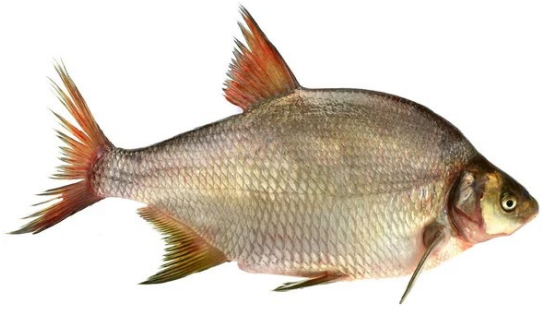
Handout with Supplementary Material for Quiz 1 Problem 2

Problem 2 uses a fish data set that has multiple measurements for multiple species of fish. We will restrict our analysis to two species, Bream and Perch, and four continuous measurements as describe in the Table 1 below. We are not completely sure of the origin of this data set, but it appears like one based on a 1917 study of fish caught in Lake Laengelmavesi near Tampere in Finland.

Table 1. Fish variables for Quiz 1 Problem 2.

Variable	**Description**
Weight	Weight of fish in grams (g)
Length3	Cross length of fish in cm
Height	Height of fish in cm
Width	Diagonal width of fish in cm
Species	Categorical variable, either Bream or Perch

For this analysis, we want to determine which of fish cross length, height, or diagonal width is the best single predictor of fish weight. We also want to explore difference between the two species. Figure 1 shows you images of typical Bream and Perch for reference.

<table>
<tr>
<td>A stock Bream
 
Picture from
https://depositphotos.com/stock-photos/sea-bream.html</td>
<td> 
A yellow Perch

Picture from
https://wiki.fishingplanet.com/Yellow_Perch/en</td>
</tr>
</table>

Figure 1. Images of a typical Bream and a typical Perch.

Handout with Supplementary Material for Quiz 1 Problem 2

Exploratory Data Analysis

Pages 3 – 8 provide various EDA output for the fish data set. All EDA includes information on the four quantitative variables overall and split out by species.

Figure 2 shown on the next page is a pairs plot for four continuous variables along with density plots for each variable. All plots also include a third variable, Species, which is color coded (Bream pink and Perch blue). For our analysis, Weight is the response and Length3, Height and Width are possible predictors. For each pair, three correlations are shown, one overall, one for Bream only and one for Perch only. The first plot in row 1 shows the density plots for Weight by species. The remaining information in row 1 are correlations between Weight and each explanatory variable overall (black), Bream (pink), and Perch (blue). Each following row is for one of the explanatory variables and with a color-coded scatterplot for Weight on each of the explanatory variables. The correlations in row 2 are between Height and Length3 and Width and Length3. Some annotations have been added to Figure 2 to help you understand how to read it.

Pages 9 – 11 show linear model output, including diagnostic plots for single variable models for Weight on each of the predictors for all data.

Pages 12 and 13 introduce a regression that includes a single quantitative variable (only presented for Width) with the addition of the categorical variable Species. In some sense, you can think of each regression as fitting a separate model for each species in a single `lm` call. Interpretation of the output is included on these pages.

