













1. Kruskal-Wallis 
2. ANCOVA 
- Q3. Spea  and Pearson's test 
4. MAN 
5. Mann-Whitney test 
6. MLR – multiple linear regression 
7. Population variance 
8. Non-parametric test 
9. It is the Beta coefficient or the slope parameter for the independent variable. We report  
is as "For one-unit increase in x, we predict that there's a B-unit increase or decrease in y"
10. Model sensitivity is the rate at which we get a true positive in our model. It tells us which proportion of positive samples were correctly classified It is the ratio between the true positives and the sum of true positives and false negatives. 

Q1. Friedman's test -1

Q4. MANCOVA -1

Q9. Partial regression coefficient -0.5

11a. and 11b.

Q11 6

Aquarium_No	Responsiveness_score								
1	10	1	10	2	44	3	70		
1	22	1	22	2	48	3	76		
1	25	1	25	2	56	3	97		
1	34	1	34	2	76	3	98		
1	50	1	50	2	40	3	80		
2	44	Average1	28.2	Average2	52.8	Average3	84.2		
2	48								
2	56								
2	76								
2	40								
3	70								
3	76								
3	97								
3	98								
3	80								
Grand Total	826								
Grand Average	55.06666667								

Q11a. G1= 3.6; G2= 7.7; G3= 12.7 -1

11a. The average mean for Aquariums 1, 2 and 3 are 28.2 52.8 and 84.2, respectively. X

11b. The grand mean of ranks is 55.1. X

Q11b. Grand mean of ranks = 8 -1

Q11c. SSamong = 207.7 -2

11e. The H-stat value for the model is found to be 10.385 and the χ^2 crit (from Chi distribution table) for that H-stat within w/c $0.05 > p < 0.1$ is found to be between 5.991 and 4.605. Since H-stat is greater than χ^2 crit, we reject the null hypothesis, which is the hypothesis that the complexity of dolphin enclosures does not affect the level of dolphin responsiveness. To state another way, we are confident that there is a statistical difference in responsiveness among dolphins in the different aquariums.

Q11e. Note that the H-stat falls between the $0.005 < p < 0.01$ range (H = 10.385; df = 2)

-1

$$\begin{aligned}
 11c) \quad SS_{\text{among}} &= \sum_{i=1}^a n_i (\bar{R}_i - \bar{\bar{R}})^2 \\
 &= 5(3.6 - 8)^2 + 5(7.7 - 8)^2 + 5(12.7 - 8)^2 \\
 &= 5(19.36) + 5(0.09) + 5(22.09) \\
 &= 207.7 \quad \checkmark R \quad \checkmark I
 \end{aligned}$$

$$\begin{aligned}
 11d) \quad H &= \frac{SS_{\text{among}}}{\sigma^2}, \quad \sigma^2 = \frac{N(N+1)}{12} \\
 H &= \frac{207.7}{20} \quad \sigma^2 = \frac{15(15+1)}{2} \\
 & \quad \quad \quad = 20
 \end{aligned}$$

$$H = 10.385$$

$$\begin{aligned}
 11e) \quad df &= K - 1 = 15 - 1 = 14 = 2 \\
 @ df = 14, \quad 0.05 > p < 0.1, \quad 4.605 > \chi^2_{\text{crit}} > 5.991
 \end{aligned}$$

$$H > \chi^2_{\text{crit}} \quad \therefore \text{reject the null}$$



12.

What type of analysis is this? (1 pt) Describe the predictors in terms of the type of effect examined.

Q12 This 2 ANCOVA analysis test. The response variable in this model is the sea otter abundance, the independent effectors are location and year. ✓

What should the researchers conclude? (2 pts)

Based on the initial ANCOVA test ($p=0.05$), location is not significant, with a p-value of 0.191, the interaction between location and the year are not significant, with a p-value of 0.197, but year is significant, with a p-value of 0.006. However, although year is significant, we are not really interested in it because we know, by nature, that number of sea otters naturally change every year. So, we will leave the location in when rerunning another ANCOVA test despite it being insignificant. The initial ANCOVA run has an R^2 of 0.8782, meaning this model accounts for 87.82% of the variance of the model. The overall p-value of the initial ANCOVA run is 4.9×10^{-9} , meaning the model is overall significant. After the ANCOVA rerun, it was found that location is in fact significant, with a p-value of 1.07×10^{-9} . This means that the presence of killer whales in both lagoon and bay have a significant impact on sea otter decline. The R^2 of the 2nd rerun is still around 0.87, and the overall the model is statistically significant ($p=1.175 \times 10^{-9}$). In addition, according to the graph, sea otter number decrease over the years in both lagoon and bay, in slightly different rates (different slopes). ✓

Killer whales were present in the bay, but not in the lagoon

Try to put a bit more emphasis on what this means for the otter populations ecologically. What could the researchers conclude from year and location being significant factors in otter population decline?

