STATS 330 - ASSIGNMENT 1

jras611

09/03/2021

# STATS330 - Assignment 1

## Question 1a

We will first examine the relationship between success and treatment using tables of counts and odds.

### i) Creata a table of counts of success and treatmnt

#Question 1a.i  
#Creating Contingency Table  
stones.tbl = with(stones.df, table(success, treatment))  
stones.tbl

## treatment  
## success A B  
## 0 77 61  
## 1 273 289

### ii) Calculate the odds of successful removal of kidney stones for study participants receiving treatment A.

Probability is the fraction of times you expect to see a particular event in however many n total trials. So the probability of a succesful kidney stone removal for study participants receiving treatment A is the succesful kidney stone removals observed under treatment A divided by the total observations under treatment A.

The odds however are an expression of relative probabilities and defined as the probability that the event will occur divided by the probability that the event will not occur. If the probability of an event occuring is p and the probability of the same event not occurring is 1-p, then the odds of an event is p divided by 1-p (p/(1-p)).

#Probability of a succesful kidney stone removal under treatment A  
success\_A = 273/(273 + 77)   
  
#Odds of a succesful kidney stone removal under treatment A  
success\_A\_odds = success\_A / (1-success\_A)  
success\_A\_odds

## [1] 3.545455

Thus, there is a 78% probability of a succesful kidney stone removal for study participants receiving treatment A while the odds of a succesful kidney stone removal for study participants receiving treatment A is 3.5 successes to 1 failure.

### iii)Calculate the odds of successful removal of kidney stones for study participants receiving treatment B.

The same formula for probability and odds would apply for study participants under treatment B.

#Probability of a succesful kidney stone removal under treatment B  
success\_B = 289/(61 + 289)   
  
#Odds of a succesful kidney stone removal under treatment B  
success\_B\_odds = success\_B / (1-success\_B)  
success\_B\_odds

## [1] 4.737705

Therefore, for study participants receiving treatment B the odds of success is 4.7 successes to 1 failures.

### iv)Calculate the ratio of the odds calculated in parts (ii) and (iii) and interpret inwords. Which treatment is more successful for the removal of kidney stones?

The odds ratio of a 2-by-2 contingency table is the ratio of row-1 odds over row-2 odds. In plain words, the estimate of the odds ratio is the product of the diagonal values in the contingency table, divided by the product of the off-diaonal values

stones.tbl = with(stones.df, table(success, treatment))  
stones.tbl

## treatment  
## success A B  
## 0 77 61  
## 1 273 289

#Calculating Odds Ratio  
odds\_ratio = (77 \* 289)/(273 \* 61)  
odds\_ratio

## [1] 1.336276

The odds of a successful kidney stone removal through treatment B is about 1.3 times higher than the odds for a successful kidney stone removal under treatment A. Therefore, the more succesful treatment for the removal of kidney stones is treatment B.

## Question 1b

### i) identify the response variable and the explanatory variable.

The response variable is success and the explanatory variable is treatment.

### ii) Justify and fit an appropriate model for this exploration.

I’ve chosen a Logistic regression model for this exploration as we are handling data with a Binomial distribution since we are assessing the probability of a Successful or Unsuccessful outcome in the operation that is repeated multiple times.

Furthermore, the Binomial distribution better reflects the nature of a discrete distribution, acknowledges that the number of successes associated with the ith observation must be an integer between 0 and and accounts for nonconstant variance.

~

#Fitting Logistic Regression Model  
stones.fit = glm(success~treatment, data = stones.df, family = 'binomial')  
   
summary(stones.fit)

##   
## Call:  
## glm(formula = success ~ treatment, family = "binomial", data = stones.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8693 0.6189 0.6189 0.7049 0.7049   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.2657 0.1290 9.809 <2e-16 \*\*\*  
## treatmentB 0.2899 0.1911 1.517 0.129   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 694.98 on 699 degrees of freedom  
## Residual deviance: 692.67 on 698 degrees of freedom  
## AIC: 696.67  
##   
## Number of Fisher Scoring iterations: 4

exp(coef(stones.fit))

## (Intercept) treatmentB   
## 3.545455 1.336276

### iii) Interpret the model, being sure to communicate the uncertainty in your interpretation

We can make the following statements: - We estimate that the odds of succesful kidney stone removal for a study participant under treatment A is 3.5 - We estimate that for every study participant, the odds of a succesful kidney removal under treatment B is multiplied by 1.3

exp(confint(stones.fit))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 2.768695 4.594822  
## treatmentB 0.920001 1.947868

100\*(exp(confint(stones.fit))-1)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 176.869523 359.48220  
## treatmentB -7.999899 94.78678

-We estimate that for every study participant, the odds of succesful kidney removal under treatment B is multiplied by between 0.92 and 1.9

* We estimate that for every study participant, the odds of a succesful kidney stone removal under treatment B is multiplied by between -8% and 94%.