Handout with Supplementary Material for Quiz 1 Problem 2

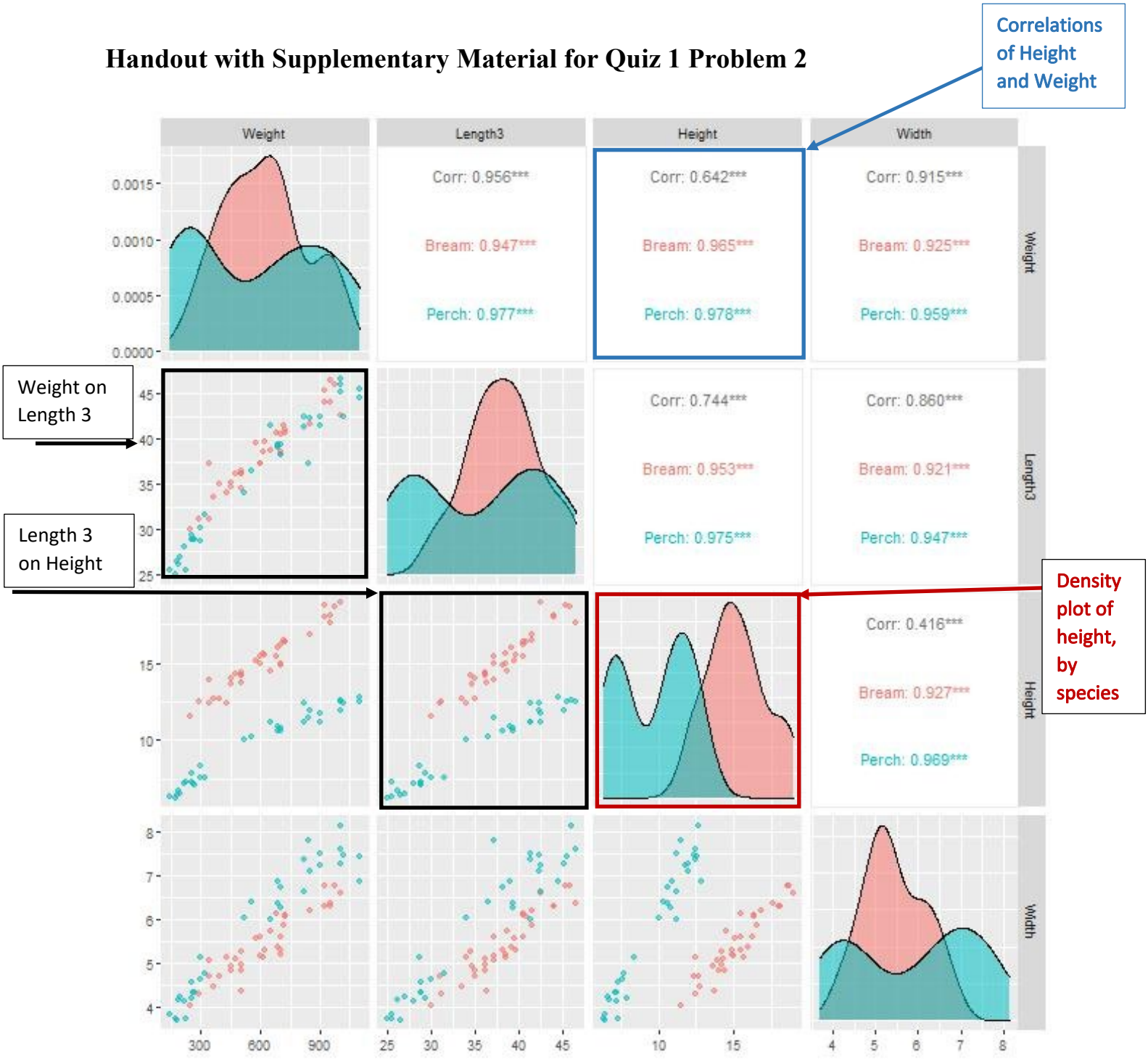


Figure 2. Pairs plot showing scatterplots, density plots and correlations for all pairwise combinations of Weight, Length3, Height, and Width. Pink indicates the species Bream and light blue indicate Perch.

Handout with Supplementary Material for Quiz 1 Problem 2

Figures 3 – 6 show individual histograms and boxplots for each variable altogether and split out by species. Table 2 shows summary statistics for each of the variables as well.

