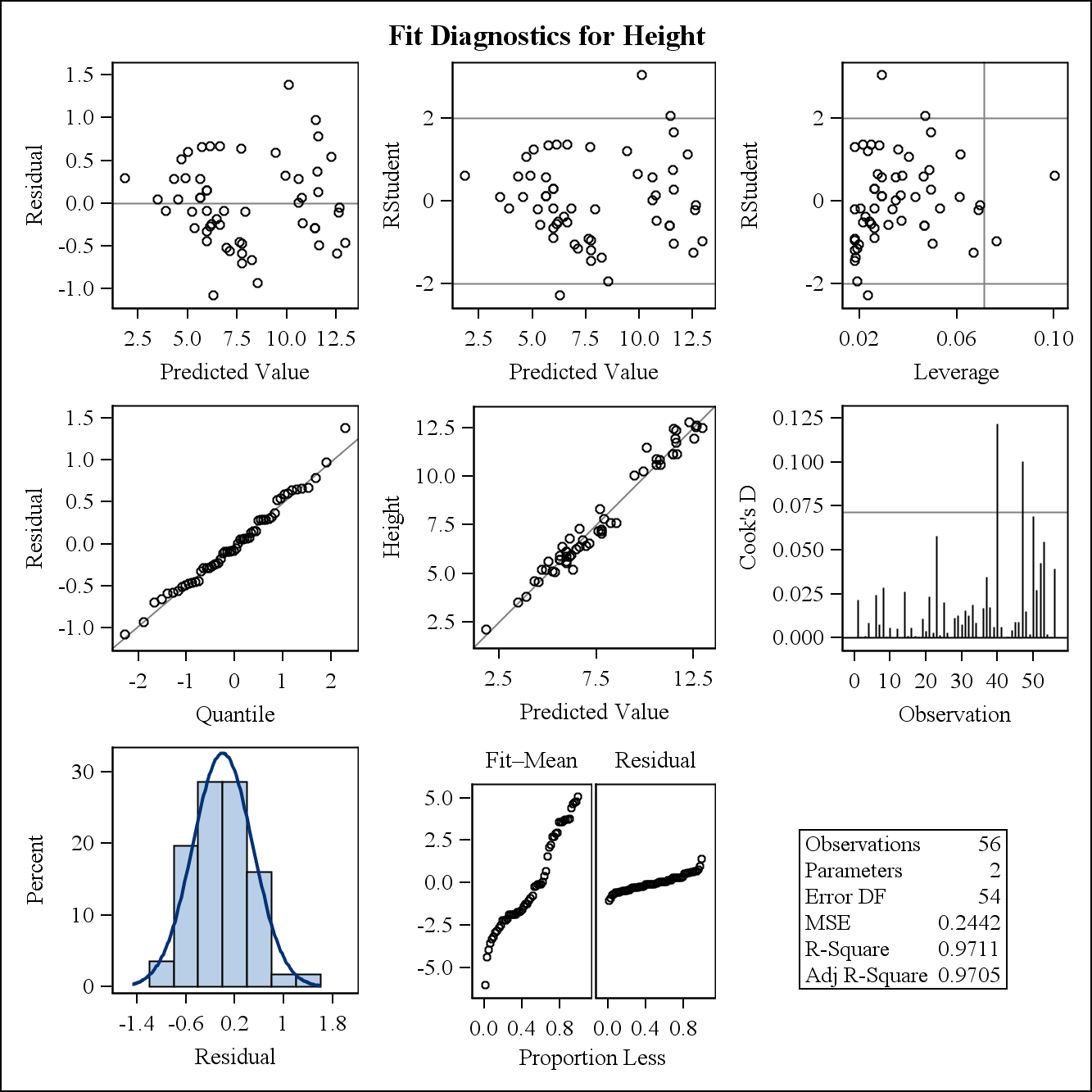
This model was discussed in problem 3.

Linear Regression Model: height = length1 for Perch

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 442.42590 | 442.42590 | 1811.47 | <.0001 |
| **Error** | 54 | 13.18873 | 0.24424 |  |  |
| **Corrected Total** | 55 | 455.61463 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.49420 | **R-Square** | 0.9711 |
| **Dependent Mean** | 7.86187 | **Adj R-Sq** | 0.9705 |
| **Coeff Var** | 6.28606 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | -0.66367 | 0.21092 | -3.15 | 0.0027 |
| **Length1** | 1 | 0.33127 | 0.00778 | 42.56 | <.0001 |



