STATS 201/8 Assignment 5

Anish Hota ahot228

Due Date: 3pm Thursday 10th October

1 Question 1 [15 Marks]

Counts of kelpfish (*Chironemus marmoratus*, or hiwihiwi) were made by diver surveys in Doubtless Bay, Northland, New Zealand. Dives were made at 55 randomly selected locations.

A researcher was interested in whether the count of hiwihiwi is influenced by depth. They wanted to know what the effect of the first 15 metres of depth was on hiwihiwi counts and also it is of interest to determine if there is an optimal depth at which the expected count of hiwihiwi is highest.

This data can be found in the file KelpFish.csv, and includes variables:

Variable	Description
Count	The number of hiwihiwi seen by SCUBA divers swimming a 25 metre transect over reef habitat.
Depth	The depth (in metres) of the dive.

Instructions:

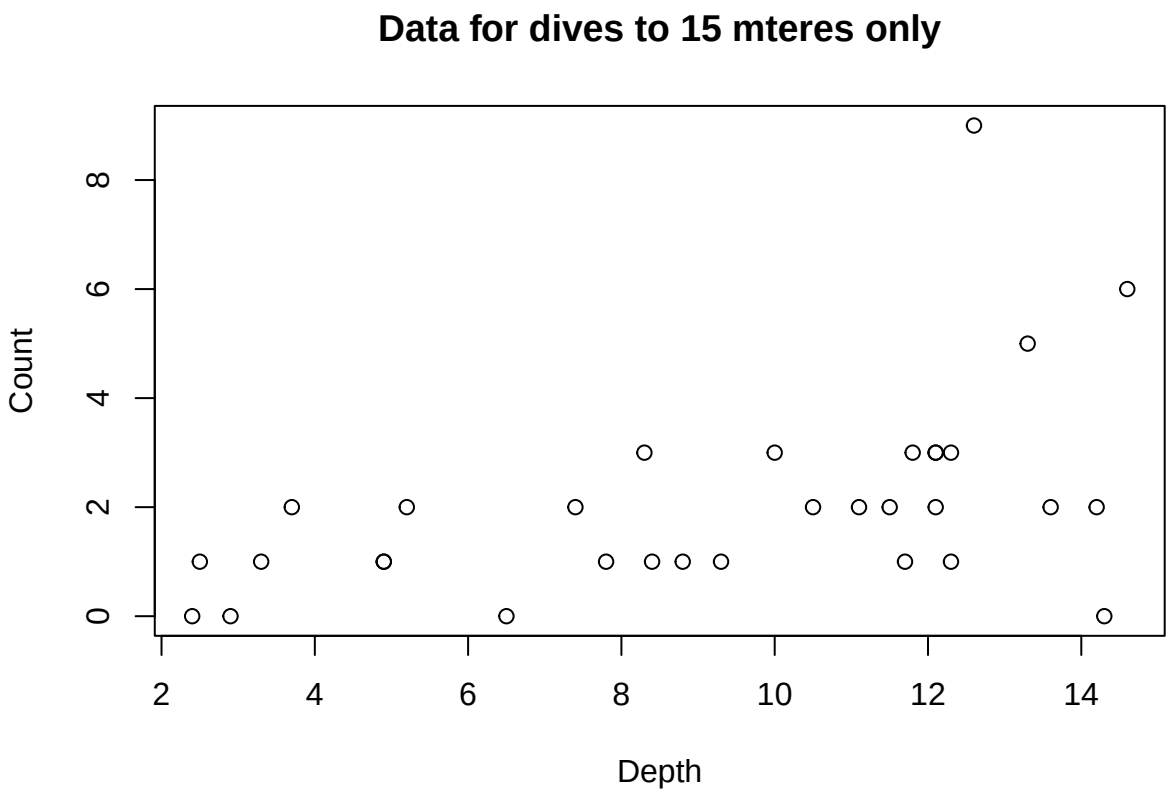
- Make sure you change your name and UPI/ID number at the top of the assignment.
- Comment on the plots.
- Fit an appropriate Poisson GLM modelling the number of hiwihiwi seen as counts for dives to a maximum of 15 metres. Generate confidence interval output for this model. Hint: As in the plot commands, replace the data statement with: `data=subset(Hiwi.df,Depth<=15)`
- Write a sentences as if for an *Executive Summary* estimating the effect of an additional 1 metre depth on the abundance of hiwihiwi (for the first 15 metres of depth).
- Using all the data, fit an appropriate Poisson GLM modelling the number of hiwihiwi seen as counts.
- Create a plot of all the data with the model superimposed over it.
- Generate prediction output to estimate the depth (to within the nearest 0.5 m) at which the expected count of hiwihiwi is highest. No confidence interval is required.
- State what the depth at which the expected count of hiwihiwi is estimated to be the highest.
- Write the fitted model as if for the *Methods and Assumption Checks*.

