BB839 Exam 2019 - model answer zzs

07 February, 2020

Introduction

This exam includes four questions that test different aspects of your learning during this course: data wrangling, data visualisation and statistics. Each question is broken down into a number sub questions.

You must answer questions 1-3; you must answer EITHER question 4 OR 5

Your work should be handed in as a single PDF. Each question should be clearly indicated

For each question you must provide the R code you used to answer the question. The code should include comments to explain what you are doing. The code should be provided as text using a fixed-width font such as `Courier`. The rest of your answers should be in another font (e.g. Times New Roman, Cambria, Ariel). You can use the Microsoft Word template provided alongside these questions as guidance.

- Plots and tables should have captions.
- Plots should be produced using `ggplot`.
- Axis labels are important.
- Reporting of any statistics should be appropriate to the type of analysis you have done.
- Reporting of methods and results should be written in the style of a scientific paper.

If you don't understand what is required for any of these questions you are encouraged to ask for help.

Hand in deadline is Friday 10th January 2020 at 12:00 CET

You MUST submit your work via Blackboard (not email!)

1) Data wrangling life history data (10 points)

The Anage database (`anage_data.csv`) is a large collection of data on the life history of animals. It includes information on life span, body size, generation time etc. for species from a wide range of taxonomic groups. In this question you will be using this data to produce a graph and some summary information. You may need to `filter`, `select`, `mutate` or otherwise manipulate the data before you use it. You may also like to rename some of the data columns for ease of use.

The length of gestation in humans is 9 months, but it varies across mammals. Gestation time (`Gestation.Incubation..days.`) is positively associated with birth weight (`Birth.weight..g.`), which makes sense since larger bodies take longer to build. It is interesting to consider if there are differences in this relationship between rodents (Order Rodentia) and their flying cousins - bats (Order Chiroptera).

a) Produce a graph showing the relationship between between the log transformed birth weight (x-axis) and log-transformed gestation length in days (y-axis) for bats and rodents.

The starting point for all of the parts of this question is the data. So I first read that in using `read.csv`.

```
setwd("/Users/jones/Dropbox/_SDU_Teaching/BB839 New Stats Course/")
anage <- read.csv("CourseData/anage_data.csv")
```

Then I can write code to subset the data to the two Orders I want using `filter` and I can rename the columns to something more user-friendly with `rename`. This step is not strictly necessary.

```
x <- anage %>%
  select(Class,Order,Family, Genus, Species,Birth.weight..g.,
         Gestation.Incubation..days.,Litter.Clutch.size) %>%
  filter(Order %in% c("Chiroptera","Rodentia")) %>%
  rename(birthWeight = Birth.weight..g.,gestation = Gestation.Incubation..days.,
         litterSize = Litter.Clutch.size)
```

After that I can use `ggplot` to make a nice Figure.

```
ggplot(x,aes(x = log(birthWeight), y = log(gestation),colour=Order)) +
  geom_point() +
  xlab("log(birth weight, grams)") +
  ylab("log(gestation duration, days)")
```

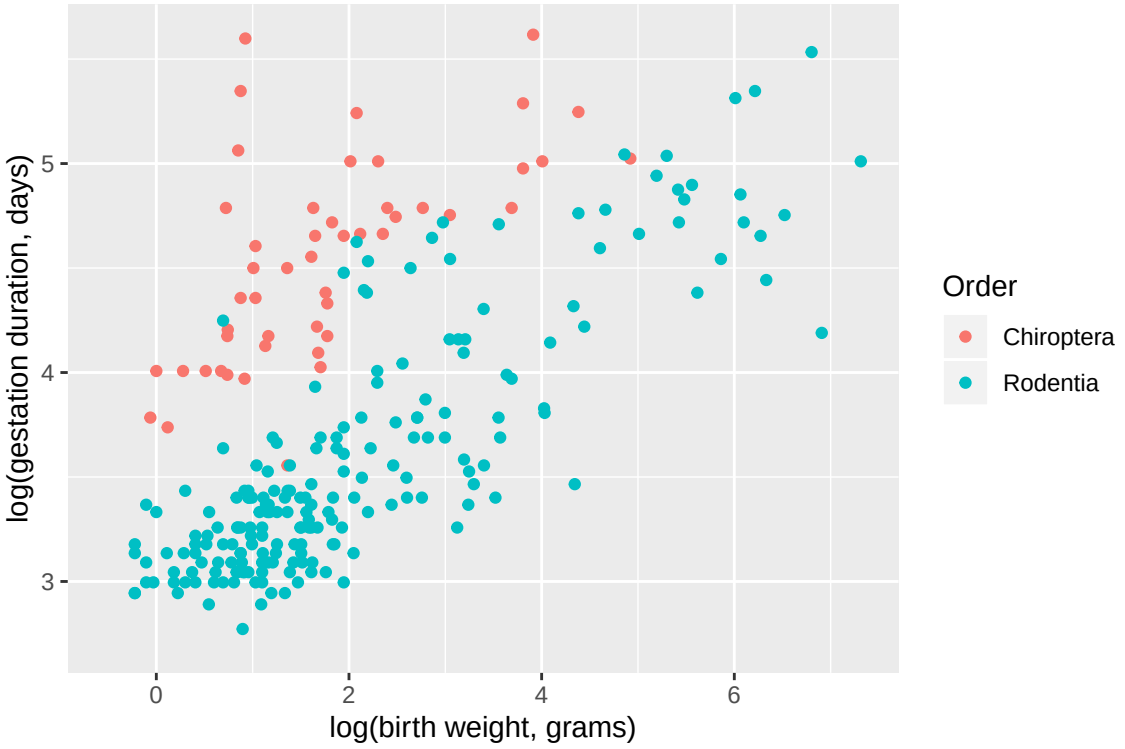


Figure 1: The association between (log transformed) birth weight and gestation duration for bats and mammals.

b) Produce a table showing the minimum, maximum, mean and median gestation times (in days) for rodents and bats.

I do this by first grouping by taxonomic Order with `group_by`, then using `summarise` to calculate the summary statistics. There are other ways to do do this. The final table could be made using word, or left like this.

```
x %>%
  group_by(Order) %>%
  summarise(minGestation = min(gestation,na.rm=TRUE),
            maxGestation = max(gestation,na.rm=TRUE),
            meanGestation = mean(gestation,na.rm=TRUE),
            medianGestation = median(gestation,na.rm=TRUE))
```

```
## # A tibble: 2 x 5
##   Order      minGestation maxGestation meanGestation medianGestation
##   <fct>             <int>        <int>         <dbl>           <dbl>
## 1 Chiroptera           35          275         110.             106.
## 2 Rodentia             15          253          45.1             30
```

You could nicely format this table like this, in Word.