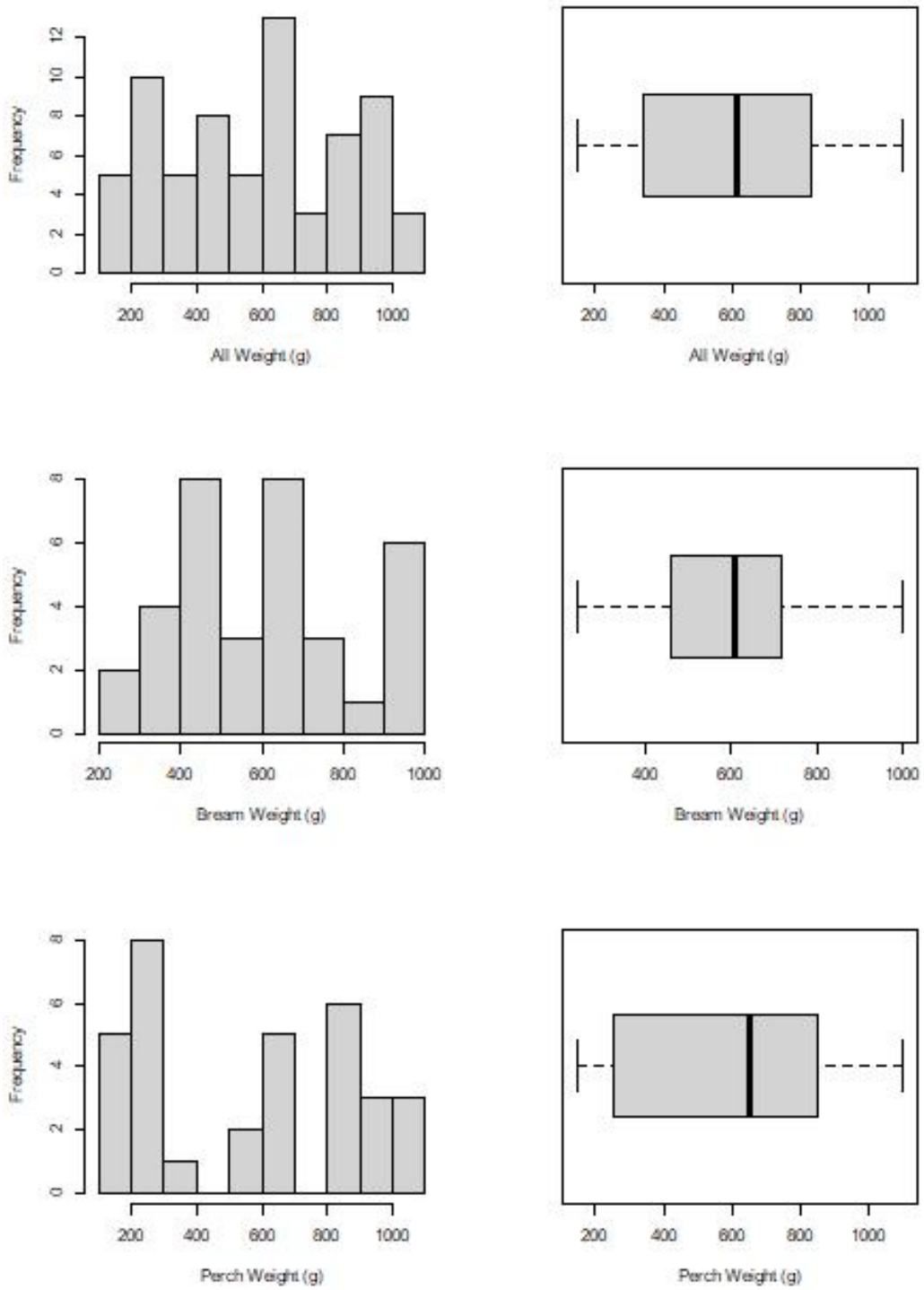


Figure 3. Histograms and boxplots for each of Weight overall and Weight by species.