-1

13.

Null Hypothesis: The number of plant species in Galapagos islands are not predicted by

Endemics20e: the number of endemic species, **Area:** the area of the island

(km²), **Elevation:** the highest elevation of the island (m), **Nearest:** the distance from the nearest

island (km), **Scruz:** the distance from Santa Cruz island (km), and **Adjacent:** the area of the

Q13

11

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

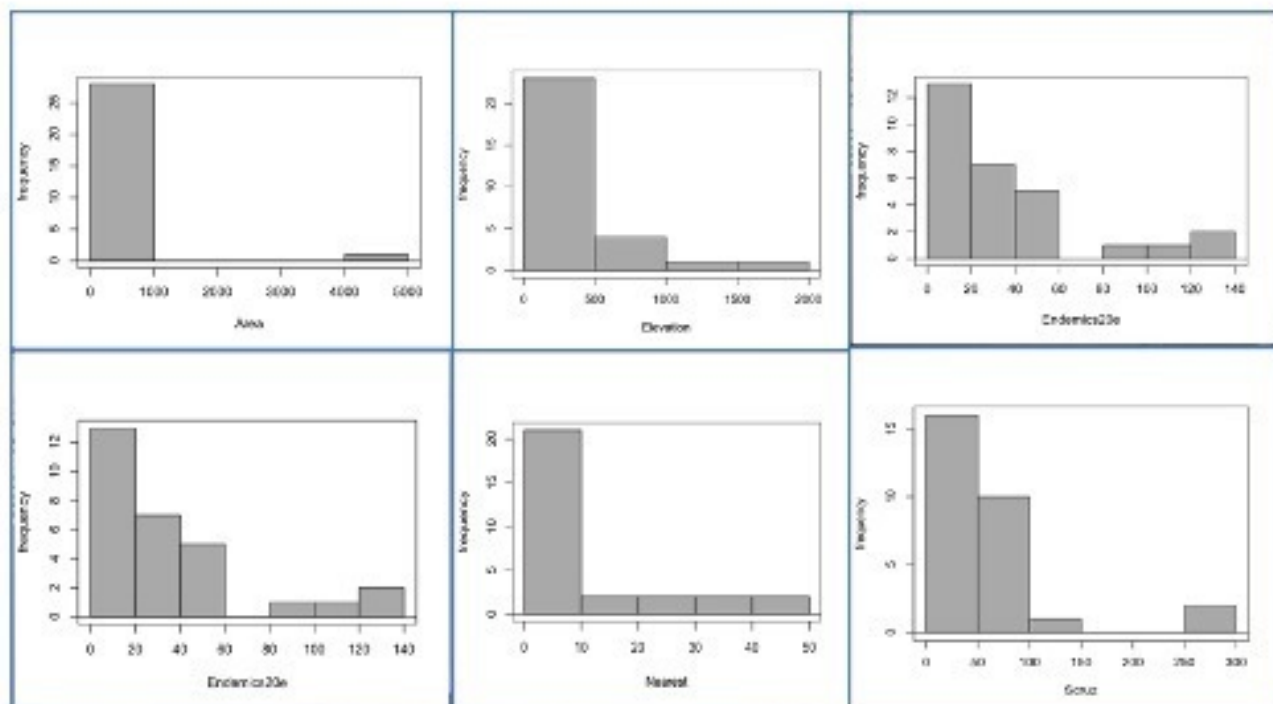


Figure 1a. Histograms: All histogram plots show non-normality because the graphs are right-skewed.