1.1 Questions of Interest:

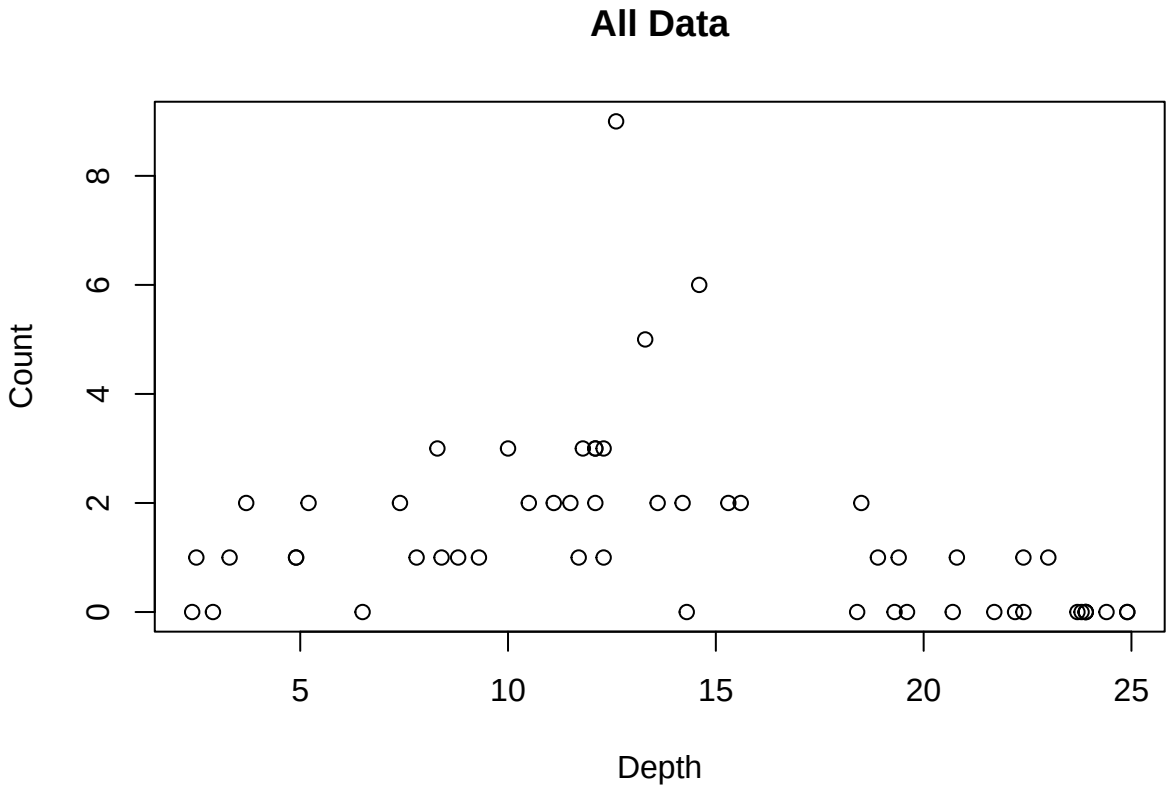
It was of interest to determine whether the count of hiwihiwi is influenced by depth. In particular, it is of interest to determine if there is an optimal depth at which the expected count of hiwihiwi is highest.

