13.

Null Hypothesis: We expect that the number of plant species in Galapogous islands are not impacted by ***Endemics20e:*** the number of endemic species, ***Area:*** the area of the island (km2), ***Elevation:*** the highest elevation of the island (m), ***Neqrest:*** the distance from the nearest island (km), ***Scruz:*** the distance from Santa Cruz island (km), ***Adjacent:*** the area of the adjacent island (km2).

Q13 11

We need to verify whether the null hypothesis is true. Also, we need to identify which is/are the best predictors.

How would you know that the total plant species richness does not include the same set of endemic species found on the island.

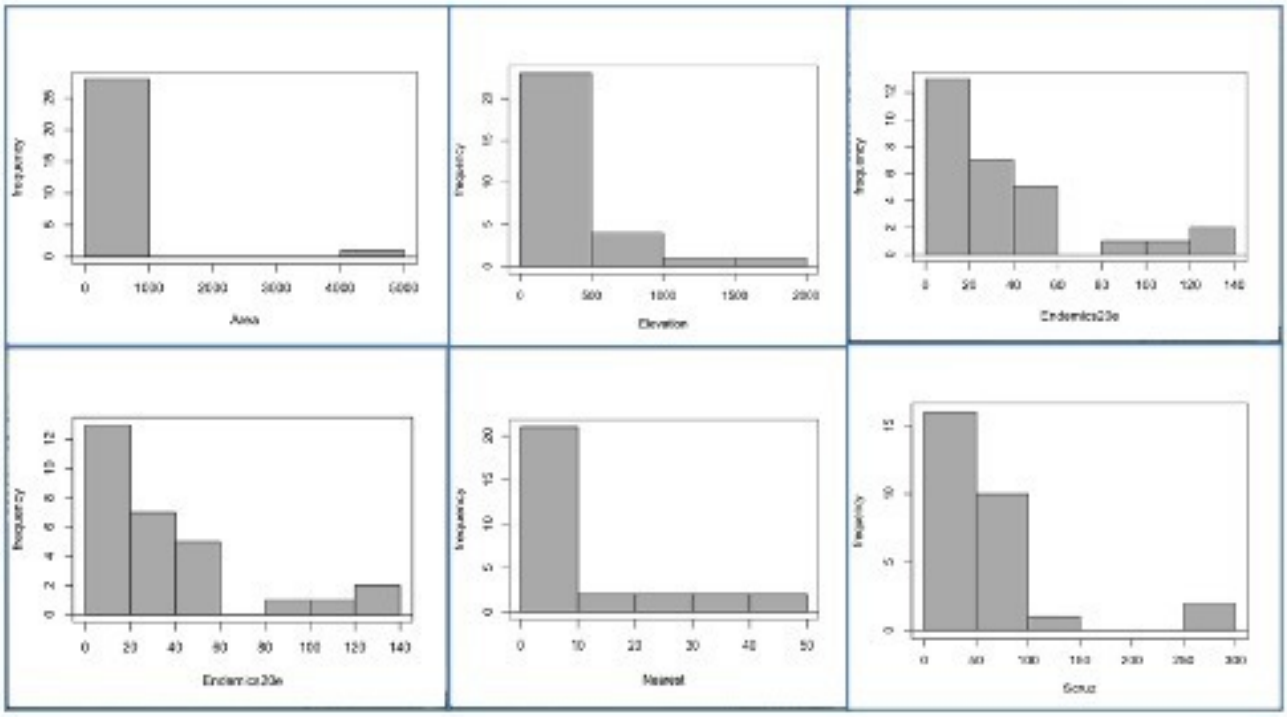

Figure 1a. Histograms: All histogram plots show non-normality because the graphs are right-skewed.