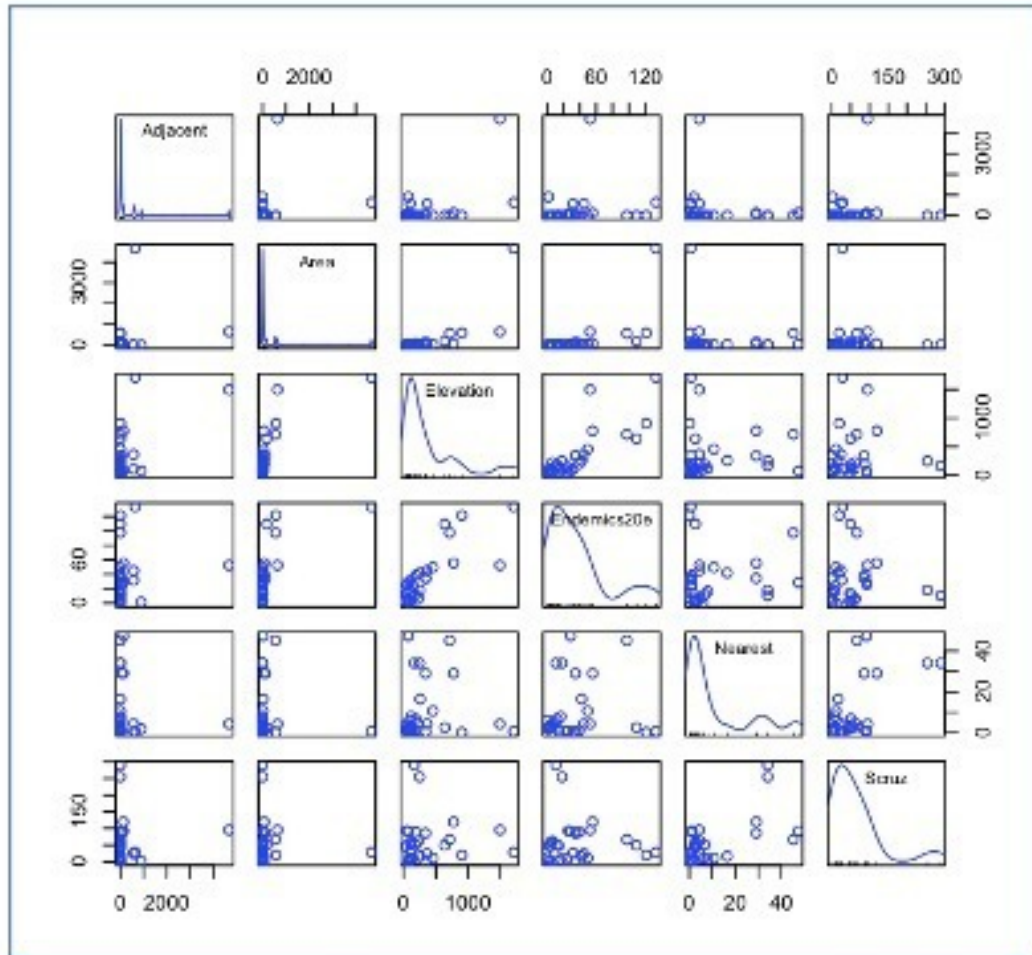


Figure 1b. Scatterplot matrix of the correlation plots of the independent variables. All the independent variable scatterplots are right skewed.

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 examining and transforming are Scruz, Area and Nearest.