Table 1: Table of summary statistics for bats and rodents

Order	minGestation	maxGestation	meanGestation	medianGestation
Chiroptera	35	275	109.7857	105.5
Rodentia	15	253	45.1280	30.0

c) Which are the species (latin names) with the minimum and maximum gestation lengths in Rodentia and Chiroptera?

I do this by first using `mutate` to create a new column of species names from `Genus` and `Species`, then I `select` the important columns and finally `filter` to those with the gestation values that I calculated in part B.

```
x %>%
  mutate(species = paste(Genus,Species)) %>%
  select(Order,species,gestation) %>%
  filter(Order == "Chiroptera" & gestation == 275)
```

```
##        Order        species gestation
## 1 Chiroptera Eidolon helvum       275
```

```
x %>%
  mutate(species = paste(Genus,Species)) %>%
  select(Order,species,gestation) %>%
  filter(Order == "Chiroptera" & gestation == 35)
```

```
##        Order          species gestation
## 1 Chiroptera Eptesicus fuscus        35
```

```
x %>%
  mutate(species = paste(Genus,Species)) %>%
  select(Order,species,gestation) %>%
  filter(Order == "Rodentia" & gestation == 253)
```

```
##      Order           species gestation
## 1 Rodentia Dinomys branickii       253
```

```
x %>%
  mutate(species = paste(Genus,Species)) %>%
  select(Order,species,gestation) %>%
  filter(Order == "Rodentia" & gestation == 15)
```

```
##      Order              species gestation
## 1 Rodentia Mesocricetus brandti        15
```

As before, you could format this information nicely in a table like this (or just write it in text):

Table 2: Species with the maximum and minimum gestation lengths in Chiroptera and Rodentia

Order	Stat	species
Chiroptera	Max	Eidolon helvum (275 days)
Chiroptera	Min	Eptesicus fuscus (35 days)
Rodentia	Max	Dinomys branickii (253 days)
Rodentia	Min	Mesocricetus brandti (15 days)

d) Explain why an ordinary least squares regression would not be the best approach to analyse this relationship.

For this question, I was looking for an understanding of one of the most important assumptions of OLS regression (which includes multiple regression, ANOVA, ANCOVA etc). This is that the data points are assumed be independent of eachother. In this case the data points are not indpendent of eachother because they are related via the phylogeny. Closely related species would be expected to cluster together and it would be "unfair" to give each point the same weight. This is a form of "pseudoreplication". There are other methods, such as "phylogenetic regression" which can account for these issues. These methods are beyond the scope of the course, but it is useful to have a good idea of where your knowledge ends and when to seek help.

This is an analagous issue to other kinds of pseudoreplication, like spatial or temporal pseudoreplication: if you are collecting data from a field, sampling from points that are too close together will be problematic. See here for a good explanation of this important topic: https://www.statisticsdonewrong.com/pseudoreplication.html

2) Industrial melanism (10 points)

Industrial melanism is a famous example of an evolutionary effect where dark pigmentation (melanism) evolves via natural selection where the environment is polluted with soot deposits. Darker individuals have a higher fitness in areas where their camouflage matches the polluted background surfaces better.

Some biologists have claimed that the knot grass moth (*Acronicta rumicis*) shows industrial melanism. You have collected data on the percentage of moths collected in light traps that are of the dark morph (`melanism.csv`). Your expectation is that you will find a higher percentage of dark morphs in the city where there is more pollution.

a) plot the data (e.g. with a box plot)

First I import the data, as usual:

```
setwd("/Users/jones/Dropbox/_SDU_Teaching/BB839 New Stats Course/")

melanism <- read.csv("CourseData/melanism.csv")
```

Then I can make a box plot like this (I could also add the jittered data points if I wanted to be flashy):

```
ggplot(melanism,aes(x = habitat,y = percentDark))+
  geom_boxplot()+
  geom_jitter()
```

b) carry out a randomisation test to determine if there is a significant difference in the frequency of the dark morph between city and countryside. Write (i) a brief method description and (ii) a summary of the results.

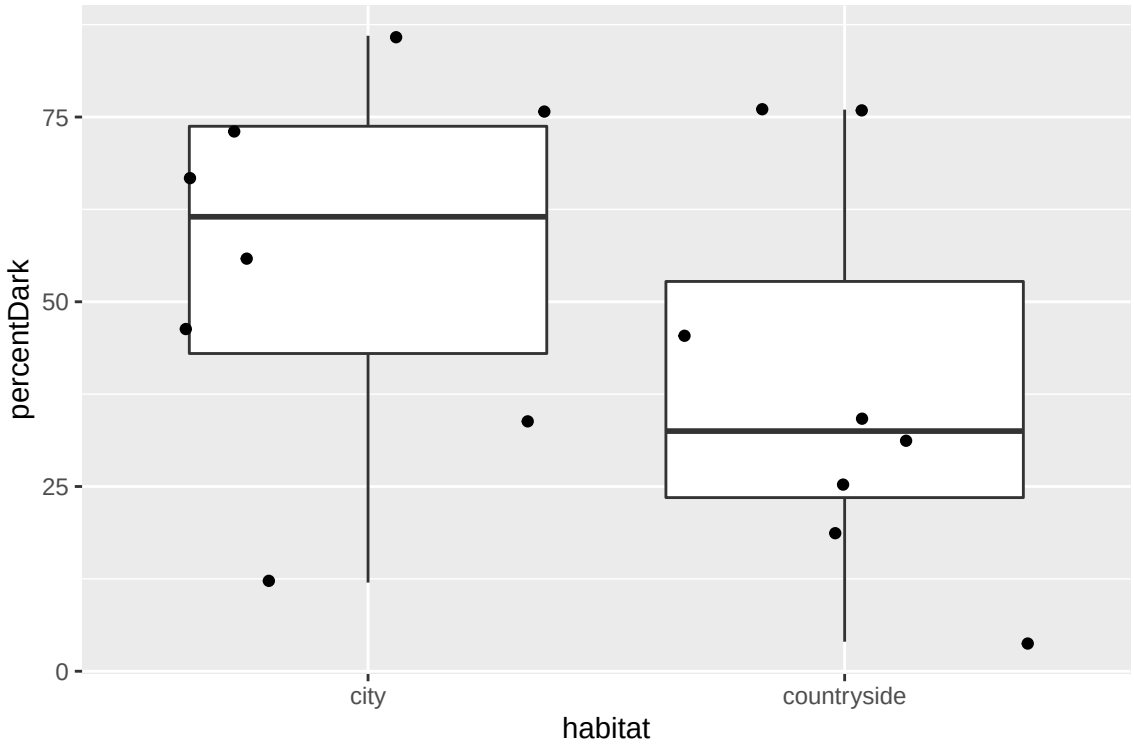


Figure 2: Box plot showing the difference in the proportion of dark morph moths in city and countryside habitats. Points show the raw data, jittered.