### iv) How do the model results compare to your answer in (a), part(iv) above?

The model results align with the answer to 1a(iv). As we found in question 1a(iv) the odds of a succesful kidney stone removal under treatment A is 3.5 and likewise, these odds are multiplied by 1.3 under treatment B.

## Question 1C

### i) Create two tables of counts of success and treatment, one for small kidney stones and one for large kidney stones.

#Creating Tables for Large and Small succesful kidney stone removals  
large.df = subset(stones.df, size == 'large')  
small.df = subset(stones.df, size == 'small')  
  
large.tbl = with(large.df, table(success, treatment))  
small.tbl = with(small.df, table(success, treatment))  
  
#Large Kidney Stones Table  
large.tbl

## treatment  
## success A B  
## 0 71 25  
## 1 192 55

#Small Kidney Stones Table  
small.tbl

## treatment  
## success A B  
## 0 6 36  
## 1 81 234

### ii) Calculate the odds of successful removal of small kidney stones for study participants receiving treatment A.

#Probability of Success for Small Kidney Stones under Treatment A  
small\_pA = 81/87  
small\_pA

## [1] 0.9310345

#Odds of Success for Small Kidney Stones under Treatment A  
small\_oddsA = small\_pA/(1-small\_pA)  
small\_oddsA

## [1] 13.5

The probability of successful removal of small kidney stones for study participants receiving treatment A is 0.93.

As such, the odds of successful removal of small kidney stones for study participants receiving treatment A is 13.28571

### iii) Calculate the odds of successful removal of small kidney stones for study participants receiving treatment B.

#Probability of Success for Small Kidney Stones under Treatment B  
small\_pB = 234/270  
small\_pB

## [1] 0.8666667

#Odds of Success for Small Kidney Stones under Treatment B  
small\_oddsB = small\_pB/(1-small\_pB)  
small\_oddsB

## [1] 6.5

The odds of successful removal of small kidney stones for study participants receiving treatment B is 6.5.

### iv) Calculate the ratio of the odds calculated in parts (ii) and (iii) and interpret in words. Which treatment is more successful for the removal of small kidney stones?

#Small Kidney Stone Table  
small.tbl

## treatment  
## success A B  
## 0 6 36  
## 1 81 234

small\_ratio = (6 \* 234)/(81 \* 36)  
small\_ratio

## [1] 0.4814815

The ratio of odds for the succesful removal of small kidney stones under treatment B is 0.48 higher than the odds for succesful removal of small kidney stones under treatment A. The more succesful treatment for the removal of small kidney stones is treatment A.

### v) Calculate the odds of successful removal of large kidney stones for study participants receiving treatment A.

#Calculating the probability of a succesful removal of a large kidney stone under treatment A  
large\_pA = 192/263  
  
#Calculating odds  
large\_oddsA = large\_pA/(1-large\_pA)  
large\_oddsA

## [1] 2.704225

The odds of a successful removal of large kidney stones for study participants receiving treatment A is 2.704225.

### vi) Calculate the odds of successful removal of large kidney stones for study participants receiving treatment B.

#Calculating the probability of a succesful removal of a large kidney stone under treatment B  
large\_pB = 55/80  
  
#Calculating odds  
large\_oddsB = large\_pB/(1-large\_pB)  
large\_oddsB

## [1] 2.2

The odds of successful removal of large kidney stones for study participants receiving treatment B is 2.2

### vii) Calculate the ratio of the odds calculated in parts (v) and (vi) and interpret in words. Which treatment is more successful for the removal of large kidney stones?

#Large Kidney Stone Removal  
large.tbl

## treatment  
## success A B  
## 0 71 25  
## 1 192 55

large\_ratio = (71 \* 55)/(192 \* 25)  
large\_ratio

## [1] 0.8135417

The ratio of odds for the succesful removal of large kidney stones under treatment B is 0.81 higher than the odds for succesful removal of large kidney stones under treatment A. The more succesful treatment for the removal of large kidney stones is treatment A.