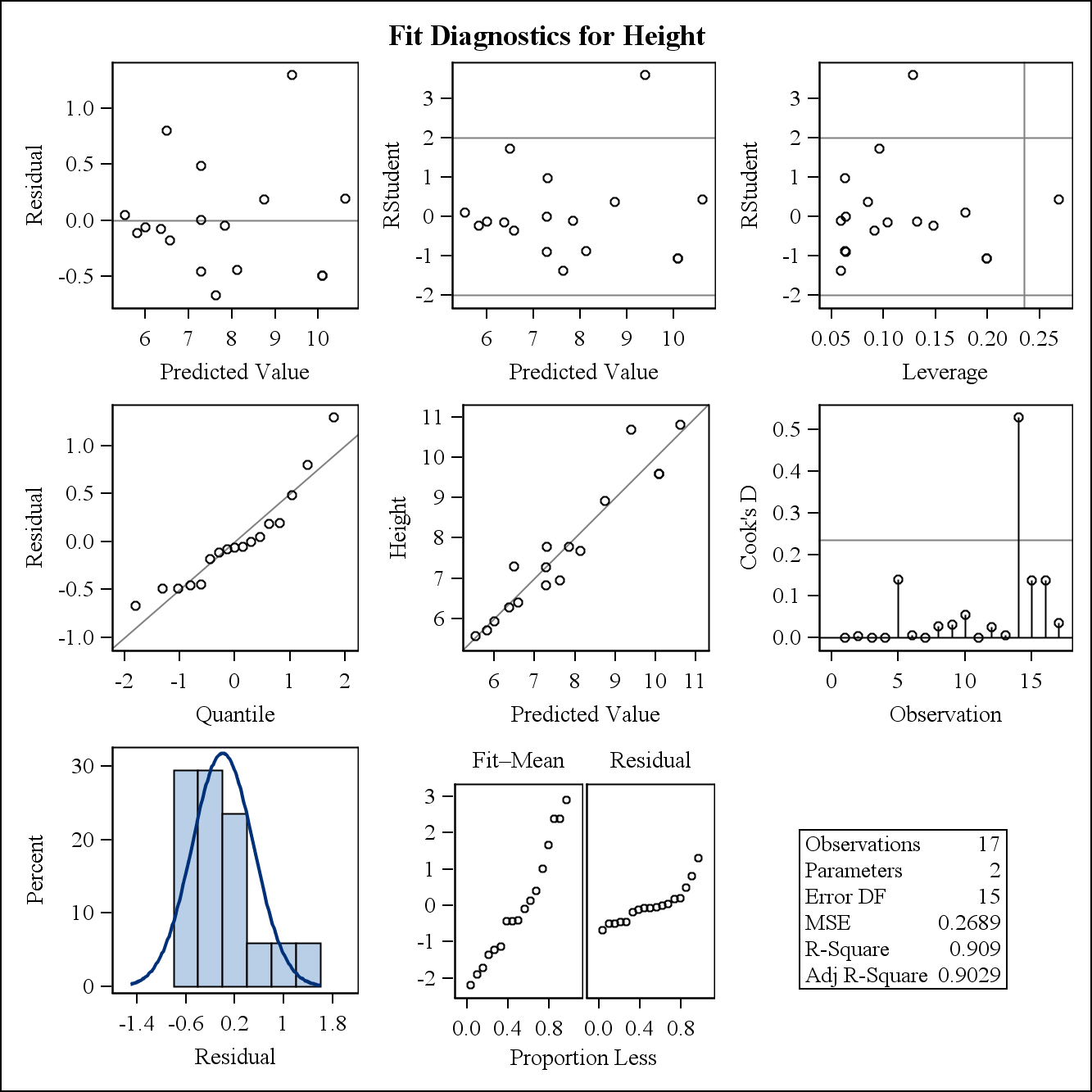
There do not appear to be any unduly influential points; therefore no observations should be removed from this model. The model is extremely statistically significant and explains 97% of the variation in height. The intercept is negative and statistically significant, which could cause problems as height cannot be negative in the real world. However, all length values for Perch in this dataset correspond to positive height values, so this intercept should not be a problem for real world Perch. As evidenced by the large amount of variation explained, the graph of predicted and fitted values appears nearly perfect. There does not appear to be any issues with the diagnostics. As length increases, so does height; however, at a slower rate compared to bream from the previous model. More specifically, for every cm increase in length, Perch increases .33 cm in height. Overall, this model is very successful.

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 40.28044 | 40.28044 | 149.78 | <.0001 |
| **Error** | 15 | 4.03402 | 0.26893 |  |  |
| **Corrected Total** | 16 | 44.31446 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.51859 | **R-Square** | 0.9090 |
| **Dependent Mean** | 7.71377 | **Adj R-Sq** | 0.9029 |
| **Coeff Var** | 6.72290 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 0.24943 | 0.62275 | 0.40 | 0.6944 |
| **Length1** | 1 | 0.17573 | 0.01436 | 12.24 | <.0001 |