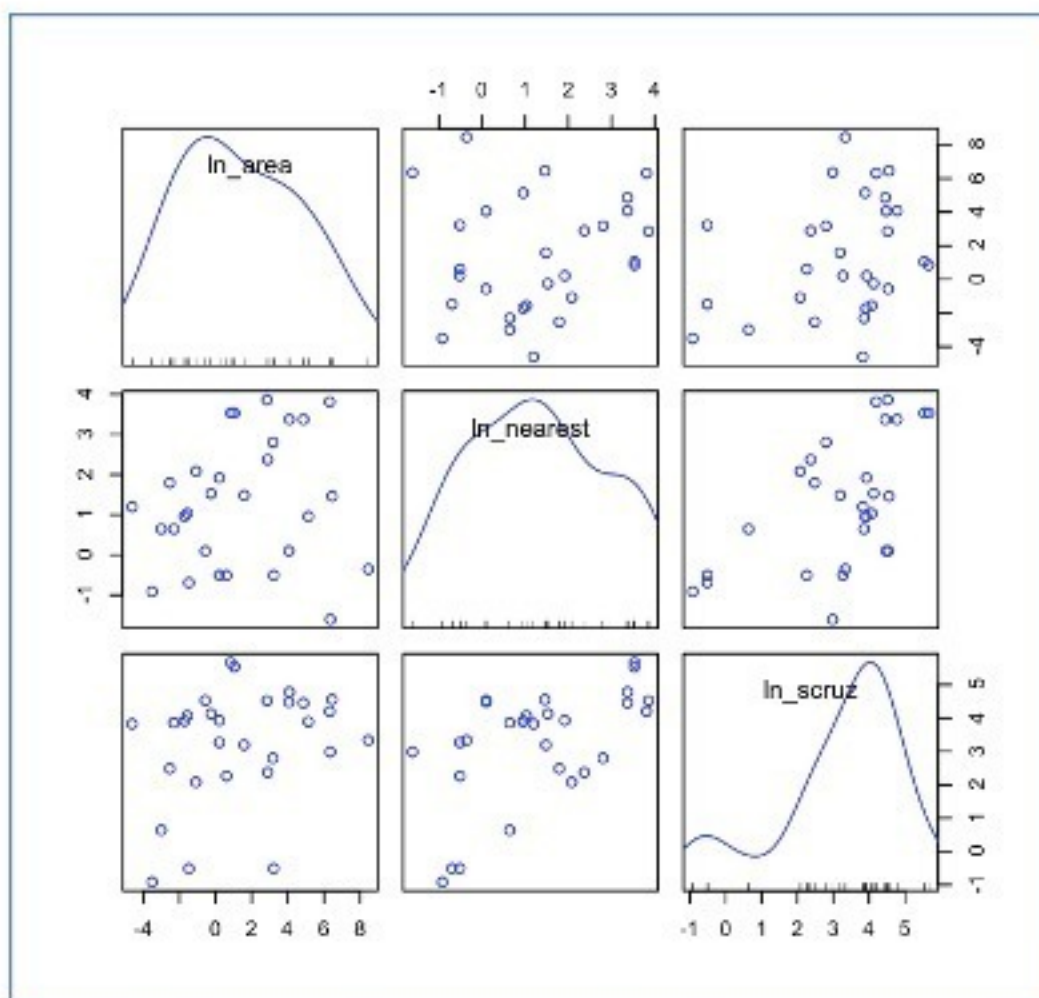


Figure 1c. The scatterplot matrix of the transformed variables. After log transformation of the variables area, nearest and Scrz, the scatterplots now exhibit normality.


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

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

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: 0.00002069
```

Figure 1d. The global model of $\ln(\text{area})$, $\ln(\text{nearest})$ and $\ln(\text{scruz})$ 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 = golapogous)

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

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

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

```
Rcmdr> summary(LinearModel.4)
```

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

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	0.00000942 ***
ln_nearest	-18.0778	14.1961	-1.273	0.214578
ln_area:ln_nearest	-0.3851	3.4812	-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: 0.00002869

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 = galapagos1)
```

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: 0.000003403

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

Figure 1a and 1b show that the population distribution is nonnormal 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 $p=2.48 \times 10^{-7}$. 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 than $\alpha=0.05$. 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 6.94×10^{-8} , which is less than the α of 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"

14.

Q14.2

13

```
glm(formula = Maturation ~ FLENmm + Lake + SEX, family = binomial(logit),
    data = wefishmaturity)

Deviance Residuals:
    1Q  Median      3Q      Max
-4.27  -0.4607  -0.1639   0.2073   4.3129

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -13.3156942    0.8156587  -16.325  <2e-16 ***
FLENmm         0.0297740    0.0005691   52.315  <2e-16 ***
Lake[T.Nipigon, L.]  0.9133472    0.7793182    1.172    0.241
Lake[T.of the Woods, L.]  1.0029213    0.7786987    1.288    0.198
Lake[T.Rainy L.]  0.8550344    0.7809833    1.095    0.274
SEX[T.male]     1.7021906    0.0673076   25.290  <2e-16 ***
Year          -0.0495413    0.0139738   -3.545  0.000392 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15175  on 12365  degrees of freedom
Residual deviance: 7327  on 12360  degrees of freedom
AIC: 7339

Number of Fisher Scoring iterations: 6

Rcmdr> exp(coef(GLM.5)) # Exponentiated coefficients ("odds ratios")
            (Intercept)          FLENmm      Lake[T.Nipigon, L.]
            0.000001648419          1.030221681375      2.492652068148
Lake[T.of the Woods, L.]      Lake[T.Rainy L.]          SEX[T.male]
            2.726234396167          2.351455204013          5.485951558611
```