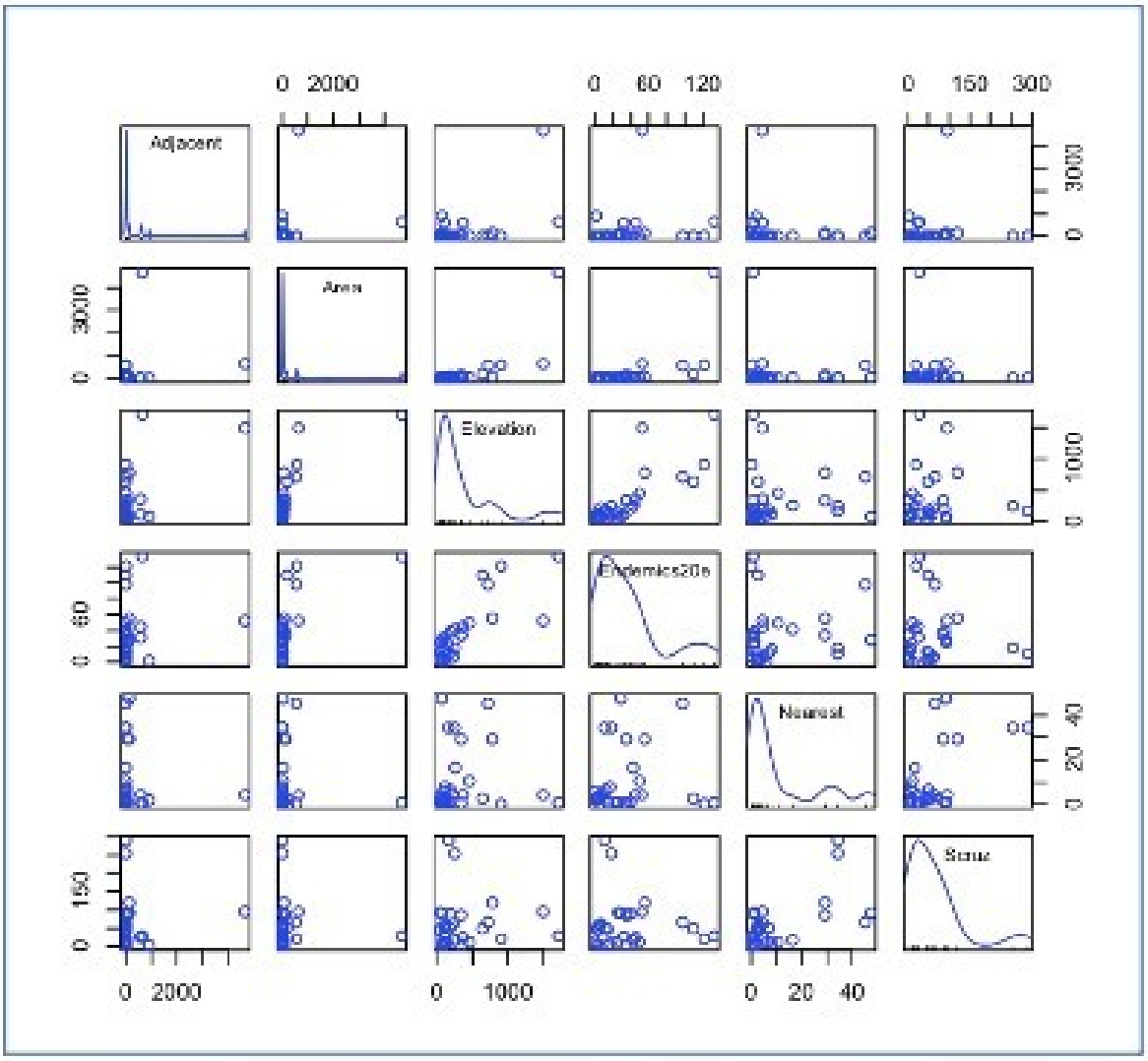

Figure 1b. Scatterplot matrix of the correlation plots of the independent variables. All the independent variable scatterplots are right skewed.

Both the histograms and scatterplots suggest non-normality, hence, a transformation is necessary. Since all the graphs are right-skewed, log transformation would be the most appropriate. However, not all graphs need to be transformed. This is because according to Island Biogeographic Theory, only the distance to the mainland (Scruz), size of an island (Area) and relative island isolation (Nearest) have the capability to have an impact on the plant species on the island. Hence, the only variables we are examinining and transforming are Scruz, Area and Nearest.

O.K. 2.75

Elevation is also a geographic variable and would have an influence both on plant species dispseral and the area of available habitat types. Some mention of elevation should be included.

important to include a rationale for why to include these variables and exclude other ones.