For the randomisation test I need to calculate the observed difference, then replicate a permutation where I shuffle the data many times (e.g. 1000 or 5000). I can then ask what proportion many of these differences were greater than the observed difference, which will give me the p-value. You could plot the frequency distribution of the differences, and with a line for the observed value, but this is not strictly necessary in my opinion (it can be useful for you to double check that things are working though).

```
obsDiff <- diff(melanism %>%
                  group_by(habitat) %>%
                  summarise(meanPerc = mean(percentDark)) %>%
                  pull(meanPerc))

diffs <- replicate(5000,diff(melanism %>%
                               mutate(habitat = sample(habitat)) %>%
                               group_by(habitat) %>%
                               summarise(meanPerc = mean(percentDark)) %>%
                               pull(meanPerc)))

sum(abs(diffs) >= abs(obsDiff))/5000
```

```
## [1] 0.186
```

I could write a description of the method something like this:

"To test whether the difference between two habitats in the frequency of the dark morph is statistically significant I did a 5000 replicate randomisation test with the null hypothesis being that there is no difference between the habitat means and the alternative hypothesis that the mean for the two habitats is different. I compared the observed difference to this null distribution to calculate a p-value in a two-sided test."

Then I could write up the results something like this:

"The observed mean values of the city and countryside were 56.25 and 38.75 respectively and the difference between them is therefore -17.5. I found that 878 of the absolute values of the 5000 null distribution replicates greater than or equal to my observed difference value. I conclude that the observed difference between the means of the two treatment groups is not statistically significant (p = 0.186). Therefore I accept the null hypothesis that there is no difference between the proportion of dark morph individuals in city and countryside habitats."

Note that you may get slightly different results for this because it is a random process. If you get very different p-values between runs of your test you should increase the number of replicates (e.g. from 1000 to 5000). A larger number of replicates should cause the p-value to vary less between runs.

3) Power in a planned experiment (10 points)

You are planning an experiment testing how the behaviour of small fish is affected by exposure to danger from predators. There are two treatments: (i) "exposed to predator cues" (from video footage of large predatory fish) and (ii) a "control" treatment, which is a safe predator-free environment. You would like to be able to detect a 20% reduction in swimming distance.

You record behaviour as the distance in meters travelled by fish as they swim around their tank within a 5 minute observation window. You hypothesise that when exposed to predator cues the fish will cover less distance and will tend to hide among the rocks and plants in the tank more. You have 5 fish tanks available for your study.

You have some preliminary data from a pilot study (`fishFear.csv`) which shows the distances moved by several individuals in a single pilot study.

a) Summarise the pilot study data to obtain mean and standard deviation.

First I import the pilot study data:

```
setwd("/Users/jones/Dropbox/_SDU_Teaching/BB839 New Stats Course/")
fish <- read.csv("CourseData/fishFear.csv")
```

Then I calculate the mean and standard deviation:

```
mean(fish$distance)
```

```
## [1] 2.038
```

```
sd(fish$distance)
```

```
## [1] 0.4767203
```

b) Conduct a power analysis based on the pilot study data to estimate the number of samples required to carry out your experiment with 80% power. Describe the results of this power analysis.

I can do a power analysis like this, where I replicate a t-test 1000 times for randomly generated samples based on the mean and standard deviation of the pilot study data. I assume that the standard deviation of the two groups is the same, but I calculate the mean for the treatment group by multiplying by 0.8 (i.e. a 20% decrease). I can then increase the sample size until I get a power of 80% (i.e. when the proportion of cases when the p-value is <0.05 exceeds 0.8).

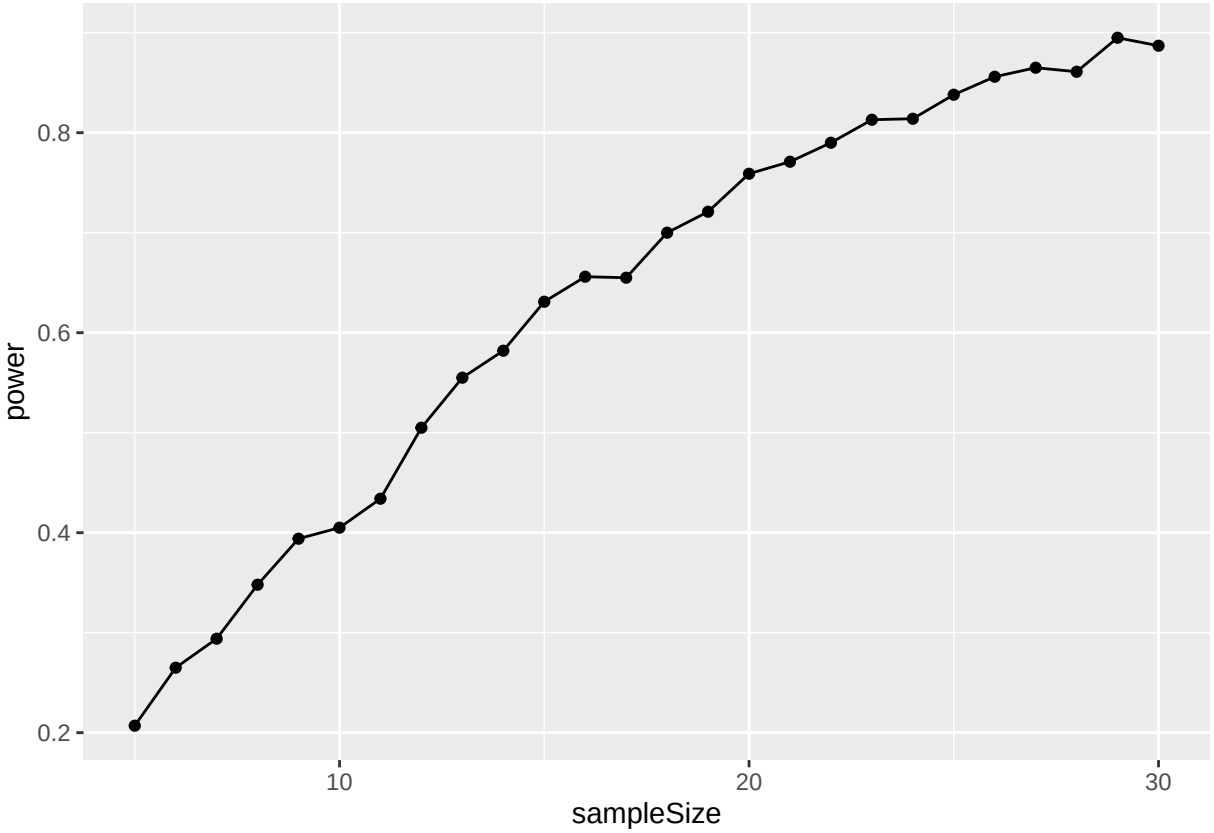
You should end up with a required sample size of about 23.

```r
sampleSize <- 23 #Set sample size here
meanDist <- mean(fish$distance)
sdDist <- sd(fish$distance)

#Repeat simulated test 1000 times
pValues <- replicate(1000,t.test(rnorm(sampleSize, mean = meanDist,sd = sdDist),
                        rnorm(sampleSize, mean = meanDist*0.8,sd = sdDist))$p.value)

#What proportion of the p-values are <0.05
sum(pValues<0.05)/1000
```

```
## [1] 0.8
```

