MAST30027 Modern Applied, Statistics Assignment2

September 15,2019

Name: Tianyi Mo Student ID: 875556 Tutorial time: Tue 2.15pm Tutor: Qiuyi Li

1.Introduction

1.1 Background

This report is about evaluating chimpanzees prosocial tendency by analysing and fitting model to the data. The experiment has has two options, one is prosocial option, when human students participate the experiment, they nearly always choose the prosocial option when another student sits on the opposite side of the table. The question is whether a focal chimpanzee behaves similarly, choosing the prosocial option more often when another animal is present.

1.2 Data

There are four attributtes in the raw data. By looking through the dataset, it is found that every chimpanzee have the same number of instance if data (72 for each chimpanzee), therefore it is not tend to bias towards particular a chimpanzee. It should also be noted the the data is also balanced in the two conditions and two prosoc left, each chimpanzee has 16 experiment on each combination of condition and prosoc left.

These are the 4 attributes in the raw dataset. actor (1 to 7) condition (0 or 1): prosoc left (0 or 1) pulled left (0 or 1) Load the dataset

```
dataset = read.delim("assign2.txt", header = TRUE, sep = " ")
```

2. Preprocessing

2.1Create New Attribute

Since the question is whether a focal chimpanzee choosing the prosocial option more often when another animal is present. Therefore, an new attribute "prosocial_action" is created, it means whether prosoc_left and pulled_left are same. If these two atributes are same, chimpanzee perform prosocially and prosocial_action has value TRUE and otherwise it has value FALSE.

```
dataset['prosocial_action'] = (dataset$prosoc_left == dataset$pulled_left)
#first10 rows in the dataset after adding new attribute prosocial_action
dataset[1:10,]
```

```
##     actor condition prosoc_left pulled_left prosocial_action
## 1       1         0           0           0             TRUE
## 2       1         0           0           1            FALSE
## 3       1         0           1           0            FALSE
## 4       1         0           0           0             TRUE
## 5       1         0           1           1             TRUE
## 6       1         0           1           1             TRUE
## 7       1         0           1           0            FALSE
## 8       1         0           1           0            FALSE
## 9       1         0           0           0             TRUE
## 10      1         0           0           0             TRUE
```
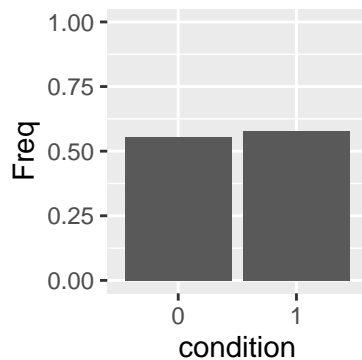
3.Visualization

3.1Plot of the percentage of prosocial action for all chimpanzees

```
prosocial = xtabs(prosocial_action ~ condition, data = dataset)
(prosocial_percentage = transform(prosocial, Freq=Freq/252))
```

```
##   condition      Freq
## 1         0 0.5555556
## 2         1 0.5793651
```

The average percentage of choose prosocial increase from 0.5555556 to 0.5793651 in condition 1.

```
# generate the bar plot
library(ggplot2)
ggplot(data=prosocial_percentage,aes(x=condition,y=Freq))+
  coord_cartesian(ylim=c(0, 1))+
  geom_bar(stat="identity",position="dodge")
```


On average, chimapnzees choose the prosocial option more frequently in condition 1 (when there is another chimpanzee in opposite). But the increase is very small and plot each chimpanzee may helps to see the trend more clearly.