Handout with Supplementary Material for Quiz 1 Problem 2

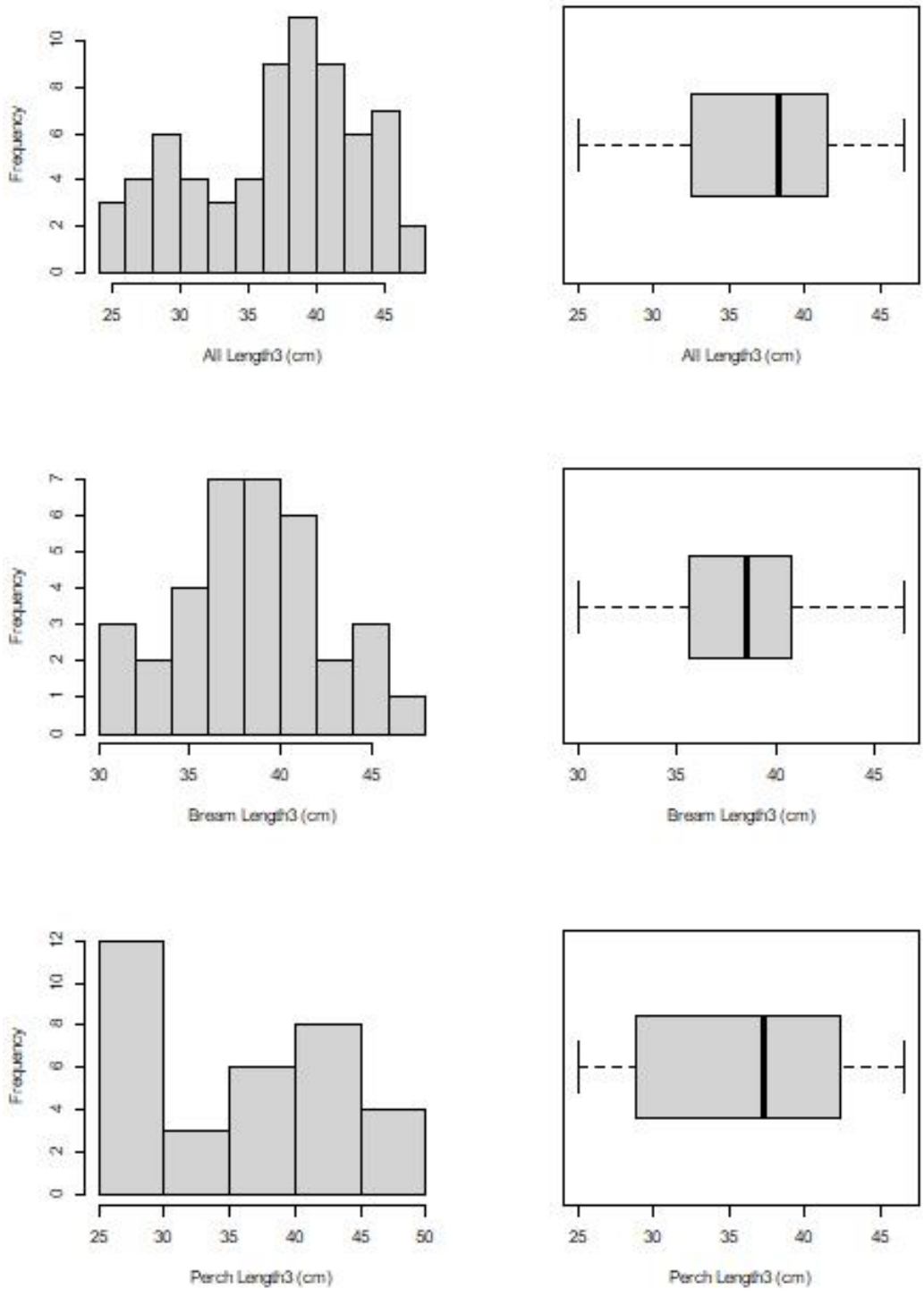


Figure 4. Histograms and boxplots for each of Length3 overall and Length3 by species.

Handout with Supplementary Material for Quiz 1 Problem 2

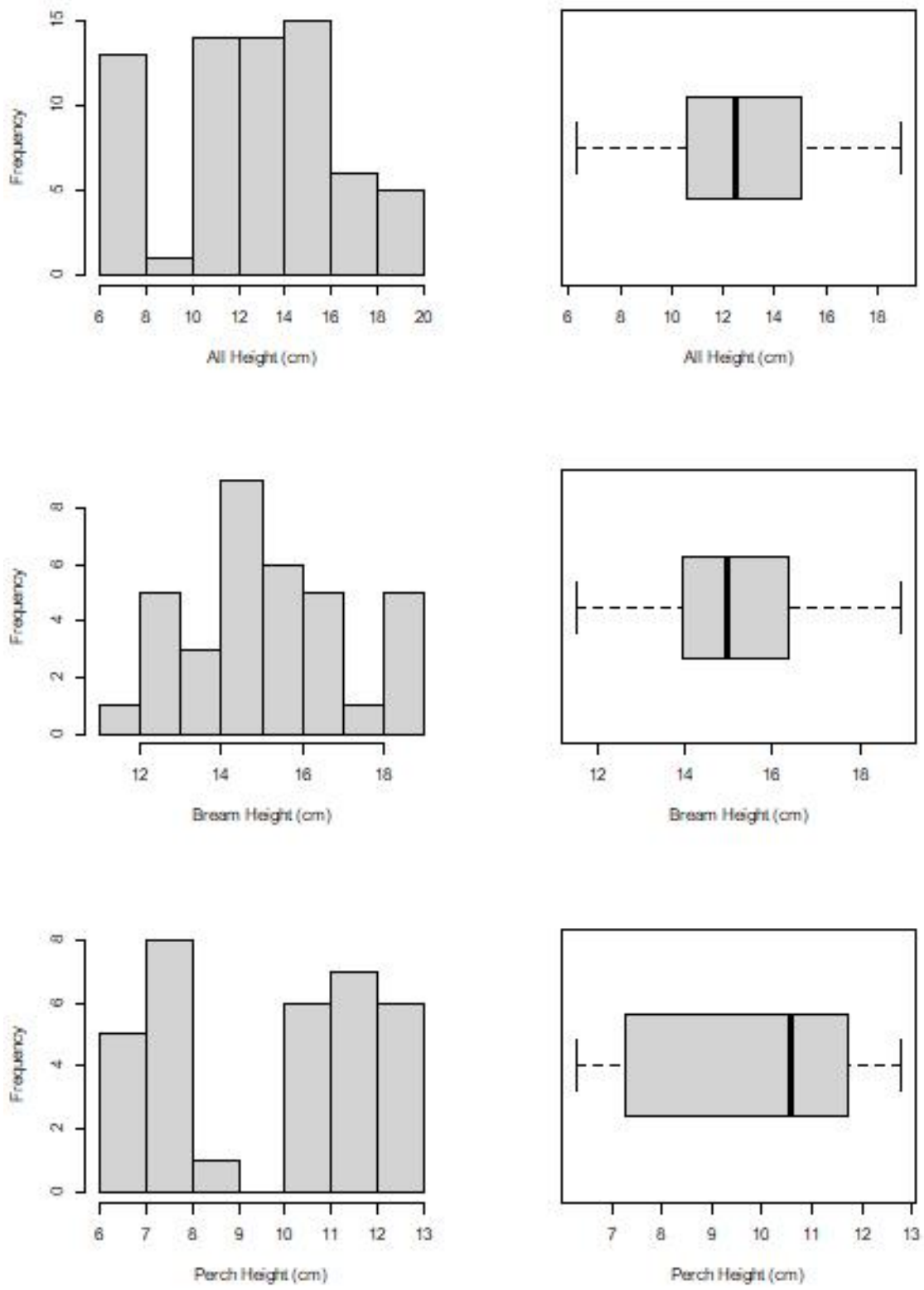


Figure 5. Histograms and boxplots for each of Height overall and Height by species.