Another clever way of doing this and producing a graph:

```r
meanDist <- mean(fish$distance)
sdDist <- sd(fish$distance)

powerResults <- data.frame(sampleSize = 5:30)

powerResults$power <- NULL
for(i in 1:nrow(powerResults)){
pvals <- replicate(1000,t.test(rnorm(powerResults$sampleSize[i],
                                     mean = meanDist,sd = sdDist),
       rnorm(powerResults$sampleSize[i],
             mean = meanDist*0.8,sd = sdDist))$p.value)

powerResults$power[i]<-sum(pvals<0.05)/1000
}

ggplot(powerResults,aes(x = sampleSize,y=power))+
  geom_point()+
  geom_line()
```



c) Describe how you might carry out your study, given your findings and the facilities available to you.

For this question I was hoping that you would come up with a workable plan, with a consideration of pseudoreplication. There are MANY ways you could have done this, I mainly wanted to see you think through a workable plan that weighed up the different factors: would putting several fish in the same tank mean that the measurements are not independent (the fish might influence eachother)? On the other hand, maybe you SHOULD put fish in groups (some fish would be freaked out by being alone). How would you capture data on several fish simulataneously? Did you realise that the sample size should be PER GROUP, not total number? Did you realise that the number you came up with is the MINIMUM, and that you could use more? This was a hard question because it was so open-ended, but it closely mimics the kind of problem you will be confronted with in real life.

4) The Titanic (20 points)

The `titanic.csv` data set includes a range of data on a subset of the passengers on the Titanic that sank in 1920 after hitting an iceberg. There were approximately 2200 passengers and crew on board and >1400 of these people died.

Your task is to analyse the data to find out how passenger class (`Pclass`) and gender (`Sex`) influenced survival probability. Survival is indicated with the numeric values 0 (died) and 1 (survived). Passenger class has values of 1, 2 or 3 and Sex is recorded as `male` or `female`.

a) Produce an appropriate graph of the raw data that illustrates survival differences among passenger classes and genders.

```
setwd("/Users/jones/Dropbox/_SDU_Teaching/BB839 New Stats Course/")

titanic <- read.csv("CourseData/titanic.csv")
```

```
titanic <- titanic %>%
  mutate(Pclass = as.factor(Pclass)) %>%
  mutate(Survived = as.factor(Survived))

(A<-ggplot(titanic,aes(x = Pclass,y = Survived,colour=Sex))+geom_jitter(alpha = 0.75))
```