Figure 2a. Generalized linear regression model of WEfishmaturity.

```
Rcmdr> stepwise(GLM.6, direction='backward/forward', criterion='AIC')

Direction: backward/forward
Criterion: AIC

Start: AIC=7328.38
Maturation ~ FLENmm + Lake + SEX + Year

      Df Deviance      AIC
<none>      7314.4  7328.4
- Lake      3   7321.3  7329.3
- Year      1   7327.0  7339.0
- SEX       1   8064.9  8076.9
- FLENmm    1  14694.1 14706.1

Call: glm(formula = Maturation ~ FLENmm + Lake + SEX + Year, family = binomial(logit),
  data = wefishmaturity)

Coefficients:
      (Intercept)              FLENmm      Lake[T.Nipigon, L.]
      85.69784              0.02994              1.02360
Lake[T.of the Woods, L.]      Lake[T.Rainy L.]      SEX[T.male]
      1.15032              1.00284              1.70630
      Year
     -0.04954

Degrees of Freedom: 12365 Total (i.e. Null); 12359 Residual
Null Deviance:      15180
Residual Deviance: 7314      AIC: 7328
```

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

```
> ll.null <- GLM.3$null.deviance/-2
> GLM.3$df.null
[1] 12365
```

Figure 2c. The null deviance of the model.

```
> ll.proposed <- GLM.3$null.deviance/-2
> GLM.3$df.residual
[1] 12359
```

Figure 2c. The residual deviance of the model.


```
> ll.null <- GLM.3$null.deviance/-2  
> GLM.3$df.null  
[1] 12365  
> ll.proposed <- GLM.3$deviance/-2  
> GLM.3$df.residual  
[1] 12359  
> (ll.null-ll.proposed)/ll.null  
[1] 0.518011
```

Figure 2d. McFadden's Pseudo R^2 is 0.518011.

```
> 1 - pchisq(2*(ll.proposed - ll.null), df=length(GLM.3$coefficients)-1)  
[1] 0
```

Figure 2e. The p-value associated with McFadden's Pseudo R^2 .