Handout with Supplementary Material for Quiz 1 Problem 2

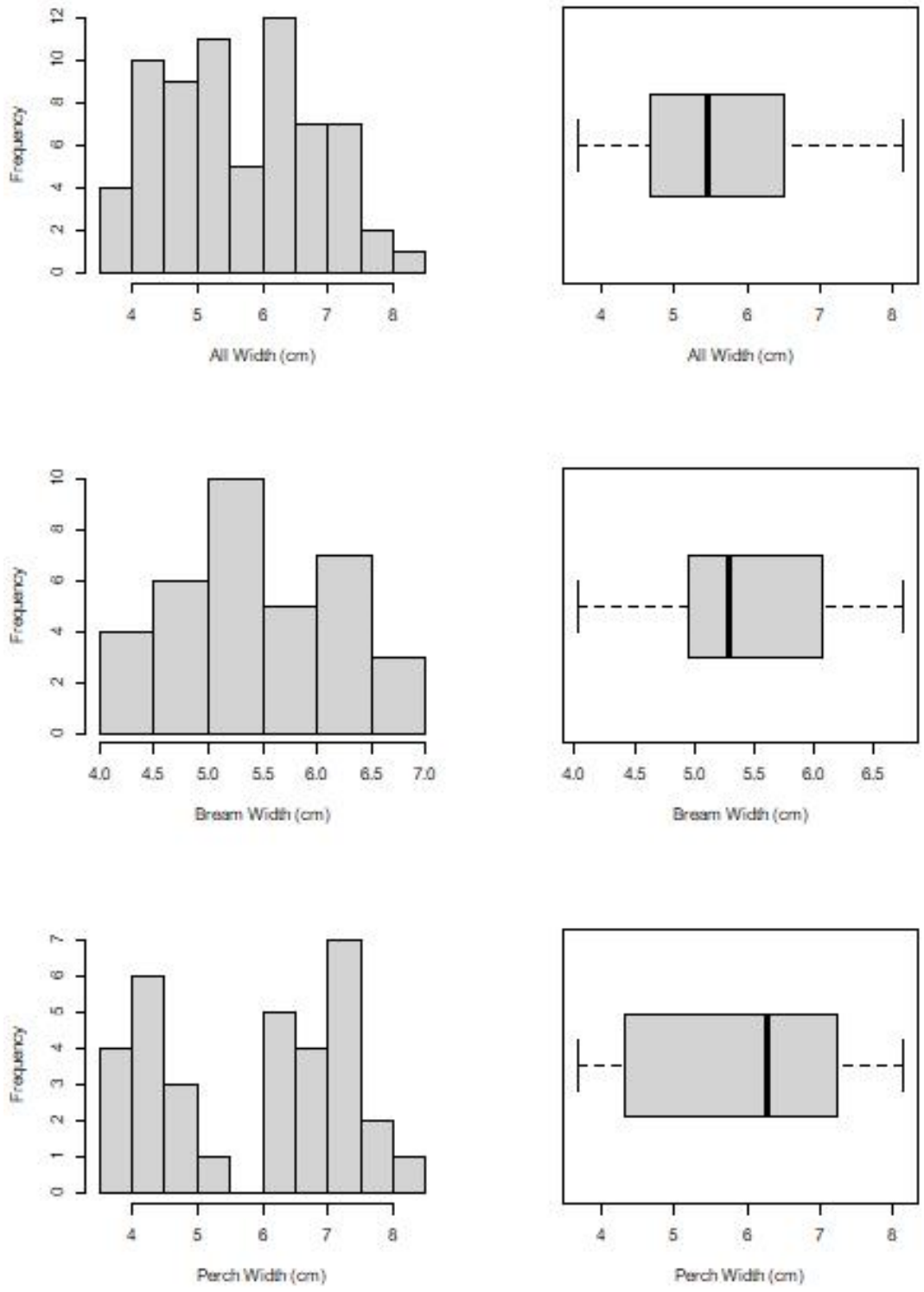


Figure 6. Histograms and boxplots for each of Width overall and Width by species.

