MODERN APPLIED STATISTICS (MAST30027): ASSIGNMENT 2

Akira Takihara Wang
School of Computing and Information Systems
The University of Melbourne
Student ID: 913391

September 15, 2019

1 Introduction

All R code is provided at the end of the report.

The data originates from an experiment conducted to evaluate the prosocial tendencies of chimpanzees.

The experiment is conducted as follows:

> A focal chimpanzee sits at the end of a long table with two separate levers on either side of the table. On this table, there are two dishes which contain desirable food items which are connected to a contraption connecting to their respective levers. When a lever is pulled by the focal chimpanzee, the corresponding food dishes are delivered towards the opposite ends of the table. In all trials, both dishes on the focal animals' side contain desirable food items, so regardless of the lever pulled, the focal chimpanzee will always receive food.

The initial experiment conducted on humans suggests that humans will almost always choose the prosocial option, so the variation done on chimpanzees will hopefully follow a similar pattern.

2 Methadology

An initial look at the given dataframe yields 504 observations with 4 attributes, where all data points are given integer types as a default. According to the given data, all attributes other than "actor" are Boolean (0 or 1). Hence, we need to justify our use of each attribute:

- The "actor" attribute is an ID, so it should be treated as a factor.
- All remaining attributes are Boolean (and not integer), so they should also be treated as factors.

It is also known that a chimpanzee is prosocial given that it satisfies the following conditions:

1. Pulls the left lever (1) **and** the prosocial option is on the left side (1)
2. Pulls the right lever (0) **and** the prosocial option is on the right side (0).

Therefore, a new attribute "is_prosoc" is created by finding the Boolean of "pulled_left" == "prosoc_left".

From an initial look at the final data, roughly 57% of the instances indicate "is_prosoc" to be True, but this includes the Control experiments. By conditioning the data to only have instances where "condition" to be (a partner is sitting across the table), a total of 58% of the instances have "is_prosoc" to be True, a measly increase from 56% during control tests. To put this into perspective, there are only an additional 6 tests during the actual experimentation that indicate "is_prosoc" to be True - pretty much no difference.

As such, a couple of questions can arise from this:

- Do the chimpanzee's care about the prosocial option or are they just pulling the lever they prefer most.
- Do chimpanzees have a natural handedness which they tend more towards to during experimentation.
- Is there any statistically significant increases that may suggest chimpanzees be prosocial.

3 Independence Test of Attributes

3.1 Handedness of a Chimpanzee

One obvious choice of testing is to see if the handedness of a chimpanzee is independent during the control experiment. The expected result is that they are dependent because like humans, chimpanzees will have a preferred handedness.

A contingency table between the two factors is built, with the resulting p-value being very significant (6.99e-09). Thus, it is safe to now assume that chimpanzees will have a preferred handedness.

3.2 More Prosocial by Nature

Besides, it is a good idea to see if a chimpanzee may be more prosocial by nature. This can be done by checking for independence between "is_prosoc" and "actor".

The contingency is built using the two new factors, with a resulting insignificant p-value (0.2756). Therefore, the observed chimpanzees will act independently of the prosocial option regardless of the experiment

3.3 Conditioning on Handedness and Prosocial

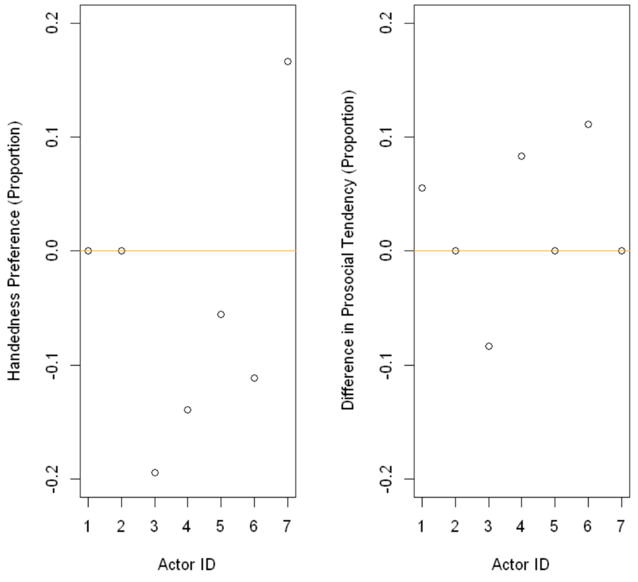


Figure 1: Dual plot of the differences in proportion of Handedness and natural prosocial tendencies.