## Question 1d

### Comment on what you discovered in (a) part (iv) and (c) parts (iv) and (vii). Are you surprised? If so, why? If not, why not?

In (a) part(iv) we discovered that the odds of a successful kidney stone removal through treatment B is about 1.3 times higher than the odds for a successful kidney removal under treatment A.

We extrapolate on that in (c) parts (v) and (vi) and learn that the ratio of odds for the succesful removal of small kidney stones under treatment B is 0.48 higher than the odds for succesful removal of small kidney stones under treatment A and also that the ratio of odds for the succesful removal of large kidney stones under treatment B is 0.81 higher than the odds for succesful removal of large kidney stones under treatment A, and that in those specified cases of succesfully removing a small or large kidney stone; treatment A was actually the preferrable option.

This is surprising to see considering that we discovered early on in Q1a (iv) and was supported through the code Q1b(iv) that the odds of success under treatment B was 1.3 times higher than treatment A.

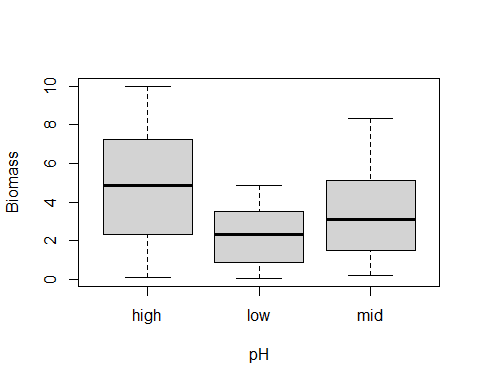
I believe this discrepancy could be due to the fact that treatment B is most typically used for small kidney stone removals with 270 out of 350 the study participants observed under treatment B (77%) undergoing a small kidney stone removal, as opposed to the majority of treatment A study participants (263 out of 350 (75%)) who had to have a large kidney stone removal.

I imagine there is a significant difference in not only the procedure, but also the risk of failure or success among the two sizes of kidney stones during removal such that even if treatment A would be the more reliable option to remove either; small kidney stones are more likely to be easily and thus succesfully removed such that they are typically handled with treatment B since there is less risk for complication with smaller kidney stones. In that case, it makes sense why despite the findings of Q1c (v) and (vi), the data for treatment B may make it seem like the best treatment possible to ensure a succesful removal of a kidney stone.

## Question 2a

### Create appropriate plots to explore the following relationships, and comment on what the plots tell you:

#Loading Species dataframe  
species.df = read.csv('species.csv', header = TRUE)  
names(species.df)[names(species.df) == 'ï..pH'] = 'pH'  
  
#Subsetting each variable  
Biomass = species.df$Biomass  
pH = species.df$pH  
Species = species.df$Species  
  
#Biomass and pH  
boxplot(Biomass ~ pH)

 As Biomass is a numeric variable and pH is a categorical variable, it is better to explore the relationship of the two variables (Biomass and pH) through a boxplot.

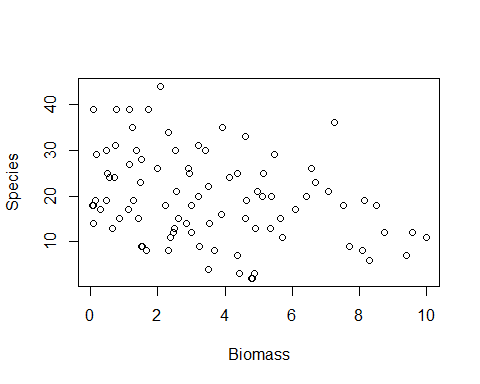
We can notice that the first boxplot for the group with a high pH level is comparatively tall, suggesting some significant difference across the other pH levels (low and mid) as well as significant variance in Biomass across the observations of the same high pH level.

Whereas the boxplot of those with low pH levels is comparatively short suggesting less variance across the Biomass values across observations of the same low pH level.

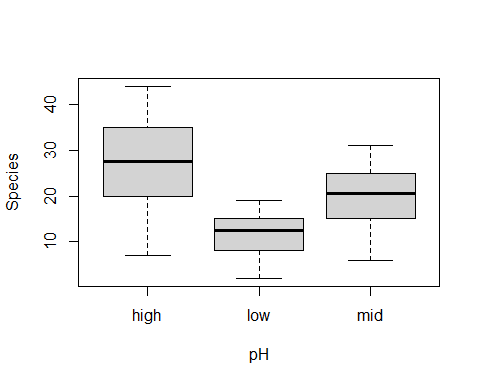
Both the boxplots for high and low pH levels are relatively symmetrical.The mid pH level group appears to have a larger upper quartile than a lower quartile, indicating more variance in observations with greater Biomass values.