Linear Regression Model: height = length1 for Pike



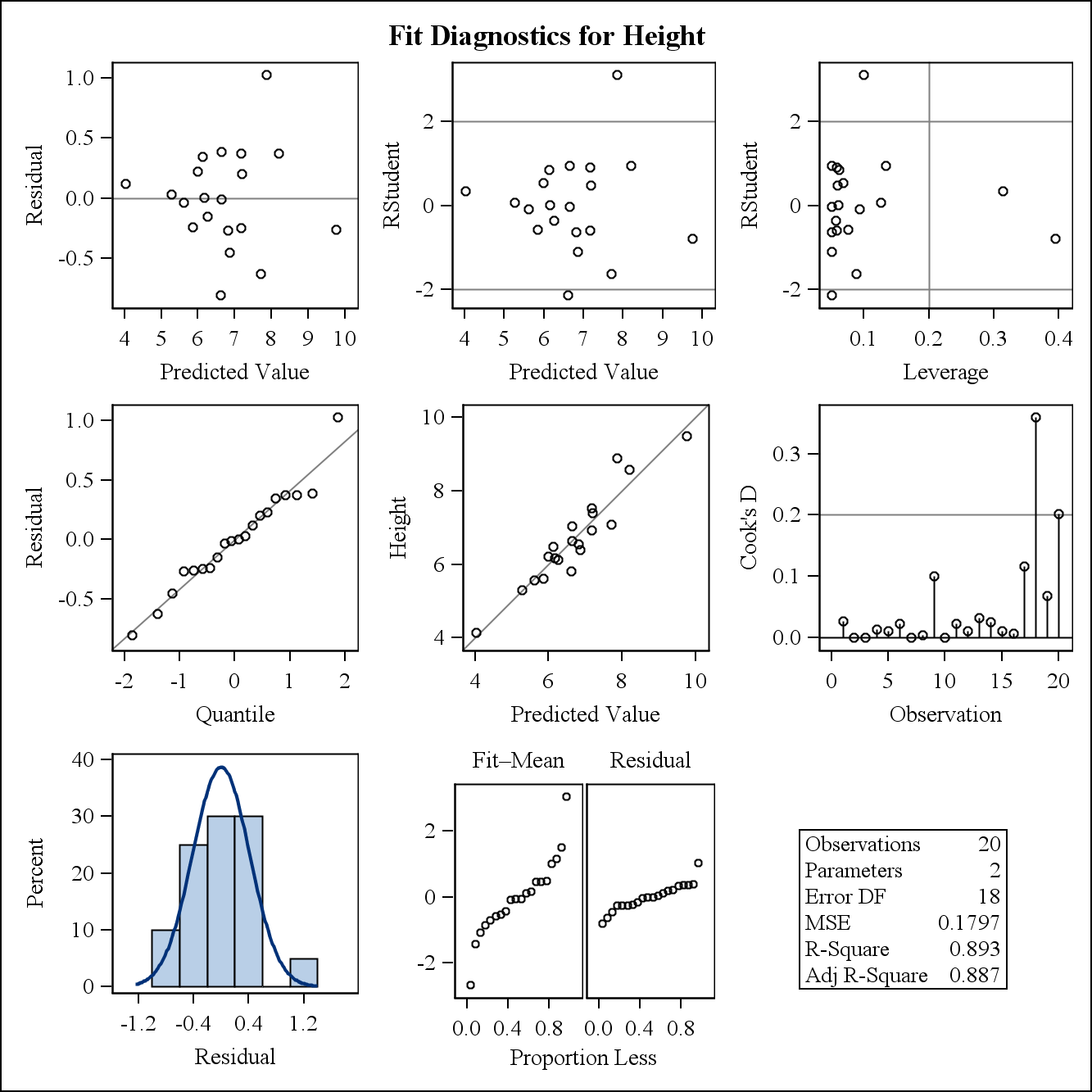
There is one observation that seems to have a large influence on the model, however; its Cook’s Distance is less than 1 so it is likely appropriate to leave it in the model. The model is very significant and explains around 90% of the variation in height. There are no problems with diagnostics in this model. Overall, the model is successful at modeling height based on length for Pike fish. The height of Pike fish increases as length increases; however, this occurs at a slower rate compared to Bream and Perch. Specifically, for every cm increase in length, Pike height increases .18 cm.

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 26.98690 | 26.98690 | 150.18 | <.0001 |
| **Error** | 18 | 3.23458 | 0.17970 |  |  |
| **Corrected Total** | 19 | 30.22149 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.42391 | **R-Square** | 0.8930 |
| **Dependent Mean** | 6.69480 | **Adj R-Sq** | 0.8870 |
| **Coeff Var** | 6.33192 |  |  |

Linear Regression Model: height = length1 for Roach

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | -0.41650 | 0.58798 | -0.71 | 0.4878 |
| **Length1** | 1 | 0.34446 | 0.02811 | 12.25 | <.0001 |



There do not appear to be any unduly influential observations. The model is very statistically significant and explains about 90% of the variance in height. There are no problems with the diagnostics. Overall, this model is very successful, similar to the other models that are conditioned on species of fish. For every cm increase in length, Roach height increases by .34 cm. Based on the model, the rate at which increases in length correspond to increases in height is very similar to Perch, however; the intercept for Roach is larger, indicating Roach have slightly higher height relative to length compared to Perch.

Conclusion:

Overall, the models for each species of fish separately were most successful. Based on the very high amount of variation explained in these models and the relatively different parameter estimates, it appears different species of fish have different height: length ratios. Furthermore, there we some issues with diagnostics in the models incorporating more than one species of fish, so these models might not be appropriate to use for inference. It appears different species of fish have different shapes, which caused problems for the models in question 2 and 3 that grouped multiple species of fish together in a model.