1.2 Read in and inspect the data:

```
Hiwi.df=read.csv("KelpFish.csv")
plot(Count~Depth,main="Data for dives to 15 mteres only", data=subset(Hiwi.df,Depth<=15) )
```




```
plot(Count~Depth,main="All Data",data=Hiwi.df)
```

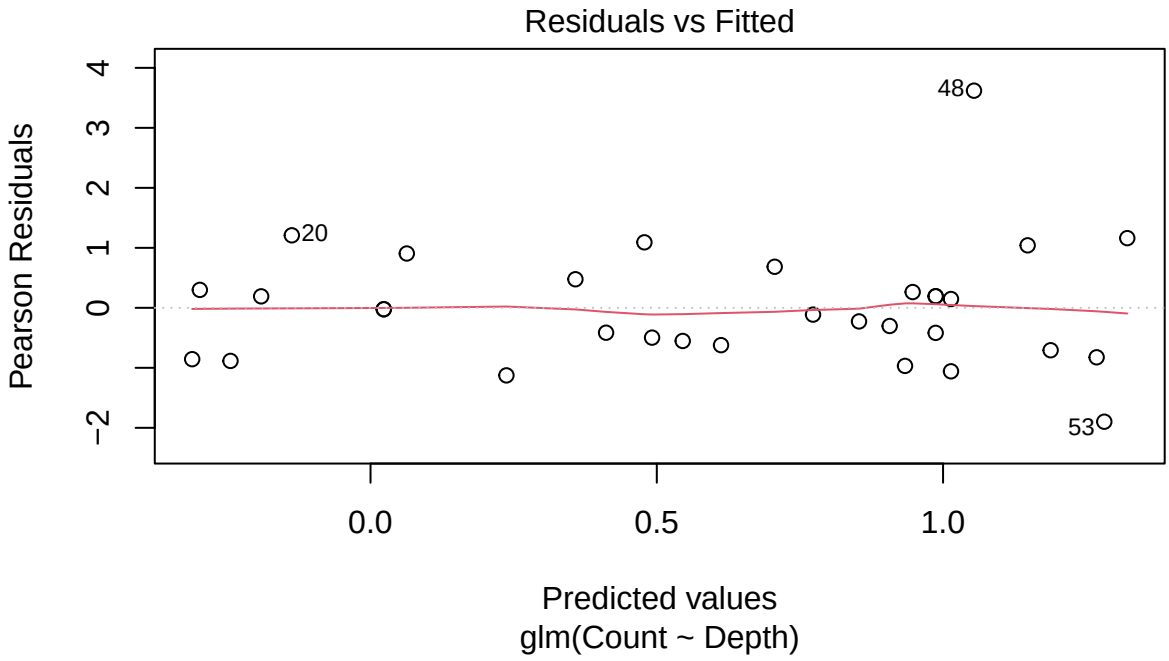


1.3 Comment on the plots:

There seems to be a general increase of count of fish to the increase in depth (a positive relationship), in the first 15 meters, more of a exponential curve. Looking at all the data the count seems to much higher around the the 13-15m level and lower on the other levels, creating a parabola sort of shape. Variability seems toe be generally good with slightly more data at the deeper levels.