Handout with Supplementary Material for Quiz 1 Problem 2

Table 2. Seven number summary statistics for the Weight (*y*) and possible predictors (*x*). The total sample size is 68 fish, 35 Bream and 33 Perch.

Variable	Min	Q1	Median	Q3	Max	Mean	SD
Weight (g)							
All	145	340	615	825	1100	599.0	273.55
Bream	242	462.5	610	717	1000	617.8	209.21
Perch	145	250	650	850	1100	579.0	330.72
Length3 (cm)							
All	25	33.0	38.3	41.4	46.6	37.0	6.03
Bream	30	35.7	38.5	40.8	46.5	38.4	4.16
Perch	25	28.9	37.3	42.3	46.6	35.6	7.34
Height (cm)							
All	6.3	10.6	12.5	15.0	19.0	12.5	3.51
Bream	11.5	14.0	15.0	16.4	19.0	15.2	1.96
Perch	6.3	7.3	10.6	11.7	12.8	9.7	2.33
Width (cm)							
All	3.7	4.7	5.5	6.4	8.1	5.6	1.16
Bream	4.0	4.9	5.3	6.1	6.7	5.4	0.72
Perch	3.7	4.3	6.3	7.2	8.1	5.8	1.48

Handout with Supplementary Material for Quiz 1 Problem 2

Single Variable Linear Model Output

Weight on Cross Length3

```
Call:  lm(formula = Weight ~ Length3, data = FDatSub2)

Coefficients:
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) -1005.264     61.596    -16.32     <2e-16
Length3        43.322      1.642     26.38     <2e-16

Residual standard error: 81.11 on 66 degrees of freedom
Multiple R-squared:  0.9134
```

Weight on Height

```
Call:  lm(formula = Weight ~ Height, data = FDatSub2)

Coefficients:
            Estimate Std. Error   t value   Pr(>|t|)
(Intercept)   -26.48      95.50    -0.277      0.782
Height         50.04       7.36     6.799   3.68e-09

Residual standard error: 211.4 on 66 degrees of freedom
Multiple R-squared:  0.4119
```

Weight on Diagonal Width

```
Call:  lm(formula = Weight ~ Width, data = FDatSub2)

Coefficients:
            Estimate Std. Error   t value   Pr(>|t|)
(Intercept)  -611.79      67.03    -9.127   2.59e-13
Width         215.27      11.67    18.440    < 2e-16

Residual standard error: 111.1 on 66 degrees of freedom
Multiple R-squared:  0.8375
```

Handout with Supplementary Material for Quiz 1 Problem 2

Scatterplots with Regression Lines

Figure 7 shows scatterplots of Weight on each the three predictor variables with regression lines. The two scatterplots in the top row have three lines. The colored lines are the regression lines fit to individual species (these were done in ggplot2). The black lines are the regression lines for the single variable regressions on the previous page. The scatterplot of Weight on Width has two additional color-coded dashed lines. These are the regression lines for each species if we assume a common slope (parallel lines) and only change the intercept. The solid color-coded lines allow for different slopes, which also implies different intercepts (some of the dashed lines are hard to see as they practically coincide with the solid lines). How to set up these different regressions will be explained on the last two pages of this document. We chose to include this to show you there is more you can do than single variable regression.