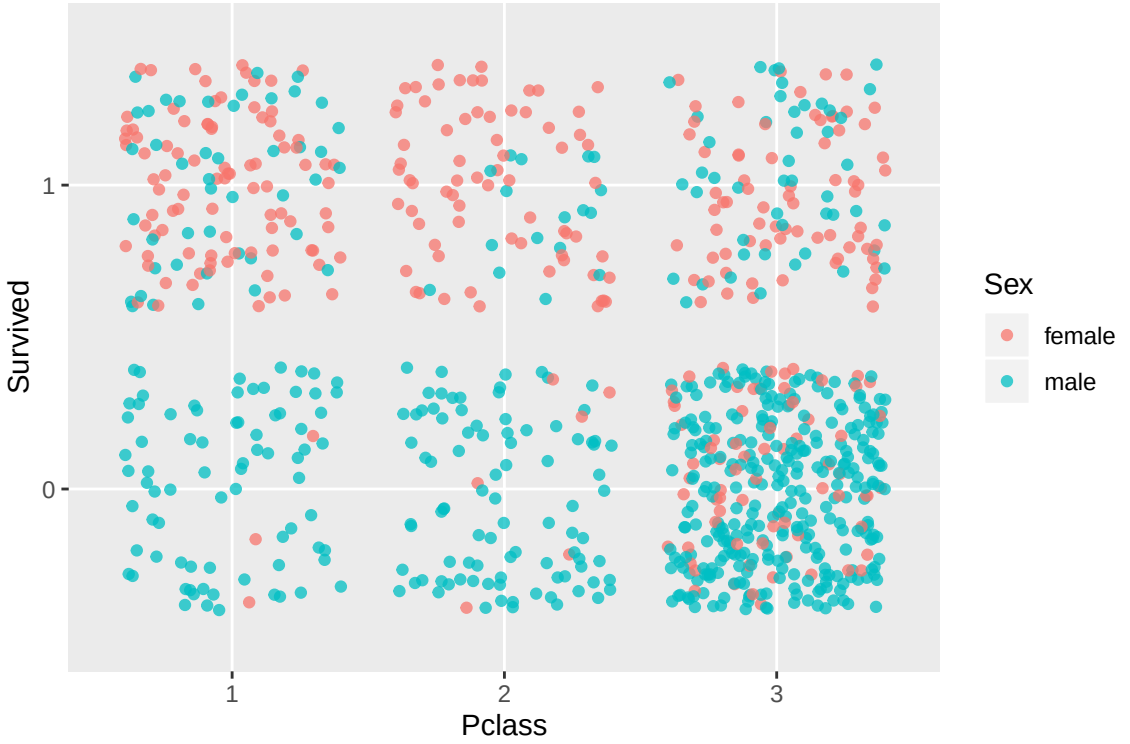
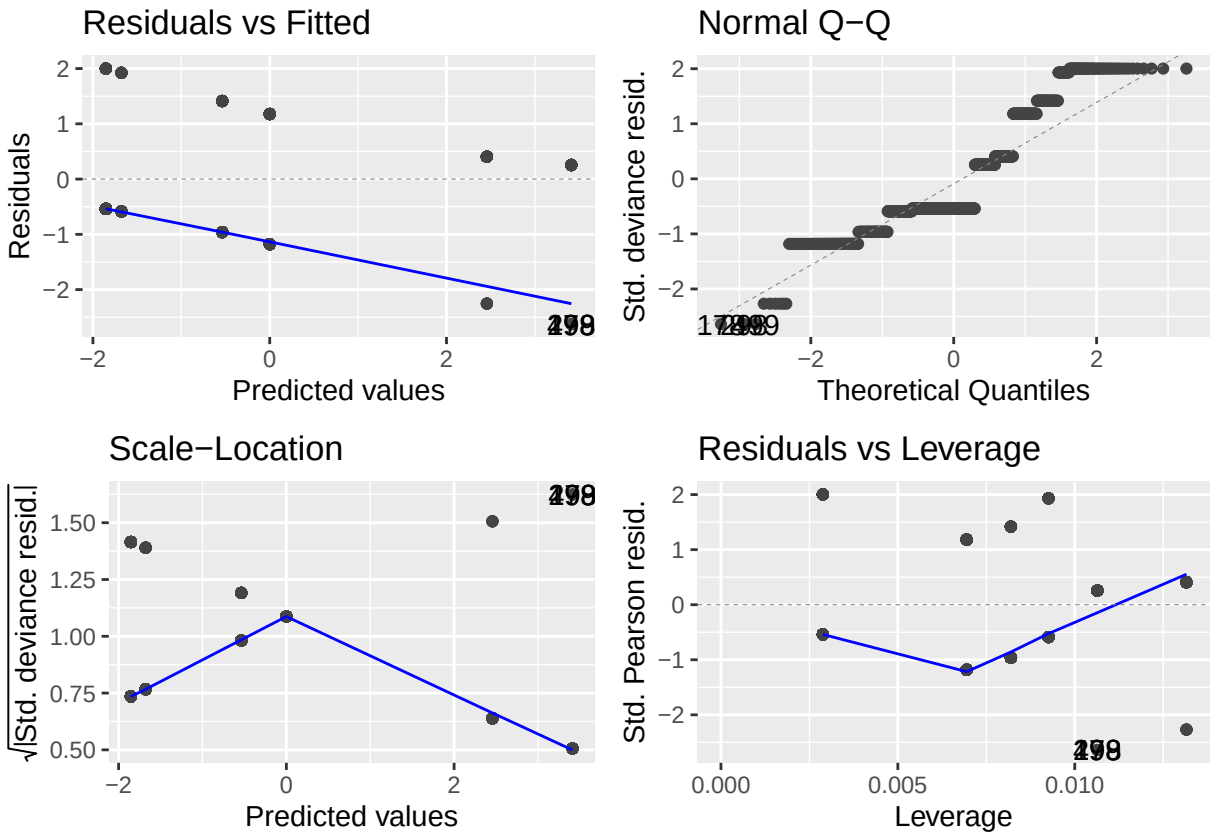


Figure 3: The fate (survival/death) of Titanic passengers. Each point represents a person and the data are divided among categories of passenger class (1, 2 or 3) and gender.

You could also have produced a bar plot, with counts of survivors/deaths.

b) Fit a suitable statistical model to estimate survival probability among passenger class and genders. Describe the method and then summarise the results produced by the model.

```
modA <- glm(Survived ~ Sex+ Pclass +Sex:Pclass,data = titanic,family = binomial)
library(ggfortify)
autoplot(modA)
```




```
anova(modA,test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                          890    1186.66
## Sex          1  268.851       889     917.80 < 2.2e-16 ***
## Pclass       2   90.916       887     826.89 < 2.2e-16 ***
## Sex:Pclass   2   28.791       885     798.10 5.598e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(modA)
```

```
##
## Call:
## glm(formula = Survived ~ Sex + Pclass + Sex:Pclass, family = binomial,
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6248  -0.5853  -0.5395   0.4056   1.9996
##
```

```
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.4122     0.5868   5.815 6.06e-09 ***
## Sexmale          -3.9494     0.6161  -6.411 1.45e-10 ***
## Pclass2          -0.9555     0.7248  -1.318  0.18737
## Pclass3          -3.4122     0.6100  -5.594 2.22e-08 ***
## Sexmale:Pclass2  -0.1850     0.7939  -0.233  0.81575
## Sexmale:Pclass3   2.0958     0.6572   3.189  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  798.1  on 885  degrees of freedom
## AIC: 810.1
##
## Number of Fisher Scoring iterations: 6
```

"I modelled the survival of the titanic passengers with a generalised linear model with binomial error structure. The explanatory variables were sex, passenger class and the interaction between them. I included the interaction because I was interested in whether the effect of sex depended on the passenger class."

"The results show that all three terms in the model were highly significant (Table 3). The Survival of males was generally much lower than that of females. In addition there was a slightly complicatef effect of passenger class: For males, the survival of 2nd and 3rd class passengers was markedly lower than that of 1st class passengers. For females, the survival of 1st and 2nd class passengers was very similar, but the survival of 3rd class passengers was much lower.|

Note that it might actually be easier to write this section after completing part (c). Then you could insert the actual values for these survival estimates. You could also use a table, like Table 4 below, which you can obtain from the input you will use to make the plot for (c)

```
## Warning in tidy.anova(temp): The following column names in ANOVA output
## were not recognized or transformed: Deviance, Resid..Df, Resid..Dev
```