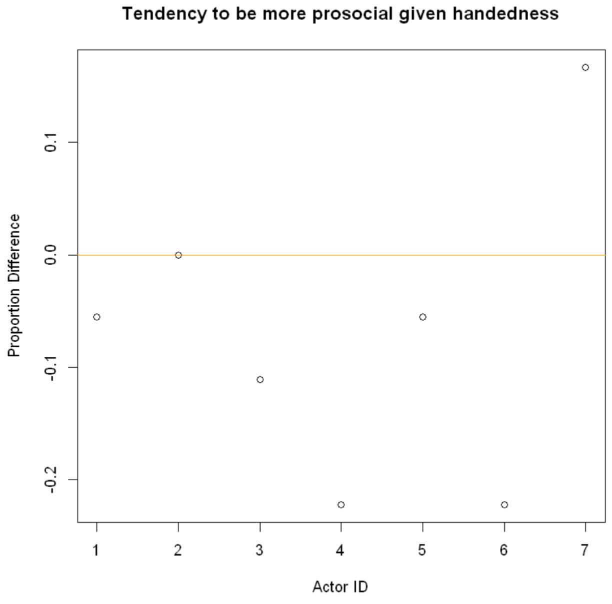


Figure 2: Plot of the tendency to be more prosocial given the handedness of the actor.

(Figure 1) is a dual plot of the above attributes and is calculated by the difference in proportion between the control and prosocial experiment conditioned on the actor. For the handedness:

- The difference in proportion for actor 1 and 2 is 0. A further look at the data suggests actor 1 has no preference (equal) when pulling a lever whilst actor 2 is completely biased on pulling the left lever.
- Actors 3 - 6 all share a generally negative difference, indicating the prosocial experiment (condition = 1) made them change their preferences in lever pulling. Perhaps this is suggestive of a minor change in preference, although the difference in proportion is only 0.2 - a very minor amount.

- Actor 7 seems to be the only chimpanzee to have a positive change, where a further look at the data reveals a biased toward the left lever (29 left / 7 right).

As for the natural prosocial tendency:

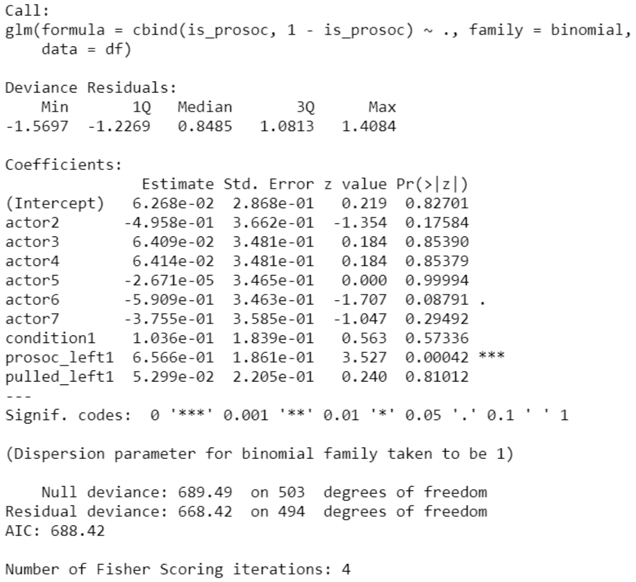
- This time, actors 2, 5 and 7 seem to have no difference in any prosocial behaviour - perhaps they do not care about the prosocial option, but more interested in pulling their favourite lever. Actor 2 is a prime example of such behaviour, opting in to pull the left lever regardless of the experiment.
- Actor 3 displays a negative difference in the proportion which connotes a potential increase in prosocial behaviour between experiments.
- In contrast, actor 1, 4 and 6 seemed to show a negative difference in prosocial behaviour, suggesting that they are choosing to not be prosocial.
- However, the difference in proportion is very small (less than 4 trials), so it seems that there is no change in prosocial behaviour between experiments.