The minimum of each pH level group is approximately similar, whereas the maximum is a lot more varied. In fact, the median of the high pH level group is approximately the same as the maximum of the low pH group and the upper quartile of the mid pH group.

#Biomass and Species  
plot(Biomass, Species)

 Biomass and Species are both numeric variables and thus can be plotted via a scatterplot. The following plot generally shows a downward trend between Species and Biomass thus indicating a negative relationship such that as Biomass increases, Species decreases.

#Species and pH  
boxplot(Species ~ pH)

 Species is a numeric variable while pH is a categorical variable, therefore a boxplot would be most appropriate to explore their relationship. Immediately, we notice a resemblance to the Biomass and pH plot,.The group with high pH tends to have a greater number in species than those of low or mid pH levels. It appears that there is a great difference in maximums between species count of high pH level observations to low and mid pH observations, that the relationship is relatively symmetrical as well and that the median of the high pH level boxplot is the same as the upper quartile of the mid pH level boxplot.

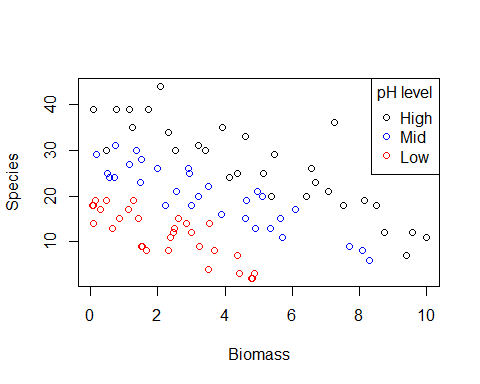
The similar minimums also bears resemblance, except for the low pH level plot which has a much lower minimum and a shorter plot with perhaps a larger lower quartile than upper quartile indicating more variance in the observations of low pH level with small species count.

Differences also lie in how the mid pH level box plot is more symmetrical for species.

## Question 2b

### Create a plot that explores the relationship between Biomass, Species and pH. Comment on what the plot tells you.

#Colouring plot by pH level  
group.cols <- c(high = "black", mid = "blue", low = "red")  
point.cols <- group.cols[as.character(pH)]  
#Plotting  
plot(Biomass, Species, col = point.cols)  
#Adding Legend  
legend("topright", legend = c("High", "Mid", "Low"), pch = rep(1, 3), col = c("black", "blue", "red"), title = 'pH level')

 From the plot we can notice that Biomass and Species has a generally negative relationship where as Biomass increases, Species decreases, but in addition to the classification by pH level we can derive the insight that observations with High pH levels tend to also have a greater amount of Species and a larger value for Biomass and that the amount of Species or Biomass tends to decrease as pH level decreases.

## Question 2c

### i) Identify the response and explanatory variables for this exploration.

As we are interested in exploring the relationship between the condition of the soil and the number of plant species, the response variable would be the number of plant species and the explanatory variables would be the condition of the soil measured by Biomass \* pH.

### ii) What model would be appropriate to fit? Choose from Linear regression, Poisson regression and Logistic regression. Justify your choice

As the response variable consists of discrete count data, the appropriate model to fit is a Poisson regression. A Poisson regression model assumes a discrete distribution for the given discrete response, accounts for nonconstant variance which we suspected in our investigation of the plots and relationships in 2a and 2b and ensures that the epected value is greater than 0 for all observations. Thus, Poisson regression would be the best choice for this model fit.

### iii) Using your answers to parts (i) and (ii), run the following code for two models making sure you replace the blanks with the appropriate terms. Briefly interpret the

.

Model 1

model1 = glm(Species~Biomass\*pH, family = "poisson", data = species.df)  
  
summary(model1)

##   
## Call:  
## glm(formula = Species ~ Biomass \* pH, family = "poisson", data = species.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4978 -0.7485 -0.0402 0.5575 3.2297   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.76812 0.06153 61.240 < 2e-16 \*\*\*  
## Biomass -0.10713 0.01249 -8.577 < 2e-16 \*\*\*  
## pHlow -0.81557 0.10284 -7.931 2.18e-15 \*\*\*  
## pHmid -0.33146 0.09217 -3.596 0.000323 \*\*\*  
## Biomass:pHlow -0.15503 0.04003 -3.873 0.000108 \*\*\*  
## Biomass:pHmid -0.03189 0.02308 -1.382 0.166954   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 452.346 on 89 degrees of freedom  
## Residual deviance: 83.201 on 84 degrees of freedom  
## AIC: 514.39  
##   
## Number of Fisher Scoring iterations: 4

In other words, we have

where is the biomass of the th soil, and and are dummy variables for the low- and mid-pH levels, respectively.