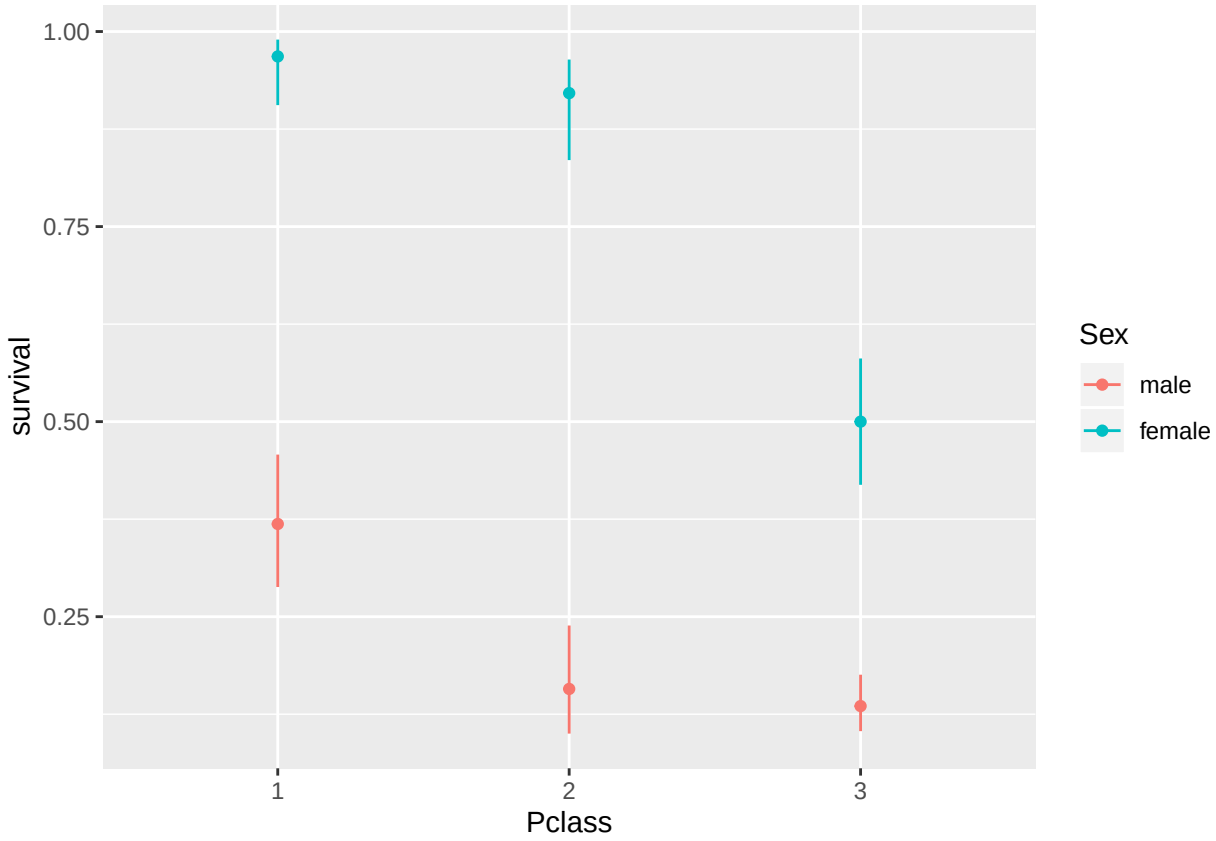
Table 3: Summary of the GLM results for the effect of sex and passenger class (Pclass), and their interaction, on survival on the Titanic

term	df	Deviance	Resid..Df	Resid..Dev	p.value
NULL	NA	NA	890	1186.6551	NA
Sex	1	268.85121	889	917.8039	0e+00
Pclass	2	90.91556	887	826.8884	0e+00
Sex:Pclass	2	28.79147	885	798.0969	6e-07

c) Make a plot showing the model's estimates and their 95% confidence intervals.

```
newDat <- expand.grid(Pclass = c("1","2","3"),Sex = c("male","female"))

pv <- predict(modA,newdata = newDat,se.fit = TRUE)
library(tidyverse)

newDat <- newDat %>%
  mutate(survival_LP = pv$fit,
         lowerCI_LP = pv$fit - 1.96*pv$se.fit,
```

```
         upperCI_LP = pv$fit + 1.96*pv$se.fit)

inverseFunction <- family(modA)$linkinv

newDat <- newDat %>%
  mutate(survival = inverseFunction(survival_LP)) %>%
  mutate(lowerCI = inverseFunction(lowerCI_LP)) %>%
  mutate(upperCI = inverseFunction(upperCI_LP))


(B <- ggplot(newDat,aes(x = Pclass,y = survival,colour=Sex))+
  geom_point() +
  geom_segment(aes(xend = Pclass,y = lowerCI,yend = upperCI)))
```




```
#ggpubr::ggarrange(A,B,ncol=2)
```

Here's a suitable table with the same information.

Table 4: Estimated survival and 95% confidence intervals passengers on the Titanic

Pclass	Sex	survival
1	male	0.369 (0.288 - 0.458)
2	male	0.157 (0.100 - 0.239)
3	male	0.135 (0.103 - 0.176)
1	female	0.968 (0.906 - 0.990)
2	female	0.921 (0.835 - 0.964)
3	female	0.500 (0.419 - 0.581)

5) Elephant poaching (20 points)

Illegal poaching activity is a major problem for African elephant conservation. You are provided with some data from some nature reserves across southern Africa (`elephantPoaching.csv`).