(Figure 2) shows the tendency of chimpanzees choosing the more prosocial option with their handedness taken into account. This means the closer to 0 a chimpanzee's data point is, the less they care about being prosocial and care more about just pulling the lever. Visibly, actor2 will always pull the left lever signifying the chimpanzee number 2 will not care about the prosocial option and just pull the left lever for unknown reasons.

Therefore, only actors 4, 6 and 7 have any noticeable difference (proportion greater than 0.18) - but even then it is not significant enough to warrant a conclusion that *there is indeed an increase* of more prosocial options being chosen when a partner is sitting across them. Actor 1, 3 and 5 do show some minor difference in proportion, but it is too small and can be attributed to natural variance in the decision making the process of the chimpanzee (proportion smaller than 0.1).

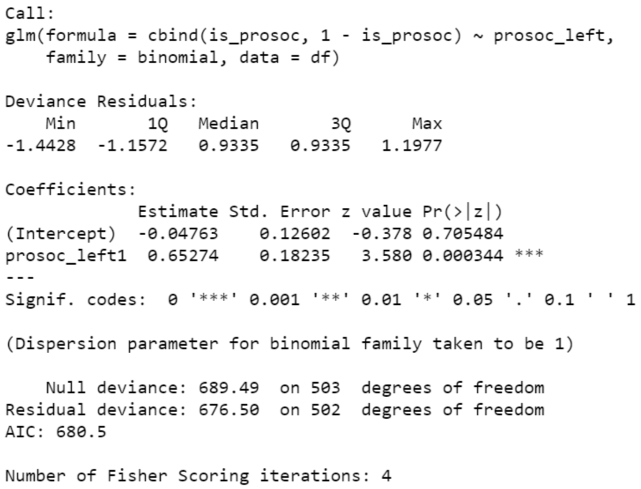
4 Fitting a Binomial Model

Since the data is True or False, a Binomial Regression model seems like a good choice for this problem. The response variable will be "is_prosoc", and the predictor variables will be "actor", "prosoc_left", "pulled_left" and "condition".



```
Call:
glm(formula = cbind(is_prosoc, 1 - is_prosoc) ~ ., family = binomial,
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5697  -1.2269   0.8485   1.0813   1.4084

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.268e-02  2.868e-01   0.219  0.82701
actor2       -4.958e-01  3.662e-01  -1.354  0.17584
actor3        6.409e-02  3.481e-01   0.184  0.85390
actor4        6.414e-02  3.481e-01   0.184  0.85379
actor5       -2.671e-05  3.465e-01   0.000  0.99994
actor6       -5.909e-01  3.463e-01  -1.707  0.08791 .
actor7       -3.755e-01  3.585e-01  -1.047  0.29492
condition1    1.036e-01  1.839e-01   0.563  0.57336
prosoc_left1  6.566e-01  1.861e-01   3.527  0.00042 ***
pulled_left1  5.299e-02  2.205e-01   0.240  0.81012
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 689.49  on 503  degrees of freedom
Residual deviance: 668.42  on 494  degrees of freedom
AIC: 688.42

Number of Fisher Scoring iterations: 4
```

Figure 3: R Summary Output of the Fitted Model (Binomial with Logit Link).



```
Call:
glm(formula = cbind(is_prosoc, 1 - is_prosoc) ~ prosoc_left,
    family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4428  -1.1572   0.9335   0.9335   1.1977

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.04763    0.12602  -0.378 0.705484
prosoc_left1  0.65274    0.18235   3.580 0.000344 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 689.49  on 503  degrees of freedom
Residual deviance: 676.50  on 502  degrees of freedom
AIC: 680.5

Number of Fisher Scoring iterations: 4
```

Figure 4: R Summary Output of the Final Fitted Model (Using AIC).