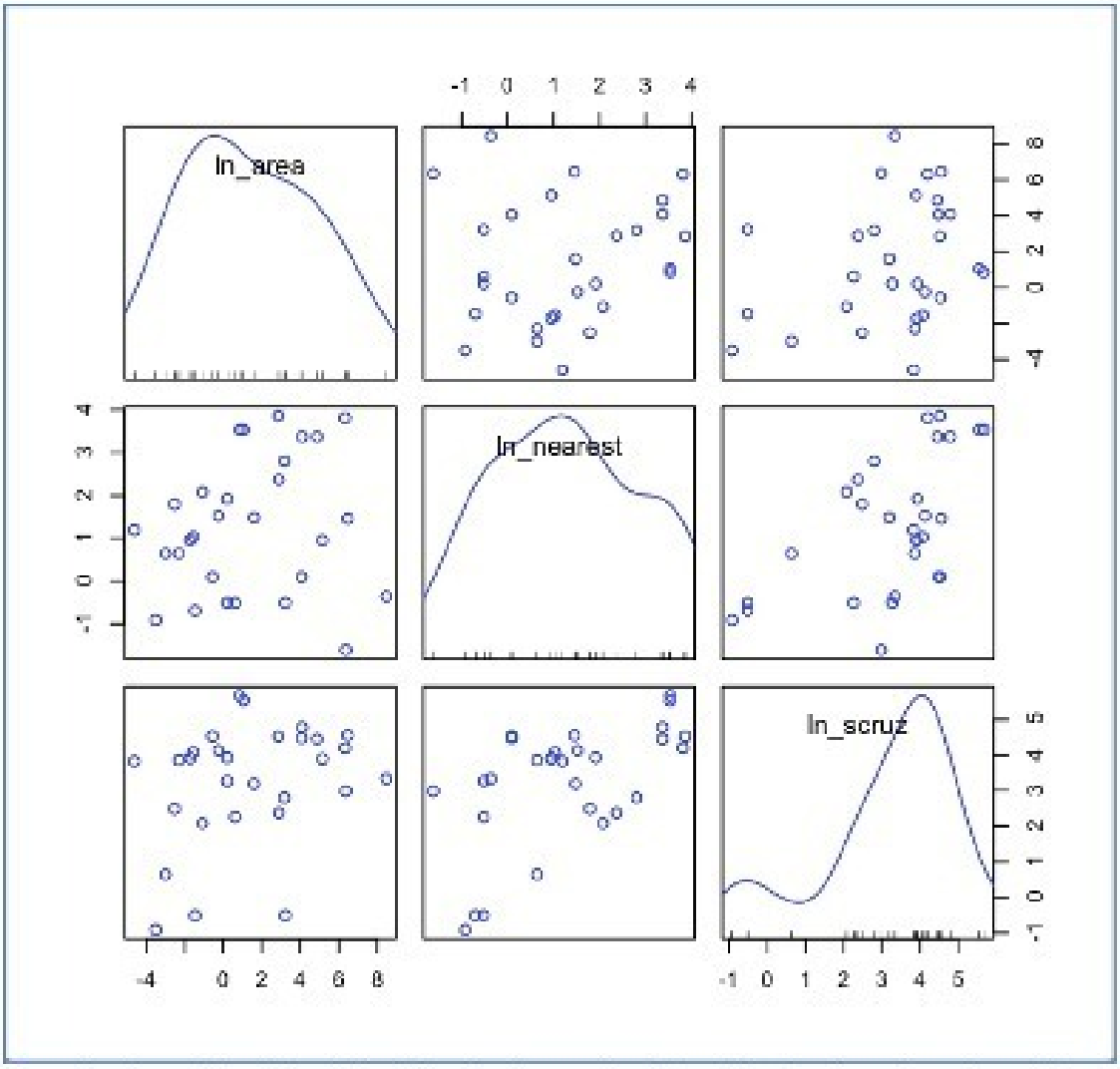

Figure 1c. The scatterplot matrix of the transformed variables. After log transformation of the variables area, nearest and Scruz, the scatterplots now exhibit normality.

```
Rcmdr>  RegModel.2 <- lm(Species20e~ln_area+ln_nearest+ln_scruz, data=galapagous)

Rcmdr>  summary(RegModel.2)

Call:
lm(formula = Species20e ~ ln_area + ln_nearest + ln

Residuals:
     Min       1Q   Median       3Q      Max
-141.878  -47.391   -8.255   31.052  190.779

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept)   83.798     34.386   2.437       0.0223 *
ln_area       34.196      4.889   6.995 0.000000248 ***
ln_nearest   -19.933     12.231  -1.630       0.1157
ln_scruz       1.313     11.711   0.112       0.9116
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.22 on 25 degrees of freedom
Multiple R-squared:  0.6821,    Adjusted R-squared:  0.644
F-statistic: 17.88 on 3 and 25 DF,  p-value: 0.000002069
```

Transform

Check response variable for normality and transform as well. If this is not done, definitely need to check residuals to see if they deviate from normality to ensure assumptions are met.