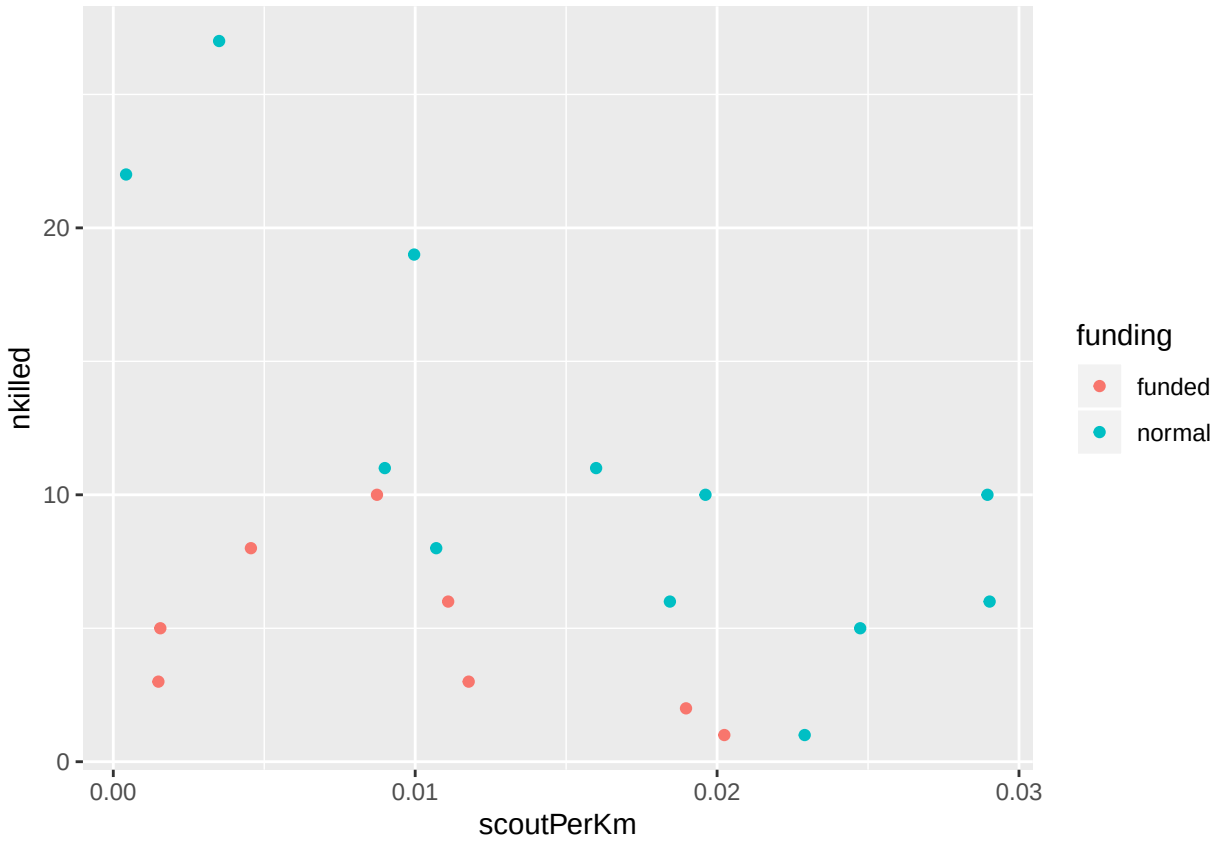
The data includes the number of elephants killed in 2016 (`nkilled`) and the number of scouts (the guards that protect the elephants) (`scoutPerKm`). A philanthropist billionnaire has been working in the area to provide additional tools to some reserves such as drones, high-tech communications and additional vehicles. This funding is recorded in the dataset column `funding` as `funded` (the ordinary parks with no extra funding are recorded as `normal`).

Use an appropriate statistical model to explore the relationship between elephants killed and the amount of cover provided by guards. Does this relationship differ depending on how well equipped the guards are?

a) Plot the data to show the relationship between the the number of scouts per unit area of park and the number of elephants killed. Colour code the points by whether the park management received extra funding or not.

Here's how to do the plot.

```
setwd("/Users/jones/Dropbox/_SDU_Teaching/BB839 New Stats Course/")

eleph <- read.csv("CourseData/elephantPoaching.csv")
ggplot(eleph,aes(x=scoutPerKm,y = nkilled,colour=funding)) +
  geom_point()
```



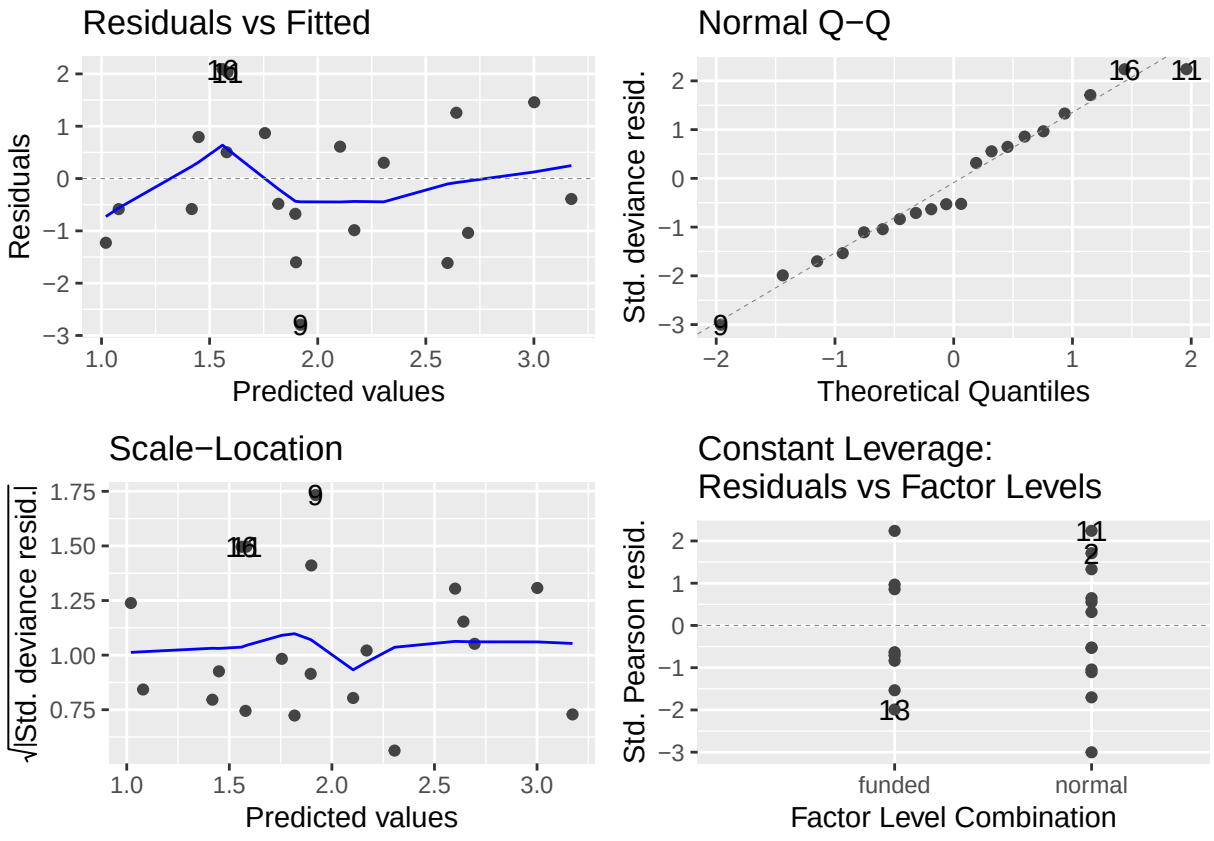
b) Fit a suitable statistical model to estimate the statistical relationship between guards per km, funding, and elephant deaths. Describe the method and then summarise the results produced by the model.