1.4 Fit an appropriate Poisson GLM modelling the number of hiwihiwi seen as counts for dives to a maximum of 15 metres. Generate confidence interval output for this model. Hint: As in the plot commands, replace the data statement with: data=subset(Hiwi.df,Depth<=15)

```
Hiwi.gfit = glm(Count~Depth, family = poisson, data=subset(Hiwi.df,Depth<=15))
plot(Hiwi.gfit, which = 1)
```




```
summary(Hiwi.gfit)
```

```
##
## Call:
## glm(formula = Count ~ Depth, family = poisson, data = subset(Hiwi.df,
##     Depth <= 15))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.63285    0.44270  -1.430 0.152854
## Depth        0.13388    0.03899   3.434 0.000596 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 46.092  on 31  degrees of freedom
## Residual deviance: 32.452  on 30  degrees of freedom
```

```
## AIC: 109.12
##
## Number of Fisher Scoring iterations: 5
```

```
  confint(Hiwi.gfit)
```

```
## Waiting for profiling to be done...
```

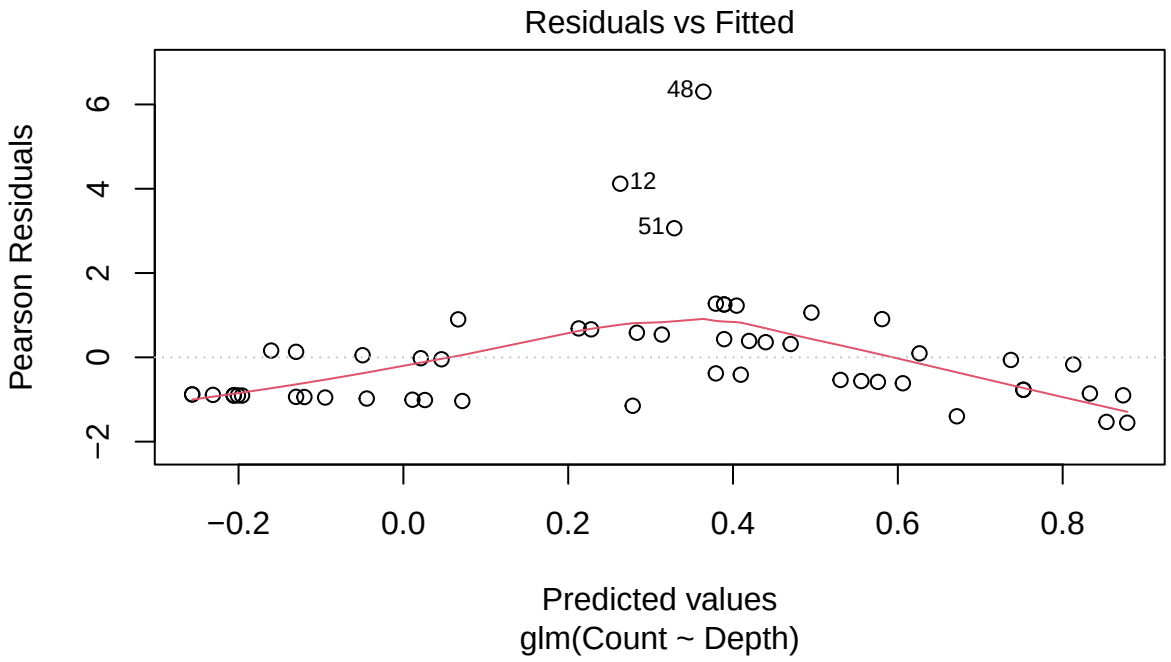
```
##                   2.5 %    97.5 %
## (Intercept) -1.56598579 0.1761627
## Depth        0.06057576 0.2140123
```

1.5 Write a sentences as if for an *Executive Summary* estimating the effect of an additional 1 metre depth on the abundance of hiwihiwi (for the first 15 metres of depth).

We estimate that with an additional 1 metre depth on the abundance of hiwihiwi (for the first 15 metres), the odds of an increase in fish count is between 6.06% and 21.40%.