(Figure 1) shows the fitted model, where it is instantly observable that the majority of parameters other than "prosoc_left" seem to be irrelevant. Interestingly, there are very large p-values in "actor", with actor5 having a very unexpected p-value of 1.

4.1 Building a Better Model

Because there seem to be several irrelevant parameters, stepwise selection using AIC as the criterion will be performed to find a final model with relevant parameters. (Figure 2) displays this model, where it seems "prosoc_left" and the intercept parameter are the only parameters of importance.

The final model then becomes:

$$\text{is\_prosocial} = -0.04763 + 0.65274 \times \text{prosoc\_left1}. \tag{1}$$

The final parameter space of the final model is quite questionable. According to the final model - a chimpanzee can be "predicted" to be prosocial depending on the side of the prosocial option. In other words, the final model does not even care about which chimpanzee is being trialled or if they pulled the lever connected to the prosocial option, but rather it only cares about which side the prosocial option is on.

4.2 Diagnostic Plots

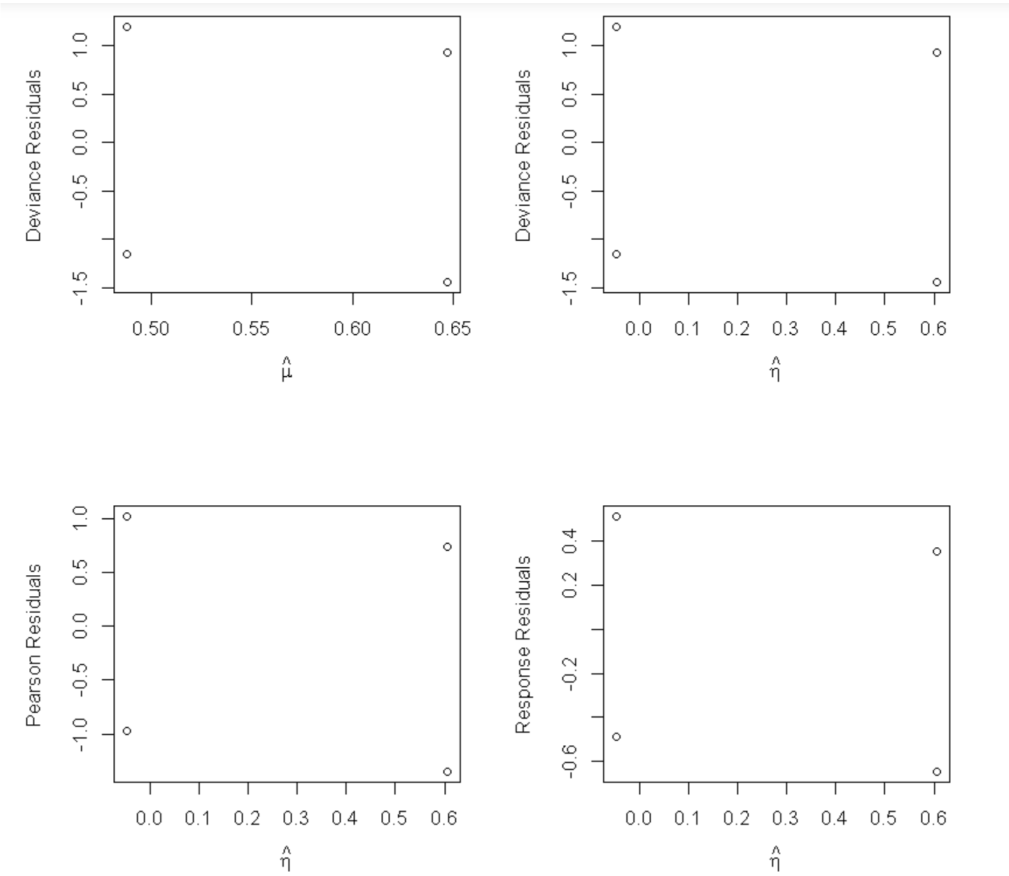


Figure 5: Diagnostic Plots of the Final Fitted Model.