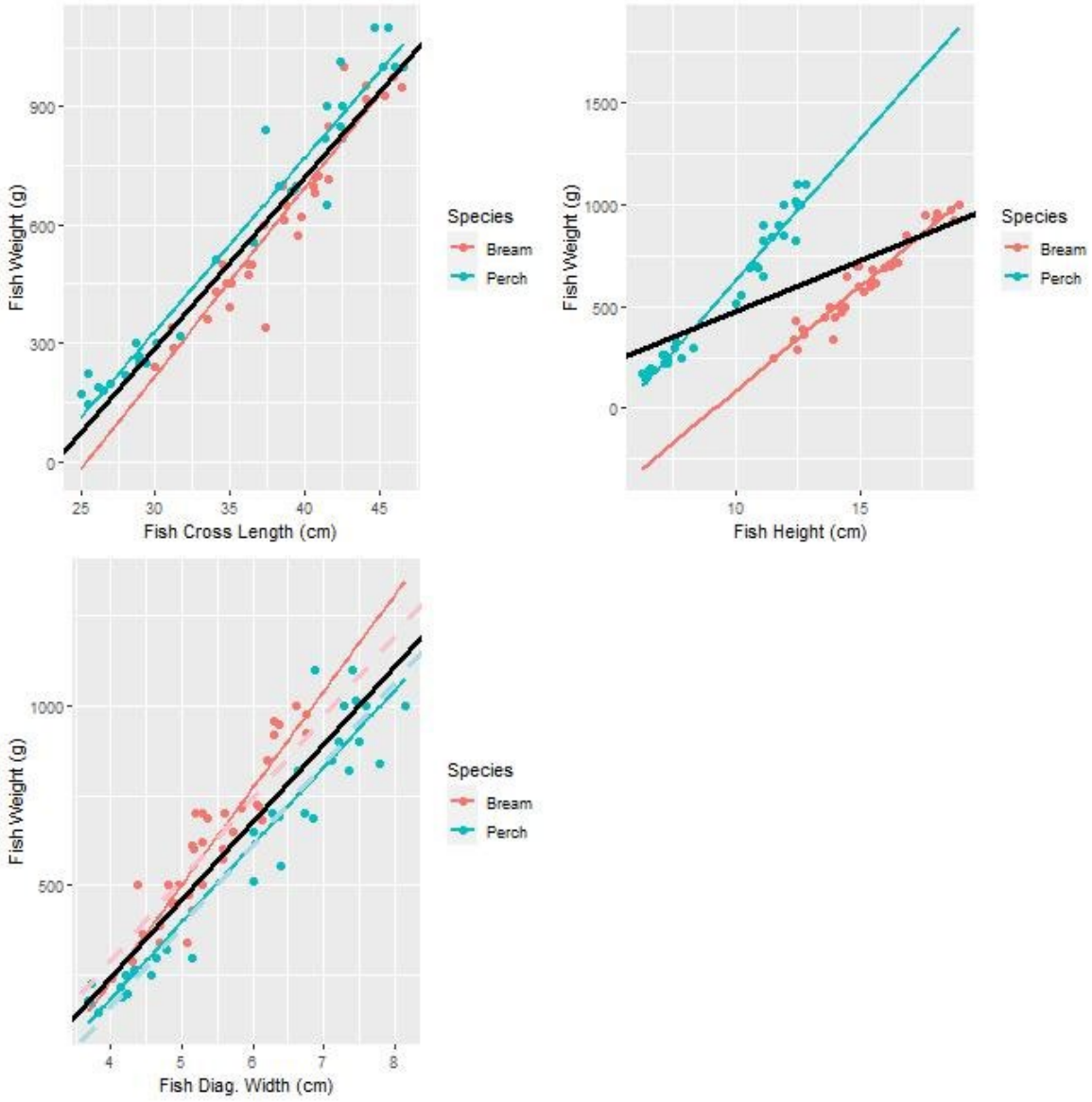


Figure 7. Scatterplots of Weight on each of the predictors with fitted regression lines.

Handout with Supplementary Material for Quiz 1 Problem 2

Residual Diagnostic Plots

Figure 8 shows residuals plots and histograms of the residuals for each of the univariate regressions (black lines in Figure 3).