25

Figure 1d. The global model of ln(area), ln(nearest) and ln(scruz) as a function of Species20e.

```
Rcmdr>  stepwise(RegModel.2, direction='backward/forward', criterion='AIC')

Direction:  backward/forward
Criterion:  AIC

Start:  AIC=260.83
Species20e ~ ln_area + ln_nearest + ln_scruz

            Df Sum of Sq    RSS    AIC
- ln_scruz   1        89 177404 258.85
<none>                  177315 260.83
- ln_nearest  1     18838 196153 261.76
- ln_area     1    347002 524317 290.27

Step:  AIC=258.85
Species20e ~ ln_area + ln_nearest

            Df Sum of Sq    RSS    AIC
<none>                  177404 258.85
- ln_nearest  1     25164 202567 260.69
+ ln_scruz    1        89 177315 260.83
- ln_area     1    376263 553667 289.85

Call:
lm(formula = Species20e ~ ln_area + ln_nearest, data = galapagous)

Coefficients:
(Intercept)      ln_area   ln_nearest
      86.87        34.34       -19.17
```

Figure 1e. The stepwise model selection (backward/forward) of the model.

6

```
Rcmdr>  LinearModel.4 <- lm(Species20e ~ ln_area + ln_nearest +ln_area*ln_nearest, data=galapogous)

Rcmdr>  summary(LinearModel.4)

Call:
lm(formula = Species20e ~ ln_area + ln_nearest + ln_area * ln_nearest,
    data = galapogous)

Residuals:
     Min       1Q   Median       3Q      Max
-139.997  -49.183   -8.327   31.532  193.518

Coefficients:
                  Estimate Std. Error t value    Pr(>|t|)
(Intercept)        86.0594    22.0468   3.903    0.000635 ***
ln_area            34.6254     5.3696   6.448 0.000008942 ***
ln_nearest        -18.0778    14.1961  -1.273    0.214578
ln_area:ln_nearest  -8.3851     3.4817  -0.111    0.912791
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.22 on 25 degrees of freedom
Multiple R-squared:  0.6821,    Adjusted R-squared:  0.644
F-statistic: 17.88 on 3 and 25 DF,  p-value: 0.000002069
```

Figure 1f. Linear regression of the model with only the relevant terms and the interaction term.

```
Call:
lm(formula = Species20e ~ ln_area + ln_nearest, data = galapogous1)

Residuals:
     Min       1Q   Median       3Q      Max
-141.001  -49.373   -7.969   27.983  189.409

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   86.869     20.398   4.259     0.000238 ***
ln_area       34.341      4.624   7.426 0.0000000694 ***
ln_nearest   -19.173      9.984  -1.920     0.065839 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 82.6 on 26 degrees of freedom
Multiple R-squared:  0.682,     Adjusted R-squared:  0.6575
F-statistic: 27.88 on 2 and 26 DF,  p-value: 0.0000003403
```

Figure 1g. Linear regression of the model with the relevant terms only but without the interaction term

13. MLR or multiple linear regression is the most appropriate test for this data because of the mechanistic property of the situation. Figure 1a and 1b show that the population distribution is non-normal for all the predictor variables. This means that a transformation is needed so meet the assumptions of normality. After the appropriate transformation was done, a linear regression was ran and it was found that the variable "Area" is the most significant, as it is below the threshold of alpha=0.05. The p-value of this variable in the global model of linear regression is p=2.48x10^-7. The variables "Nearest" and "Scruz" have p-values of 0.1157 and 0.9116, respectively. Hence, these two variables are considered non-significant. This global model also has an adjusted R-squared of 0.644, which means 64% of the error is accounted for by the model. To identify the relevance of the variables "Area", "Nearest" and "Scruz", a stepwise backward/forward selection was ran. It was identified that all variables are relevant except for "Scruz". This means that the variable "Scruz" is statistically irrelevant and must be removed from the global model because it is taking up the degrees of freedom that the other variables should be taking. Figure 1f shows the linear regression model of the relevant variables, "Area" and "Nearest", as well as the interaction term of the two. We can see that the interaction term between these 2 variables have a p-value of 0.91, a value that is a lot greater that alpha=0.05. Therefore, we can also say that the interaction between the "Area" and "Nearest" is not significant. Figure 1g shows the linear regression model that only includes the variables "Area" and "Nearest". To sum this up, the area of the island in km^2 is highly significant because it has a p-value of 6.94x10^-8, which is less than the alpha 0f 0.05. For every additional unit of the island area, plant species will increase an estimated 34.341 units, controlling for the variable "nearest", which is the distance to the nearest island in km. In addition, the distance to the nearest island is marginally significant, with a p-value of 0.065. For every one additional unit of the variable "nearest", the plant species decrease by approximately 19.173, controlling for the variable "area". The adjusted R-squared value for this final model is 0.6575, which means the final regression table accounts for about 66% of the error of the model.