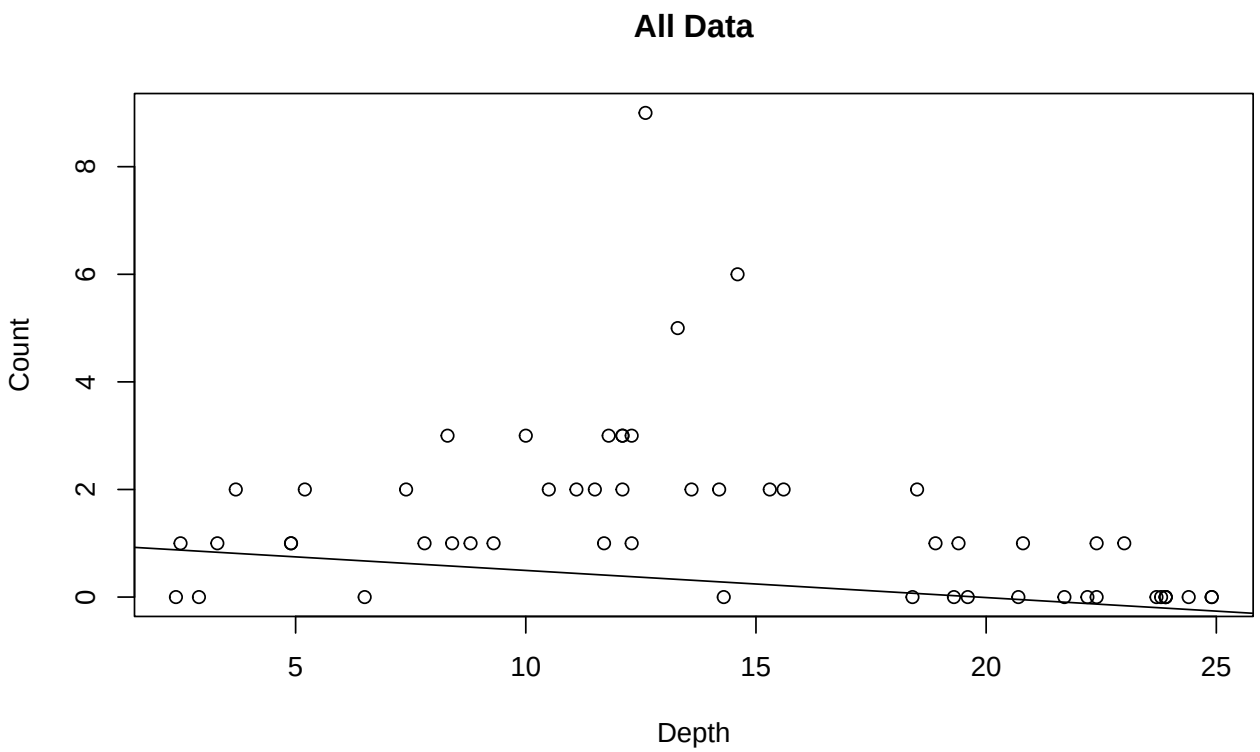
1.6 Using all the data, fit an appropriate Poisson GLM modelling the number of hiwihiwi seen as counts.

```
  Hiwi.gfit1 = glm(Count~Depth,family = poisson, data=Hiwi.df)
  plot(Hiwi.gfit1,which = 1)
```



1.7 Create a plot of all the data with the model superimposed over it.

```
plot(Count~Depth,main="All Data",data=Hiwi.df)
abline(Hiwi.gfit1)
```



1.8 Generate prediction output to estimate the depth (to within the nearest 0.5 m) at which the expected count of hiwihiwi is highest. No confidence interval is required.

```
depth_range <- seq(min(Hiwi.df$Depth), max(Hiwi.df$Depth))
predHiwi.df <- data.frame(Depth = depth_range)
predictcount <- predict(Hiwi.gfit1, newdata = predHiwi.df, type = "response")
maxdepth <- depth_range[which.max(predictcount)]
```

1.9 State what the depth at which the expected count of hiwihiwi is estimated to be the highest.

The depth at which the count is predicted to be the highest is 2.5 metres.