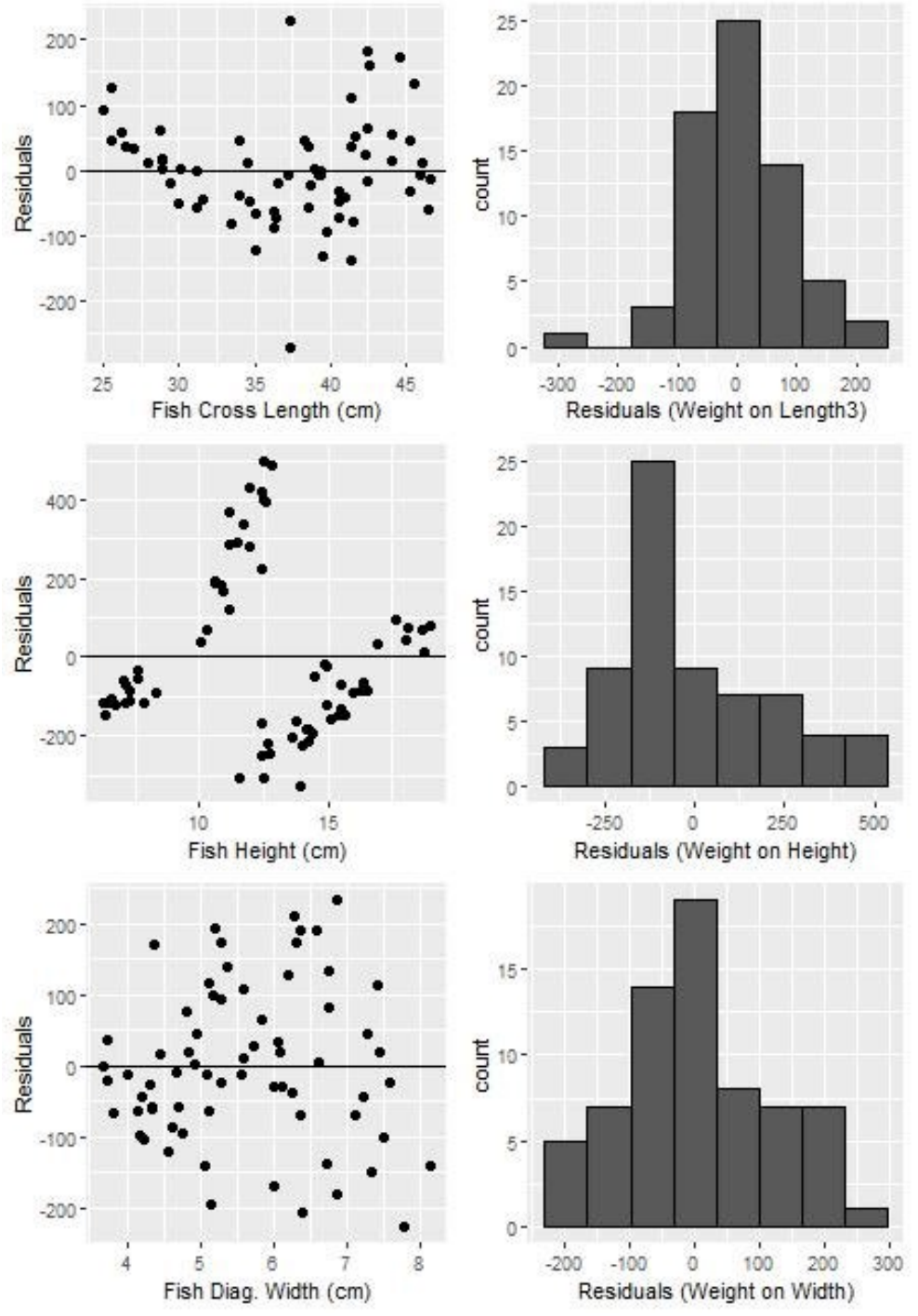


Figure 8. Residual plots and histograms of residuals for each of the three univariate regressions.

Handout with Supplementary Material for Quiz 1 Problem 2

Linear Model Output Accounting for Species

Weight on Width and Species allowing for different intercepts only (same slope/parallel lines)

In this regression, we add a second categorical variable to the model. This allows us to fit two lines, one for each species. In R, the first level of the category is included in the y-intercept. Coefficients for the other levels show up in the output. Thus, in the output below, we have an additional line the Perch species (SpeciesPerch) Since Species is categorical, this coefficient is added to the y-intercept to get the intercept for Perch. The two regression lines that result from this model are shown underneath the output. When we simply add a categorical variable, we are assuming parallel line, so only one overall slope for Weight and different y-intercepts for each Species.

```
Call:  lm(formula = Weight ~ Width + Species, data = FDatSub2)

Coefficients:
             Estimate Std. Error
(Intercept)  -604.299     54.799
Width         225.169      9.692
SpeciesPerch -130.145     22.385

Residual standard error: 90.82 on 65 degrees of freedom
Adjusted R-squared:  0.8898
```

Regression equation for Bream: $\hat{y} = -604.299 + 225.169(Width)$

Regression equation for Perch: $\hat{y} = (-604.299 - 130.145) + 225.169(Width)$ or
$\hat{y} = -734.444 + 225.169(Width)$

Handout with Supplementary Material for Quiz 1 Problem 2

Weight on Width and Species allowing for different slopes (and thus different intercepts)

In this regression, we add a second categorical variable to the model, but we use an asterisk instead of a plus. This allows us to fit two lines, one for each species, but allowing for different slopes for each species. Thus, in the output below, we have two additional lines for the Perch species: 1) SpeciesPerch, which is used to get the y-intercept for Perch, and 2) Width:SpeciesPerch, which allows us to fit a different slope for Perch. This is called in interaction term. The two regression lines that result from this model are shown underneath the output. You can see how the y-intercept and SpeciesPerch were combined to get the y-intercept for the Perch line and Width and Width:SpeciesPerch were combined to the slope for the Perch line.

```
Call:  lm(formula = Weight ~ Width * Species, data = FDatSub2)

Coefficients:
                    Estimate Std. Error
(Intercept)          -838.39     114.42
Width                 268.30      20.90
SpeciesPerch          167.45     130.69
Width:SpeciesPerch    -54.02      23.39

Residual standard error: 87.94 on 64 degrees of freedom
Adjusted R-squared:  0.8967
```

Regression equation for Bream: $\hat{y} = -838.39 + 268.30(Width)$

Regression equation for Perch: $\hat{y} = (-838.39 + 167.45) + (268.30 - 54.02)(Width)$ or
$\hat{y} = -670.94 + 214.28(Width)$