14.

```
Call:
glm(formula = Matuation ~ FLENmm + Lake + SEX + Year, family = binomial(logit),
    data = wefish)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
Q14. 13  -0.4402  -0.1531   0.2246   3.4625

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             137.657259  85.055495   1.618    0.106
FLENmm                    0.030159   0.001789  16.860   <2e-16 ***
Lake[T.Nipigon, L.]       0.523483   1.363650   0.384    0.701
Lake[T.of the Woods, L.]  0.776647   1.361743   0.570    0.568
Lake[T.Rainy L.]          0.120281   1.380243   0.087    0.931
SEX[T.male]               1.593754   0.212273   7.508    6e-14 ***
Year                     -0.075298   0.042581  -1.768    0.077 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1536.26  on 1257  degrees of freedom
Residual deviance:  732.59  on 1251  degrees of freedom
AIC: 746.59

Number of Fisher Scoring iterations: 6


Rcmdr>  exp(coef(GLM.2))  # Exponentiated coefficients ("odds ratios")
             (Intercept)                    FLENmm      Lake[T.Nipigon, L.]
            6.078379e+59             1.030619e+00             1.687896e+00
Lake[T.of the Woods, L.]        Lake[T.Rainy L.]              SEX[T.male]
            2.174171e+00             1.127814e+00             4.922190e+00
                    Year
            9.274667e-01
```

Figure 2a. Generalized linear regression model of WEfishmaturity.

```
Rcmdr>  stepwise(GLM.2, direction='backward/forward', criterion='AIC')

Direction:  backward/forward
Criterion:  AIC

Start:  AIC=746.59
Matuation ~ FLENmm + Lake + SEX + Year

         Df Deviance     AIC
- Lake    3   738.44  746.44
<none>        732.59  746.59
- Year    1   735.74  747.74
- SEX     1   797.67  809.67
- FLENmm  1  1500.22 1512.22

Step:  AIC=746.44
Matuation ~ FLENmm + SEX + Year

         Df Deviance     AIC
<none>        738.44  746.44
+ Lake    3   732.59  746.59
- Year    1   741.58  747.58
- SEX     1   800.79  806.79
- FLENmm  1  1523.01 1529.01

Call:  glm(formula = Matuation ~ FLENmm + SEX + Year, family = binomial(logit),
    data = wefish)

Coefficients:
(Intercept)       FLENmm  SEX[T.male]         Year
  135.49486      0.02974      1.54730     -0.07383

Degrees of Freedom: 1257 Total (i.e. Null);  1254 Residual
Null Deviance:      1536
Residual Deviance: 738.4        AIC: 746.4
```

Figure 2b. The stepwise model selection (backward/forward) of the model.

2

```
Rcmdr>  summary(GLM.3)

Call:
glm(formula = Matuation ~ FLENmm + SEX + Year, family = binomial(logit),
    data = wefish)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1210  -0.4453  -0.1527   0.2258   3.4160

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 135.494856  83.625359   1.620   0.1052
FLENmm        0.029736   0.001765  16.847  < 2e-16 ***
SEX[T.male]   1.547297   0.209879   7.372 1.68e-13 ***
Year         -0.073830   0.041818  -1.766   0.0775 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1536.26  on 1257  degrees of freedom
Residual deviance:  738.44  on 1254  degrees of freedom
AIC: 746.44

Number of Fisher Scoring iterations: 6
```

Figure 2c. Generalized linear model with only the relevant variables (FLENmm, SEX and Year) included.

```
Rcmdr>  exp(coef(GLM.3))  # Exponentiated coefficients ("odds ratios")
 (Intercept)       FLENmm  SEX[T.male]         Year
6.993075e+58 1.030183e+00 4.698754e+00 9.288295e-01
> ll.null <- GLM.3$null.deviance/-2
> GLM.3$df.null
[1] 1257
> ll.proposed <- GLM.3$deviance/-2
> GLM.3$df.residual
[1] 1254
> (ll.null - ll.proposed) / ll.null
[1] 0.5193245
> 1 - pchisq(2*(ll.proposed - ll.null), df=length(GLM.3$coefficients)-1)
[1] 0
```

Figure 2d. The calculation of the null deviance, residual deviance, McFadden's Pseudo R^2, and p-value.