The appropriate model is a Poisson GLM. This is because the data are *counts* and failure to use an appropriate GLM would lead to predictions of negative numbers of elephants killed (not realistic!).

```
modA <- glm(nkilled ~ scoutPerKm + funding + scoutPerKm:funding,
            data = eleph,family = poisson)
anova(modA,test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: nkilled
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                    19     95.886
## scoutPerKm           1   18.879        18     77.007 1.393e-05 ***
## funding              1   44.544        17     32.463 2.487e-11 ***
## scoutPerKm:funding   1    0.101        16     32.362    0.7511
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(ggfortify)
autoplot(modA)
```




```
summary(modA)
```

```
##
```

```
## Call:
## glm(formula = nkilled ~ scoutPerKm + funding + scoutPerKm:funding,
##     family = poisson, data = eleph)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7942  -0.9999  -0.4369   0.8122   2.0903
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.9699     0.2586   7.619 2.56e-14 ***
## scoutPerKm                -46.9275    25.8355  -1.816   0.0693 .
## fundingnormal               1.2269     0.2949   4.160 3.18e-05 ***
## scoutPerKm:fundingnormal   -8.8041    27.6248  -0.319   0.7500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 95.886  on 19  degrees of freedom
## Residual deviance: 32.362  on 16  degrees of freedom
## AIC: 114.61
##
## Number of Fisher Scoring iterations: 4
```

"To examine the effect of number of scouts and funding on elephant deaths in parks I fitted a GLM with Poisson error structure. The explanatory variables were funding status (funded and normal), and the number of scouts per square kilometer. I also included the interaction between the two to check whether the effectiveness of scouts (i.e. the reduction in deaths per additional scout) depended on their funding.".

"The results show that both the number of scouts and funding had a significant effect on elephants killed. The addition of funding reduced the number of elephants killed significantly, as did the number of scouts. However, the fact that the interaction term was not significant indicates that the effect of increasing the amount of scouts did not depend on the amount of funding available (Table 4)"

Table 5: Summary of the GLM results for the effect of scout density and funding (and their interaction) on the number of elephant deaths.

term	df	Deviance	Resid..Df	Resid..Dev	p.value
NULL	NA	NA	19	95.886	NA
scoutPerKm	1	18.879	18	77.007	1.39e-05
funding	1	44.544	17	32.463	2.49e-11
scoutPerKm:funding	1	0.101	16	32.362	7.51e-01

c) Produce a plot that shows (in addition to the raw data points) the fitted values produced by your model and the uncertainty in those estimates.

You can make this plot like this:

```
#Create a data frame to predict from.
newData <- expand.grid(scoutPerKm = seq(0,0.03,0.001),funding = c("funded","normal"))

#Predict from the model using newData.
pv <- predict(modA,newData,se.fit =TRUE)

#Add predicted values (on scale of linear predictor).
```

```r
newData <- newData %>%
  mutate(nKilled_LP = pv$fit) %>%
  mutate(lower_LP = pv$fit - 1.96*pv$se.fit) %>%
  mutate(upper_LP = pv$fit + 1.96*pv$se.fit)

#get inverse function.
inverseLink <- family(modA)$linkinv

#Apply inverse function to the predicted values.

newData <- newData %>%
  mutate(nkilled = inverseLink(nKilled_LP)) %>%
  mutate(lowerCI = inverseLink(lower_LP)) %>%
  mutate(upperCI = inverseLink(upper_LP))

#plot with ribbons, lines and points.
ggplot(x3,aes(x=scoutPerKm,y = nkilled,colour=funding)) +
  geom_ribbon(data = newData,aes(x = scoutPerKm,ymin = lowerCI,
                                 ymax = upperCI,fill = funding),
              inherit.aes=FALSE,alpha = 0.4) +
  geom_line(data = newData,aes(x = scoutPerKm,y = nkilled,colour = funding))+
  geom_point() +
  xlab("Scouts per km^2") +
  ylab("Number of elephants killed in 2016")+
  NULL
```

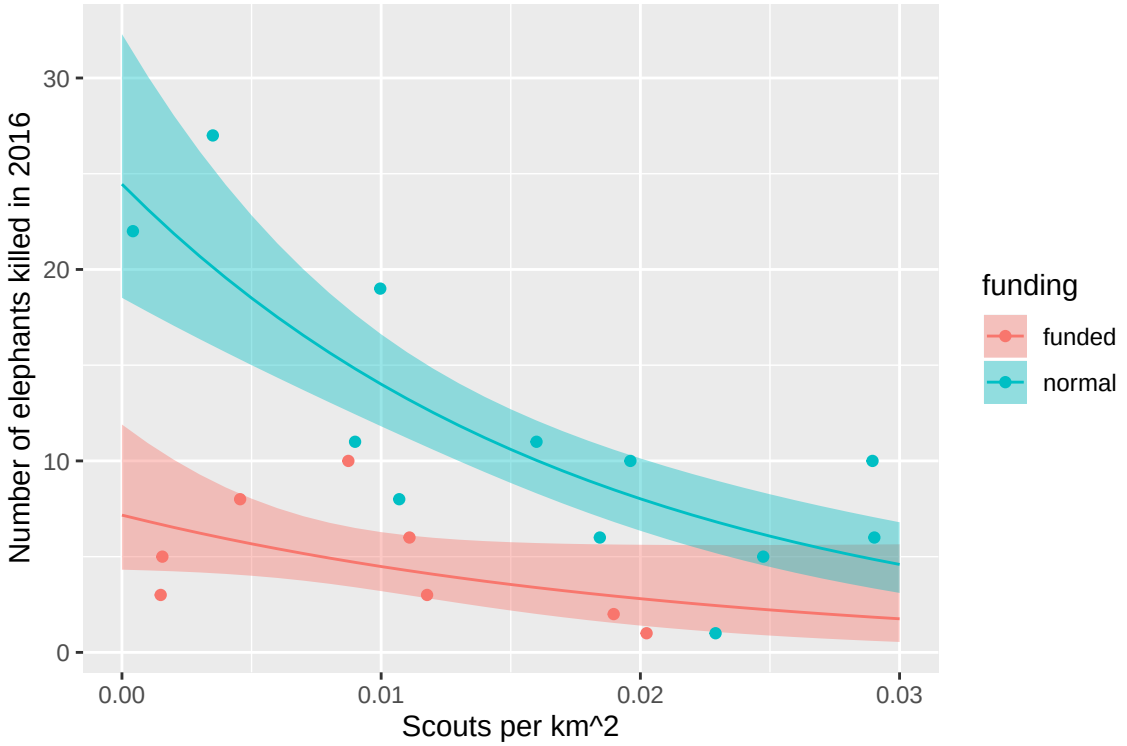


Figure 4: The association between funding, scout density and the number of elephants killed in 2016. Each point indicates the number of elephants killed in a particular park. The lines show the results of the fitted GLM.