1.10 Write the fitted model as if for the *Methods and Assumption Checks.*

The response variable, is a count, so we fitted a generalised linear model with a Poisson response distribution. We have one explanatory variable Depth. The scatterplot of Depth vs Count showed an exponentially increasing trend for both locations. We fitted a Poisson model with interaction between depth and count.

---

2 Question 2 [19 Marks]

Sniffer dogs, also called detector dogs, are used in New Zealand airports to detect contraband items such as drugs or fruit that cannot be brought into the country undeclared. Researchers aimed to assess how close the detector dogs needed to be to contraband items to detect them successfully. They used two dogs, named Alfie and Bert, both of whom were beagles trained for contraband detection. Contraband items were concealed at distances ranging from 0 to 20 metres away from a walked route through an airport luggage collection

hall. Gaps between items were random, and the ordering of near and far items was randomized, to avoid the dogs picking up on patterns in concealed items. Alfie and Bert were walked along the route one day apart to ensure they were not influenced by each other's scents. The success or failure of each dog at detecting each item was recorded.

The data can be found in the file Dog.csv, with variables:

Variable	Description
detect	Whether the dog succeeded (1) or failed (0) to detect the item.
distance	Distance, in metres, of the concealed item from the dog's route.
dogName	Name of the dog (Alfie or Bert).

We want to:

- assess a dog's detection ability over different distances.
- assess whether there were differences in this relationship between the different dogs.
- more specifically, what was the detection probability at distances of 0, 5, 10, 15, and 20 metres?
- and, at what distance did the dogs detect only half of the contraband items?

Instructions:

- Comment on the two plots of the data.
- We have provided output for three models. Choose the most appropriate of these models for the data and generate confidence interval and prediction output required.
- Why did we **not** need to test for overdispersion on this data?
- Write the fitted model as if for the *Methods and Assumption Checks*.
- Was the any evidence that the detection ability differed between the two dogs or that how their detection ability changed at different distances differed? Justify your answer including relevant P-values.
- If we were to superimpose our model onto the plot of the data, would it look like (a) one single line, (b) two parallel lines, (c) two non-parallel lines, (d) one S-curve, (e) two parallel S-curves, or (f) two non-parallel S-curves. **Justify your answer.**
- Write a sentences as if for an *Executive Summary* estimating the effect of a 1 metre increase in distance on the dog's detection ability.
- What are the detection probabilities at 0, 5, 10, 15, and 20 metres? **Note:** Report point estimates only. You do not need to calculate confidence intervals.
- At what distance did the dogs detect only half of the contraband items?

2.1 Questions of Interest:

We wish to study the ability of sniffer dogs to detect contraband items from various distances away. We also wish to assess evidence for different abilities among the two dogs in the study, and to quantify the probability of detection for items at various distances.