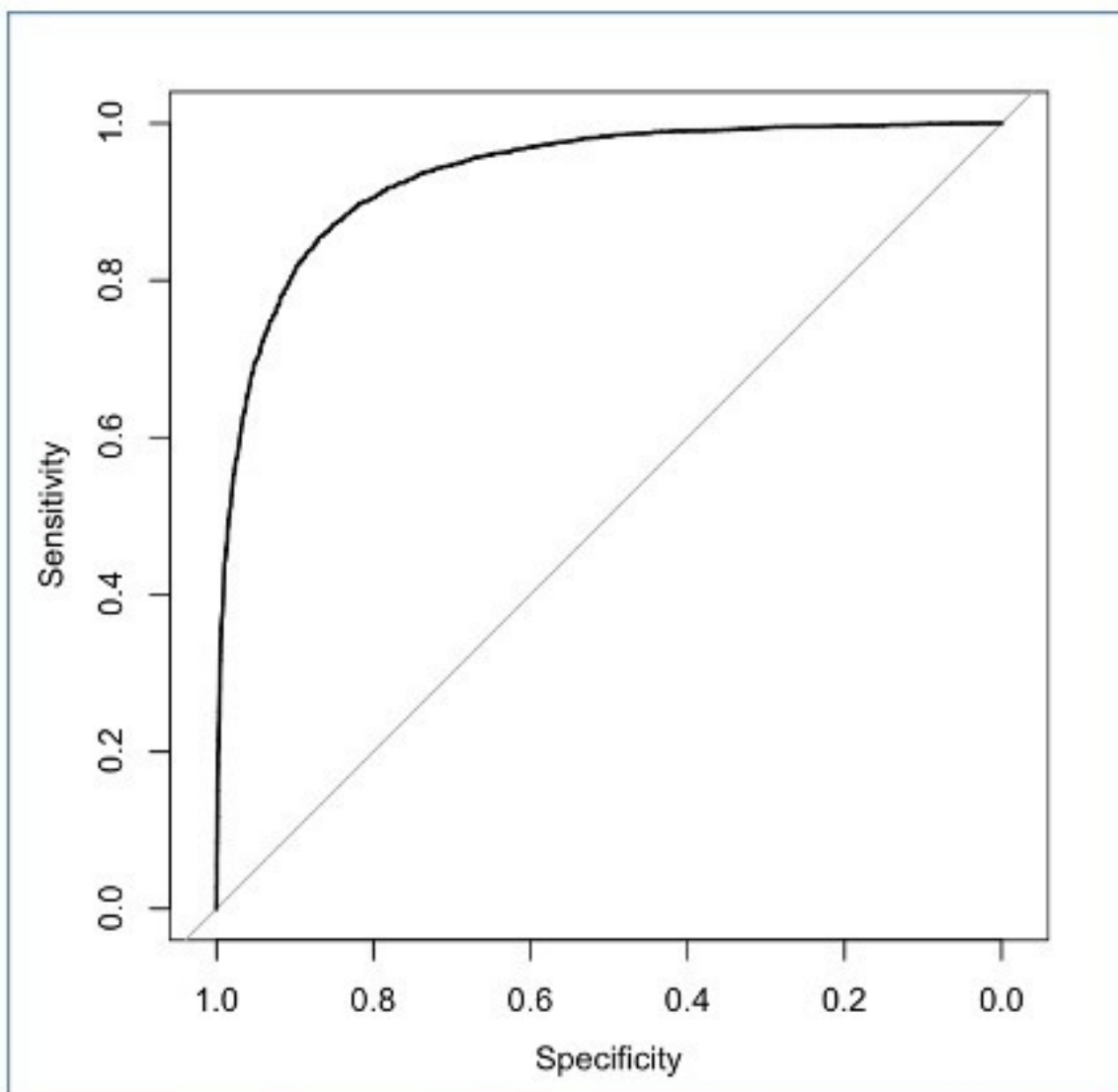


Figure 2f. The ROC curve of Walleye Fish Maturity.

```
Call:
roc.default(response = wefishmaturity$Maturation, predictor = GLM.35fitted.values, plot = TRUE)

Data: GLM.35fitted.values in 8616 controls (wefishmaturity$Maturation 0) < 3759 cases (wefishmaturity$Maturation 1).
Area under the curve: 0.9549
```

Figure 2g. The area under the ROC curve.

14. Figure 2a shows the global linear regression model of the Walleye Fish Maturity data. It shows that only the year, fork length and the sex are significant because they have p-values of 0.000392, $<2e-16$ and $<2e-16$, which are all a lot less than $\alpha=0.05$. After running a stepwise backward/forward selection, it was found that all variables (fork length, lake, sex and the year) are in fact relevant. From this, we can say that only fork length and sex can tell the PROBABILITY of catching a mature Walleye fish, while the other two variables, lake and year, can only tell the LIKELIHOOD of catching a mature Walleye fish. We can say that for every one unit increase in the fork length, the probability of catching a mature fish will increase by about 0.0298, controlling for the lake, sex and year; and for every one unit increase in the male sex population, the likelihood of catching a mature fish is 1.70, controlling for fork length, lake and year. On the other hand, the odd ratio of sex is 5.48, which is greater than 1. This means that there is an increase likelihood of catching a mature Walleye if a male one is caught. The null deviance of the model is 12365 and the residual deviance of the model is 12359. From these 2 values, the McFadden's Pseudo R^2 is calculated to be 0.518011. This is a relatively good number as a fit value. It shows that about 50% of the model reflect the ability of catching a mature Walleye based on all of the variables. The p-value of McFadden's Pseudo R^2 is exactly 0, which means that it is statistically significant ($\alpha=0.05$). Figure 2f shows the ROC curve of Walleye fish maturity. The graph is above the gray line, which means that we have a reasonable model. The area under the curve (AUC) is also identified to be 0.9349, which is a value that is very close to 1. We can conclude that our model is almost perfect, meaning it exhibits a prediction model that is less likely to have false negatives and false positives.

Q15

0.5

I feel like cat #7 right now :/