We look at the diagnostic plots to see if there are any trends in the residuals. (Figure 3) shows the plots, and there seems to be an absence of any sign of trends. It is safe to say that the final model seems like a good fit, although the background data will suggest that this final model is a very questionable fit.

4.3 Model Deviance

If the model explains the data well, it should approximately fit the χ^2 distribution. Calculations yield a significant p-value (2.139e-147≈0 due to Machine Error) which advises there is a considerable amount of unexplainable information in the model. There may be a possibility of extra variability in the data which the model is unable to capture, so we should estimate the dispersion parameter to check.

4.4 Checking for Overdispersion

Since this model assumes $\hat{\phi} = 1$, we can check for overdispersion using the Pearson χ^2 statistic:

$$X^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y})}{n - p}, \tag{2}$$

where n is the number of data points and p is the number of parameters in the fitted model.

After calculation, the dispersion parameter is found to be 1.0039, which is approximately 1. Hence, there seems no presence of any extra variability in the data from the fitted model.

5 Conclusion

Ultimately, it seems the chimpanzees do not really care about the prosocial option, but simply care about pulling a lever. The final fitted model also reaches the same conclusion, with only the constant intercept parameter and "prosoc_left" being relevant. But, within the experiment, the "prosoc_left" being True or False should not matter. This is because the experiment is only interested in whether or not a chimpanzee also pulled the lever corresponding to the prosocial option and **not** the side of which the prosocial option is on. Obviously, one would expect the final model to capture information such as "pulled_left" and "prosoc_left" which determine if the chimpanzee chose the prosocial option. Without both being present, there is information missing and the fitted model will act similar to a random classifier dependent on a single attribute. Future considerations of the study include obtaining a larger sample size as the deviance of the fitted model clearly concluded that something was missing, with overdispersion being non-existent as calculated.

In [12]:

```
df = read.table("assign2.txt", header=TRUE)
str(df)
head(df)
```

```
'data.frame':	504 obs. of  4 variables:
 $ actor      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ condition  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prosoc_left: int  0 0 1 0 1 1 1 1 0 0 ...
 $ pulled_left: int  0 1 0 0 1 1 0 0 0 0 ...
```

actor	condition	prosoc_left	pulled_left
1	0	0	0
1	0	0	1
1	0	1	0
1	0	0	0
1	0	1	1
1	0	1	1

In [13]:

```
# Create attribute
df$actor = factor(df$actor)
df$condition = factor(df$condition)
df$prosoc_left = factor(df$prosoc_left)
df$pulled_left = factor(df$pulled_left)
df$is_prosoc = as.integer(df$prosoc_left == df$pulled_left)
```

In [14]:

```
# Also create only control and only experiment
single = df[which(df$condition == 0),]
pair = df[which(df$condition == 1),]
```

Binomial Regression

In [15]:

```
binom.model = glm(cbind(is_prosoc, 1 - is_prosoc) ~ ., df, family=binomial)
summary(binom.model)
```

```
Call:
glm(formula = cbind(is_prosoc, 1 - is_prosoc) ~ ., family = binomial,
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5697  -1.2269   0.8485   1.0813   1.4084

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.268e-02  2.868e-01   0.219  0.82701
actor2       -4.958e-01  3.662e-01  -1.354  0.17584
actor3        6.409e-02  3.481e-01   0.184  0.85390
actor4        6.414e-02  3.481e-01   0.184  0.85379
actor5       -2.671e-05  3.465e-01   0.000  0.99994
actor6       -5.909e-01  3.463e-01  -1.707  0.08791 .
actor7       -3.755e-01  3.585e-01  -1.047  0.29492
condition1    1.036e-01  1.839e-01   0.563  0.57336
prosoc_left1  6.566e-01  1.861e-01   3.527  0.00042 ***
pulled_left1  5.299e-02  2.205e-01   0.240  0.81012
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 689.49  on 503  degrees of freedom
Residual deviance: 668.42  on 494  degrees of freedom
AIC: 688.42

Number of Fisher Scoring iterations: 4
```