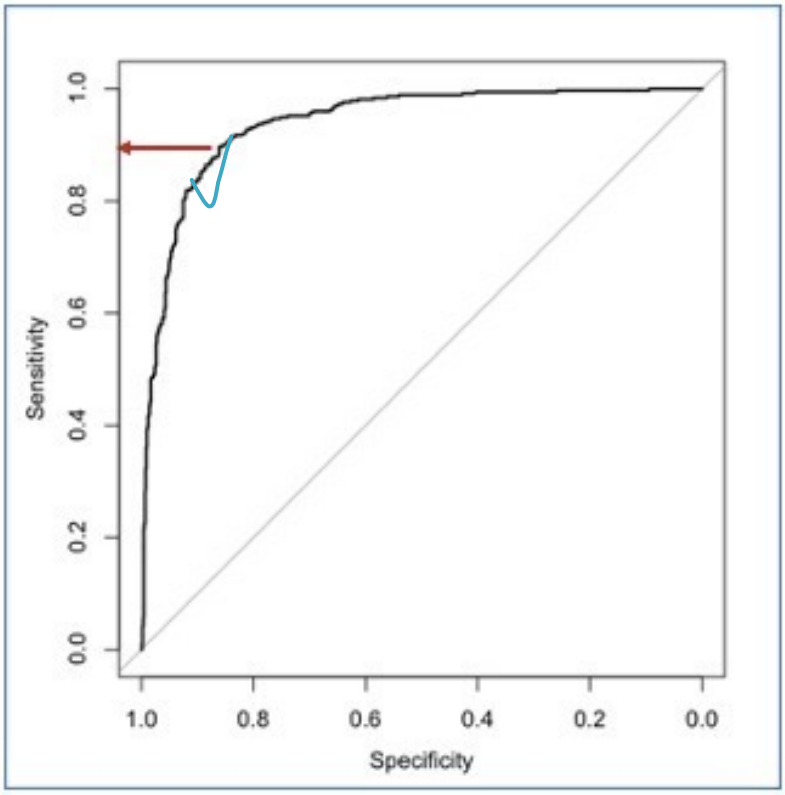

Figure 2e. ROC curve of the Walleye Fish Maturity. The optimal threshold is around 0.9.

```
Call:
roc.default(response = wefish$Matuation, predictor = GLM.3$fitted.values,     plot = TRUE)

Data: GLM.3$fitted.values in 881 controls (wefish$Matuation 0) < 377 cases (wefish$Matuation 1).
Area under the curve: 0.94
```

Figure 2f. The area under the ROC curve is identified to be 0.94.

Conclusion

After running the initial logistic regression, it was found that the fork length (FLENmm) and sex (SEX) have the most significant impact, with p-values of <2e-16 and 6e-14 (alpha=0.05). Year is also marginally significant, with a p-value of 0.077. A backward/forward stepwise selection was ran to identify the best predictor variable(s) of Walleye fish maturity. A backward/forward stepwise selection was run instead of forward/backward because the number of variables under consideration is smaller than the sample size. Also, this stepwise selection considers the effects of all the variables simultaneously. It was found that all variables except for the lake are the most relevant predictors of fish maturity, with an AIC of 746.4. Another logistic regression was ran and it was found that sex (p=1.68e-13) and fish length (p<2e-16) still have the most significant impact (alpha=0.05), and that year is still also marginally significant (p=0.0775). We can then conclude that for every one additional change in fork length, the probability of catching a mature fish is about 0.0297, controlling for the sex and year; and for every one unit increase in the number of the sex male, the probability of getting a mature fish is approximately 1.55, controlling for the fork length and year. The null deviance of the model is 1257 and the residual deviance of the model is 1254. From these 2 values, the McFadden's Pseudo R^2 is calculated to be 0.5193, which indicates a relatively optimal but not particularly strong likelihood value for the model. Moreover, the p-value of Mcfadden's Pseudo R^2 is exactly 0, which means that the model describes a significant relationship in the model (alpha=0.05). To reiterate, this is a confirmation that sex and fork length are the best predictors of fish maturity. Figure 2e shows the ROC curve of Walleye fish maturity, with ~0.9 as the optimal threshold. The graph is above the gray line, which means that we have a reasonably good fit model. The area under the curve (AUC) is also identified to be 0.94, which means that the model has an accuracy of 94%.