2.2 Read in and inspect the data:

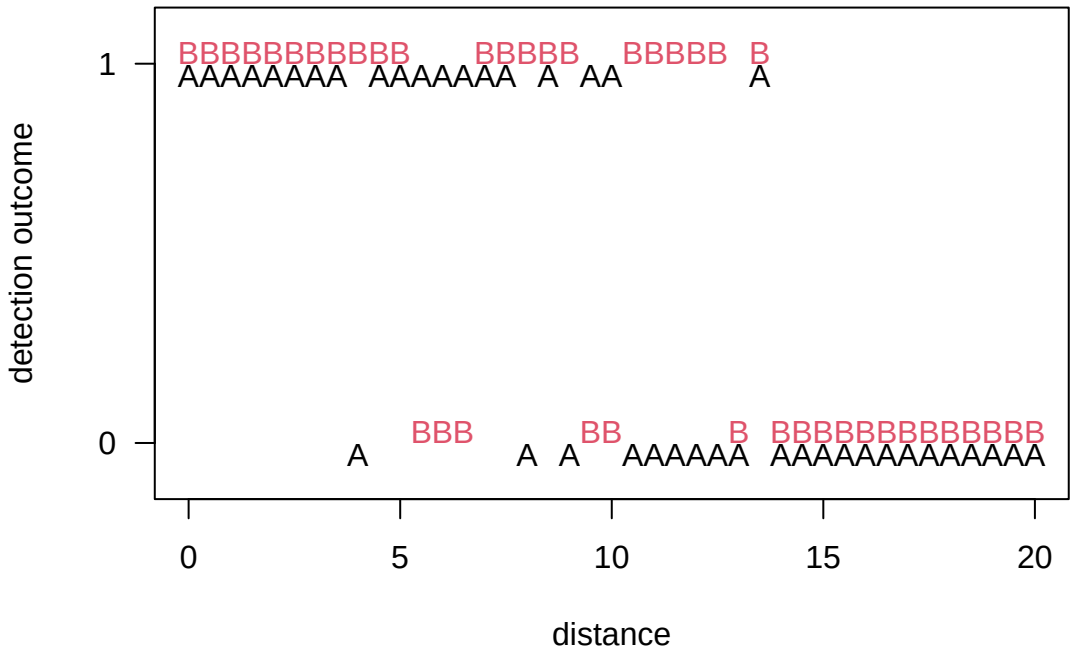
```
dog.df = read.csv("Dog.csv", stringsAsFactors=TRUE)
head(dog.df)
```

```
##   dogName distance detect
## 1   Alfie      0.0      1
## 2   Alfie      0.5      1
## 3   Alfie      1.0      1
## 4   Alfie      1.5      1
## 5   Alfie      2.0      1
## 6   Alfie      2.5      1
```

```
tail(dog.df)
```

```
##    dogName distance detect
## 77    Bert     17.5      0
## 78    Bert     18.0      0
## 79    Bert     18.5      0
## 80    Bert     19.0      0
## 81    Bert     19.5      0
## 82    Bert     20.0      0
```

```
# For clarity, plot Alfie's results slightly below 0 and 1, and Bert's results slightly above:
with(dog.df[dog.df$dogName=="Alfie",], plot(distance, detect-0.03, pch="A", ylim=c(-0.1, 1.1), yaxt="n"
with(dog.df[dog.df$dogName=="Bert",], points(distance, detect+0.03, pch="B", col=2))
axis(2, at=c(0, 1), las=1)
```



2.3 Comment on the plot:

The dogs in general tend to not detect food when the distance is further away. There seems to be not much difference between the two dogs when it comes to detection, although it seems that Bert can detect at a generally higher distance than Alfie.

2.4 Fit three models and generate summary output

```
dog.fit1 <- glm(detect ~ distance * dogName, family=binomial, data=dog.df)
summary(dog.fit1)
```

```
##
## Call:
## glm(formula = detect ~ distance * dogName, family = binomial,
##     data = dog.df)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)            4.7419     1.5147   3.131  0.00174 **
## distance              -0.5131     0.1536  -3.339  0.00084 ***
## dogNameBert           -0.9970     1.9035  -0.524  0.60045
## distance:dogNameBert   0.1662     0.1837   0.905  0.36567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113.676  on 81  degrees of freedom
## Residual deviance:  59.264  on 78  degrees of freedom
## AIC: 67.264
##
## Number of Fisher Scoring iterations: 6
```

```
dog.fit2 <- glm(detect ~ distance + dogName, family=binomial, data=dog.df)
summary(dog.fit2)
```

```
##
## Call:
## glm(formula = detect ~ distance + dogName, family = binomial,
##     data = dog.df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.80926    0.91322   4.171 3.03e-05 ***
## distance    -0.41281    0.08453  -4.883 1.04e-06 ***
## dogNameBert  0.63776    0.66231   0.963    0.336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113.676  on 81  degrees of freedom
## Residual deviance:  60.152  on 79  degrees of freedom
## AIC: 66.152
##
## Number of Fisher Scoring iterations: 5
```

```
dog.fit3 <- glm(detect ~ distance, family=binomial, data=dog.df)
summary(dog.fit3)
```

```
##
## Call:
## glm(formula = detect ~ distance, family = binomial, data = dog.df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.05731    0.88856   4.566 4.97e-06 ***
## distance    -0.40573    0.08276  -4.903 9.46e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 113.676  on 81  degrees of freedom
## Residual deviance:  61.101  on 80  degrees of freedom
## AIC: 65.101
##
## Number of Fisher Scoring iterations: 5
```