3.2Plot of the percentage of prosocial action for each chimpanzee

```
prosocial = xtabs(prosocial_action ~ actor + condition, data = dataset)
prosocial_percentage = transform(prosocial, Freq=Freq/36)
```

```
# generate the bar plot
library(ggplot2)
ggplot(data=prosocial_percentage,aes(x=actor,y=Freq,fill=condition))+
  coord_cartesian(ylim=c(0, 1))+
  geom_bar(stat="identity",position="dodge")+
  ggtitle("Plot of the percentage of prosocial action for each actor")+
  geom_hline(yintercept=0.5, linetype=2, color = "red",size = 1)
```

Plot of the percentage of prosocial action for each actor
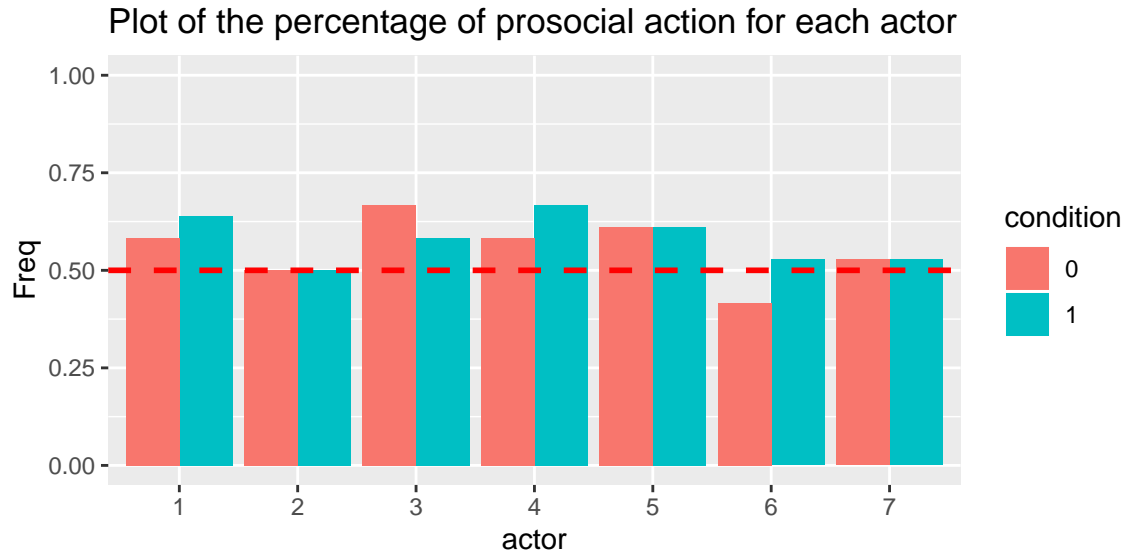
The plot show the percentage of prosocial action for each chimpanzee on two conditions. It can be seen that most chimpanzee tend to choos the prosocial optional in both condition. 3 of 7 chimpanzees dose not change the frequency of choosing. Three actors choose prosocial option more frequently while one actor perform differently. However, we cannot make conclusion based on the plot and models need to be fit.

3.3 Outlier

From the plot in 3.2, the second chimpanzee (actor 2) show strange result. In both conditions, it chooses proaction option at exactly at 0.5. By looking back through the dataset. It can be found that the it always choose the left one regardless of other factors. There is no similar behaviour like this among other chimpanzees and it might be an outlier.

```r
# Second chimpanzee always pulled the left lever
xtabs(~prosoc_left+pulled_left, data = dataset[dataset$actor == 2,])
```

```
##            pulled_left
## prosoc_left  1
##           0 36
##           1 36
```

```r
# Other chimpanzees pulled both left and right lever
# for example chimpanzees 1
xtabs(~prosoc_left+pulled_left, data = dataset[dataset$actor == 1,])
```

```
##            pulled_left
## prosoc_left  0  1
##           0 25 11
##           1 17 19
```

This might be an error in data collection process and might be nomal behaviour among chimpanzee species. Based on the data, I cannot decided which one it should be. This problem should be checked with the zoologist and experts with domain knowledges to decided whether to remove data from chimpanzee 2. In this report, assume it is not an error and thus the include data from chimpanzee 2.

3.Test dependency of two factors

H0 = prosocial_action & condition are independent H2 = prosocial_action & condition are dependent

```r
x = data.frame(xtabs( ~ prosocial_action + condition, data = dataset))

#using poisson regression to build model
model = glm(Freq~prosocial_action + condition, family = poisson,data = x )
deviance(model)
```

```
## [1] 0.2910418
```

```r
df.residual(model)
```

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

```r
pchisq(deviance(model), df = df.residual(model), lower.tail = FALSE)
```

```
## [1] 0.5895537
```

The Chi-square statistics is 0.5895537, it is larger than 0.05, thus it is insignificant and prosocial_action & condition are independent. However, there might be confounding factors that are ignored and we need to test them.