There are two kinds of effects: The main effects of Biomass and the respective pH levels which only involve a single variable and the interaction effects such as the interaction between Biomass and pH level which involves more than one variable.

We can interpret the main effects in the same way as we can for models that do not have an interaction as the baseline level is another variable which is a factor.

#Interpreting Main effect of Biomass  
100\*(exp(10\*confint(model1)[2, ])-1)

## Waiting for profiling to be done...

## 2.5 % 97.5 %   
## -73.21439 -56.28961

We estimate that, for every 10m/a increase in the biomass of soil, the expected number of plants in species is associated with a decrease of between 73.21439% and 56.28961%.

#Interpreting Low pH levels  
100\*(exp(confint(model1)[3, ])-1)

## Waiting for profiling to be done...

## 2.5 % 97.5 %   
## -63.89244 -45.95685

We estimate that low levels of pH is associated with a decrease in the expected number of plant species of between 63.89244% and 45.95685%.

#Interpreting Mid pH levels  
100\*(exp(confint(model1)[4, ])-1)

## Waiting for profiling to be done...

## 2.5 % 97.5 %   
## -40.09656 -14.01968

We estimate that mid levels of pH is associated with a decrease in the expected number of plant species of between 40.09656% and 14.01968%.

100\*(exp(confint(model1)[5, ])-1)

## Waiting for profiling to be done...

## 2.5 % 97.5 %   
## -20.889224 -7.440173

100\*(exp(confint(model1)[6, ])-1)

## Waiting for profiling to be done...

## 2.5 % 97.5 %   
## -7.444508 1.320015

An interaction effect, measures how the effect of one variable changes depending on the other. and measures the differences in the intercept and slope respectively between the mid level pH and high level pH and the low level pH and high level pH respectively. From the calculation above we can estimate that the effect of mid pH level decreases by between approximately 21% and 7%, whilst the effect of low pH level changes from between a 7% decrease to a 1% increase.

Model 2

model2 = glm(Species~Biomass+pH, family = "poisson", data = species.df)  
summary(model2)

##   
## Call:  
## glm(formula = Species ~ Biomass + pH, family = "poisson", data = species.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5959 -0.6989 -0.0737 0.6647 3.5604   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.84894 0.05281 72.885 < 2e-16 \*\*\*  
## Biomass -0.12756 0.01014 -12.579 < 2e-16 \*\*\*  
## pHlow -1.13639 0.06720 -16.910 < 2e-16 \*\*\*  
## pHmid -0.44516 0.05486 -8.114 4.88e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 452.346 on 89 degrees of freedom  
## Residual deviance: 99.242 on 86 degrees of freedom  
## AIC: 526.43  
##   
## Number of Fisher Scoring iterations: 4

In other words, we have

where is the biomass of the th soil, and and are dummy variables for the low- and mid-pH levels, respectively.

As opposed to Model 1. Model 2 fits an interaction effect which measures how the effect of one variable changes depending on the other. In this case, the high pH level group.

100\*(exp(confint(model2))-1)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 4128.41371 5100.97582  
## Biomass -13.71583 -10.21640  
## pHlow -71.88953 -63.41456  
## pHmid -42.47504 -28.67012

Holding the pH of the soil constant, we estimate that the number of plant species in a plant plot - in the mid pH level group is between 42$ and 28% lower than an observation in the high pH level group. - in the low pH level group is between 71% and 63% lower than an observation in the high pH level group

### iv) Evaluate the adequacy of the models fitted in part (iii). Comment briefly

From the plot, we can discern that the residual deviance of model 1 is 83.201 whereas the degrees of freedom is 84. Having the residual deviance be smaller than the degress of freedom is typically a good sign for adequacy.

In comparison, model 2 has a residual deviance of 99.242 on 86 degrees of freedom so we can infer that model 1 would likely be the more adequate model to fit.

### v) Calculate the change in residual deviance from model1 to model2 and calculate the probability of observing such a change given we have ‘lost’ 2 degrees of freedom (i.e., we have reduced the number of parameters in the model from 6 to 4).