2.5 Generate inference and prediction output from cheosen model

```
summary(dog.fit1)
```

```
##
## Call:
## glm(formula = detect ~ distance * dogName, family = binomial,
##     data = dog.df)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            4.7419     1.5147   3.131  0.00174 **
## distance              -0.5131     0.1536  -3.339  0.00084 ***
## dogNameBert           -0.9970     1.9035  -0.524  0.60045
## distance:dogNameBert   0.1662     0.1837   0.905  0.36567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 113.676  on 81  degrees of freedom
## Residual deviance:  59.264  on 78  degrees of freedom
## AIC: 67.264
##
## Number of Fisher Scoring iterations: 6
```

```
confint(dog.fit2)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %     97.5 %
## (Intercept)  2.2189676  5.8555974
## distance    -0.6050765 -0.2679705
## dogNameBert -0.6402734  1.9977192
```

```
predictGLM(dog.fit3, data.frame(distance = c(0,5,10,15,20)),type="repsonse")
```

```
## ***Estimates and CIs are on the link scale***
```

```
##             fit        lwr        upr
## 1  4.057313e+00  2.3157594  5.7988675
## 2  2.028657e+00  0.9992162  3.0580972
## 3 -1.332268e-15 -0.6340296  0.6340296
## 4 -2.028657e+00 -3.0580972 -0.9992162
## 5 -4.057313e+00 -5.7988675 -2.3157594
```

2.6 Why did we not need to test for overdispersion on this data?

Our data is also a binomial model meaning that overdispersion check is not needed (It is not a Poisson model so there is most likely not really any overdispersion).

2.7 Write the fitted model as if for the *Methods and Assumption Checks.*

The data recorded was whether or not each dog would detect food at different distances. Therefore we fitted a Binomial GLM model with a single predictor of detection. The response was treated as grouped data, with each group responding to a distance.

2.8 Was the any evidence that the detection ability differed between the two dogs or that how their detection ability changed at different distances differed? Justify your answer including relevant P-values.

There is no evidence that there was a difference in detection ability between the two dogs, with a large p-value of 0.3657.

2.9 If we were to superimpose our model onto the plot of the data, would it look like (a) one single line, (b) two parallel lines, (c) two non-parallel lines, (d) one S-curve, (e) two parallel S-curves, or (f) two non-parallel S-curves. Justify your answer.

It would be two parallel S-Curves as they both have low detection rates at higher distances and form the previous question the difference between the two dogs is not significant, so it would start high anthen drop low for both dogs as the distance gets higher.

2.10 Write a sentences as if for an *Executive Summary* estimating the effect of a 1 metre increase in distance on the dog's detection ability.

The effect of a 1 metre increase in distance on a dog's detection ability is a probability of between 64% and 200% that the dog will not detect the food.

2.11 What are the detection probabilities at 0, 5, 10, 15, and 20 metres? Note: Report point estimates only. You do not need to calculate confidence intervals.

0 metres = 4.057 - 1 5 metres = 2.029 - 1 10 metres = -1.332e-15 - 0 15 metres = -2.029 - 0 20 metres = -4.057 - 0

2.12 At what distance did the dogs detect only half of the contraband items?

Looking at the data the the distance in which the dogs detected only half of the contraband was at 4 metres.