4.Testing confounding factors

actor, prosoc_left and pulled_left might be confounding factors

4.1 Testing whether actor is confounding factor

H0 : prosocial_action + condition + actor are independet H1 : prosocial_action + condition + actor are dependet

```
#test all 3 factor are independent
x = xtabs( ~ prosocial_action + condition + actor, data = dataset)
model1 = glm(Freq~ prosocial_action + condition + actor, family = poisson,data = x )
deviance(model1)
```

```
## [1] 9.704807
```

```
pchisq(deviance(model1), df = df.residual(model1), lower.tail = FALSE)
```

```
## [1] 0.9598556
```

p-value = 0.9598556, thus prosocial_action, condition and actor are independent, thus actor is not and confounding a factor.

4.2 Testing whether pulled_left is confounding factor

H0 : prosocial_action + condition + pulled_left are independet H1 : prosocial_action + condition + pulled_left are dependet

```
#test all 3 factor are independent
x = xtabs( ~ prosocial_action + condition + pulled_left, data = dataset)
model1 = glm(Freq~ prosocial_action + condition + pulled_left, family = poisson,data = x )
deviance(model1)
```

```
## [1] 1.685609
```

```
pchisq(deviance(model1), df = df.residual(model1), lower.tail = FALSE)
```

```
## [1] 0.7933302
```

p-value = 0.7933302, thus prosocial_action, condition and pulled_left are independent, thus pulled_left is not and confounding a factor.

4.3 Testing whether prosoc_left is confounding factor

H0 : prosocial_action + condition + prosoc_left are independet H1 : prosocial_action + condition + prosoc_left are dependet

```
#test all 3 factor are independent
x = xtabs( ~ prosocial_action + condition + prosoc_left, data = dataset)
model1 = glm(Freq~ prosocial_action + condition + prosoc_left, family = poisson,data = x)
deviance(model1)
```

```
## [1] 14.43789
```

```
pchisq(deviance(model1), df = df.residual(model1), lower.tail = FALSE)
```

```
## [1] 0.006020994
```

p-value = 0.006020994,reject H0, thus 3 factors prosocial_action, condition and prosoc_left are dependent.

H0 : prosoc_left is independent of prosocial_action and condition H1 : prosoc_left is dependent of prosocial_action and condition

```
#test if prosoc_left is independent of prosocial_action and condition
model2 = glm(Freq~ prosocial_action * condition + prosoc_left, family = poisson,data = x )
deviance(model2)
```

[1] 14.14685

```
pchisq(deviance(model2), df = df.residual(model2), lower.tail = FALSE)
```

[1] 0.002711932

p-value = 0.002711932, reject H0,thus prosoc_left is dependent of prosocial_action and condition

H0:prosocial_action and condition are conditionally independent of prosoc_left H1:prosocial_action and condition are conditionally dependent of prosoc_left

```
#test if prosocial_action and condition are conditionally independent of prosoc_left
model3 = glm(Freq~ prosocial_action * prosoc_left + condition * prosoc_left, family = poisson,data = x )
deviance(model3)
```

[1] 1.443903

```
pchisq(deviance(model3), df = df.residual(model3), lower.tail = FALSE)
```

[1] 0.4858032

p-value = 0.4858032, p-value > 0.05,fail to reject H0,thus prosocial_action and condition are independent conditional on prosoc_left

5.Conclusion

Question of interest is the relationship between condition and pulled_left and prosoc_left, we can combine the data over all 7 chimpanzees.

```
cta = xtabs( ~ prosoc_left + condition + pulled_left, data = dataset)
##pearson's chisquared test
summary(cta)
```

```
## Call: xtabs(formula = ~prosoc_left + condition + pulled_left, data = dataset)
## Number of cases in table: 504
## Number of factors: 3
## Test for independence of all factors:
##   Chisq = 10.877, df = 4, p-value = 0.02798
```

The chisquare statistic is 10.877, with 4 degrees of freedom and p-value 0.02798. It is less than 0.05 thus is significant. Thus the three factors dependent.