We know from the summary() function performed for each model in Q2 iii) that the residual deviance for model 1 and model 2 is 83.20114 and 99.2415 respectively.

#model1 deviance  
print('Model 1 Deviance')

## [1] "Model 1 Deviance"

dev1 = deviance(model1)  
dev1

## [1] 83.20114

#model2 deviance  
print('Model 2 Deviance')

## [1] "Model 2 Deviance"

dev2 = deviance(model2)  
dev2

## [1] 99.2415

Since we know that model 1 deviance < model 2 deviance, we proceed to calculate the change in residual deviance from model 2 to model 1.

#residual deviance calculation  
dev2 - dev1

## [1] 16.04036

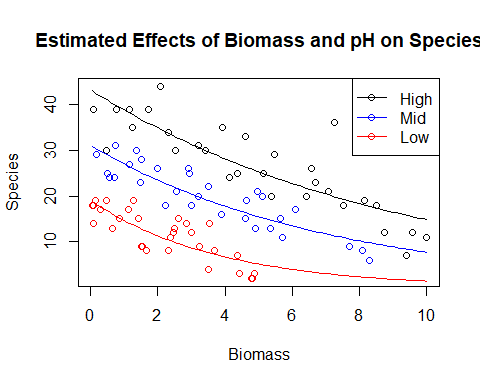
#or alternatively  
anova(model1, model2)

## Analysis of Deviance Table  
##   
## Model 1: Species ~ Biomass \* pH  
## Model 2: Species ~ Biomass + pH  
## Resid. Df Resid. Dev Df Deviance  
## 1 84 83.201   
## 2 86 99.242 -2 -16.04

The change in residual deviance is 16.04036.

### vi) Construct a plot displaying the estimated effects of your preferred model (i.e. model1 or model2) on the response . Clearly communicate what your plot displays.

plot(Biomass, Species, col = point.cols, main = 'Estimated Effects of Biomass and pH on Species')  
## A vector of 90 Biomass values between 0.05 m/a and 10m/a.  
xx <- seq(0.05, 10, length.out = 90)  
## A data frame for the high pH level group.  
new.data.high <- data.frame(Biomass = xx, pH = rep("high", 90))  
## A data frame for the low pH level group.  
new.data.low <- data.frame(Biomass = xx, pH = rep("low", 90))  
## A data frame for the mid pH level group.  
new.data.mid<- data.frame(Biomass = xx, pH = rep("mid", 90))  
  
## Fitted values for each of the three groups.  
yy.mid <- predict(model1, new.data.mid, type = "response")  
yy.low <- predict(model1, new.data.low, type = "response")  
yy.high <- predict(model1, new.data.high, type = "response")  
## Adding lines to the plot.  
lines(xx, yy.high, col = "black")  
lines(xx, yy.low, col = "red")  
lines(xx, yy.mid, col = "blue")  
## Adding the legend.  
legend("topright", legend = c("High", "Mid", "Low"), pch = rep(1, 3), col = c("black", "blue", "red"), lty = rep(1, 3))

 The plot above shows the estimated effects of model 1 on the response variable Species. In particular, it estimates that as Biomass increases, Species will decrease and the separate lines across the pH levels distinguishes the estimated effect of Biomass on species by pH, with each respective pH generally following the same downward trend albeit at differring values of Biomass and Species corresponding to their pH level (High pH Level = Greater number of Biomass and Species vs. Low pH level = Smaller number of Biomass and species)

### vii) Based on your model, calculate and interpret 95% confidence intervals for the mean number of plant species observed in each of three separate plots, all of which have Biomass=5 but each with different levels of pH: low, mid, and high.

#Creating new data with given conditions  
newdata = data.frame(Biomass = 5, pH = c("high", "mid", "low"))  
#Predicting  
predict.glm(model1, newdata = newdata, type = "response", se.fit = TRUE)

## $fit  
## 1 2 3   
## 25.342293 15.511039 5.164264   
##   
## $se.fit  
## 1 2 3   
## 0.9443154 0.9069546 0.7151710   
##   
## $residual.scale  
## [1] 1

The average number of plant species observed in three separate plots of 5m/a in Biomass and differing pH levels (high, mid and low) is estimated to be: *25 (0dp) with a standard error of 0.9 for observations with high pH levels* 16 (0dp) with a standard error of 0.9 for observations with mid pH levels \* 5 with a standard error of 0.7 for observations with low pH levels.