Then test it for each chimpanzee.

Chimpanzee 1

```
# generate data for champanzee 1
data1 = dataset[dataset$actor == 1,]
x = data.frame(xtabs( ~ prosocial_action + condition, data = data1))

#using poisson regression to build model
model = glm(Freq~prosocial_action + condition, family = poisson,data = x )
deviance(model)
```

[1] 0.2339193

```r
df.residual(model)
```

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

```r
pchisq(deviance(model), df = df.residual(model), lower.tail = FALSE)
```

```
## [1] 0.6286328
```

Chimpanzee 2

Chimpanzee 2 performed a little wired compare to other chimpanzees, it always pull on the left lever and might whatever other factors are. It might be an outlier

```r
# generate data for champanzee 1
data2 = dataset[dataset$actor == 2,]
x = data.frame(xtabs( ~ prosocial_action + condition, data = data2))

#using poisson regression to build model
model = glm(Freq~prosocial_action + condition, family = poisson,data = x )
deviance(model)
```

```
## [1] -1.332268e-15
```

```r
df.residual(model)
```

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

```r
pchisq(deviance(model), df = df.residual(model), lower.tail = FALSE)
```

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

Chimpanzee 3

```r
# generate data for champanzee 3
data3 = dataset[dataset$actor == 3,]
x = data.frame(xtabs( ~ prosocial_action + condition, data = data3))

#using poisson regression to build model
model = glm(Freq~prosocial_action + condition, family = poisson,data = x )
deviance(model)
```

```
## [1] 0.534171
```

```r
df.residual(model)
```

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

```r
pchisq(deviance(model), df = df.residual(model), lower.tail = FALSE)
```

```
## [1] 0.4648585
```

Chimpanzee 4

```r
# generate data for champanzee 4
data4 = dataset[dataset$actor == 4,]
x = data.frame(xtabs( ~ prosocial_action + condition, data = data4))

#using poisson regression to build model
```

```
model = glm(Freq~prosocial_action + condition, family = poisson,data = x )
deviance(model)
```

[1] 0.534171

```
df.residual(model)
```

[1] 1

```
pchisq(deviance(model), df = df.residual(model), lower.tail = FALSE)
```

[1] 0.4648585

Chimpanzee 5

```
# generate data for champanzee 1
data5 = dataset[dataset$actor == 5,]
x = data.frame(xtabs( ~ prosocial_action + condition, data = data5))

#using poisson regression to build model
model = glm(Freq~prosocial_action + condition, family = poisson,data = x )
deviance(model)
```

[1] 1.332268e-15

```
df.residual(model)
```

[1] 1

```
pchisq(deviance(model), df = df.residual(model), lower.tail = FALSE)
```

[1] 1

Chimpanzee 6

```
data6 = dataset[dataset$actor == 6,]
x = data.frame(xtabs( ~ prosocial_action + condition, data = data6))

#using poisson regression to build model
model = glm(Freq~prosocial_action + condition, family = poisson,data = x )
deviance(model)
```

[1] 0.8935135

```
df.residual(model)
```

[1] 1

```
pchisq(deviance(model), df = df.residual(model), lower.tail = FALSE)
```

[1] 0.344527

Chimpanzee 7

```
data7 = dataset[dataset$actor == 7,]
x = data.frame(xtabs( ~ prosocial_action + condition, data = data6))

#using poisson regression to build model
```

```r
model = glm(Freq~prosocial_action + condition, family = poisson,data = x )
deviance(model)
```

```
## [1] 0.8935135
```

```r
df.residual(model)
```

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

```r
pchisq(deviance(model), df = df.residual(model), lower.tail = FALSE)
```

```
## [1] 0.344527
```
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

ctrl + alt + I