In [16]:

```
bm.step = step(binom.model, trace=FALSE)
summary(bm.step)
```

```
Call:
glm(formula = cbind(is_prosoc, 1 - is_prosoc) ~ prosoc_left, 
    family = binomial, data = df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.4428  -1.1572   0.9335   0.9335   1.1977  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -0.04763    0.12602  -0.378 0.705484    
prosoc_left1  0.65274    0.18235   3.580 0.000344 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 689.49  on 503  degrees of freedom
Residual deviance: 676.50  on 502  degrees of freedom
AIC: 680.5

Number of Fisher Scoring iterations: 4
```

It seems there may be a preference as to where the prosocial option is. We can try see if the mean tells us anything.

In [17]:

```
pchisq(668.48, 1, lower.tail=FALSE)
```

2.13875460832085e-147

In [18]:

```
singlebm.step.mu = predict(bm.step, newdata=single, se.fit=TRUE, type="link")
singlebm.step.muFit = singlebm.step.mu$fit
singlebm.step.muError = singlebm.step.mu$se.fit
alpha = dnorm(0.975)

# Compute fitted CI
widthsingle = alpha * singlebm.step.muError

CIfitsingle = cbind(singlebm.step.muFit-widthsingle, singlebm.step.muFit+widthsingle)
CIsingle = data.frame(CIfitsingle[,1], singlebm.step.muFit, CIfitsingle[,2])

# Rename for clarity
names(CIsingle)[names(CIsingle)=="CIfitsingle...1."] = "2.5%"
names(CIsingle)[names(CIsingle)=="singlebm.step.muFit"] = "Fitted Value (single)"
names(CIsingle)[names(CIsingle)=="CIfitsingle...2."] = "97.5%"
```

In [19]:

```
head(CIsingle)
```

2.5%	Fitted Value (single)	97.5%
-0.07888433	-0.04762805	-0.01637177
-0.07888433	-0.04762805	-0.01637177
0.57242526	0.60511383	0.63780240
-0.07888433	-0.04762805	-0.01637177
0.57242526	0.60511383	0.63780240
0.57242526	0.60511383	0.63780240

In [20]:

```
pairbm.step.mu = predict(bm.step, newdata=pair, se.fit=TRUE, type="link")
pairbm.step.muFit = pairbm.step.mu$fit
pairbm.step.muError = pairbm.step.mu$se.fit
alpha = dnorm(0.975)

# Compute fitted CI
widthpair = alpha * pairbm.step.muError

CIfitpair = cbind(pairbm.step.muFit-widthpair, pairbm.step.muFit+widthpair)
CIpair = data.frame(CIfitpair[,1], pairbm.step.muFit, CIfitpair[,2])

# Rename for clarity
names(CIpair)[names(CIpair)=="CIfitpair...1."] = "2.5%"
names(CIpair)[names(CIpair)=="pairbm.step.muFit"] = "Fitted Value (pair)"
names(CIpair)[names(CIpair)=="CIfitpair...2."] = "97.5%"
```

In [21]:

```
head(CIpair)
```

	2.5%	Fitted Value (pair)	97.5%
37	-0.07888433	-0.04762805	-0.01637177
38	-0.07888433	-0.04762805	-0.01637177
39	0.57242526	0.60511383	0.63780240
40	-0.07888433	-0.04762805	-0.01637177
41	-0.07888433	-0.04762805	-0.01637177
42	-0.07888433	-0.04762805	-0.01637177

In [22]:

```
F = binom.model$family$linkinv
```

In [23]:

```
CI2logit = data.frame(F(CIsingle[,1]), F(CIsingle[,2]), F(CIsingle[,3]))
# Rename for clarity
names(CI2logit)[names(CI2logit)=="F.CIsingle...1.."] = "2.5%"
names(CI2logit)[names(CI2logit)=="F.CIsingle...2.."] = "Fitted Value (pair)"
names(CI2logit)[names(CI2logit)=="F.CIsingle...3.."] = "97.5%"

head(CI2logit)
```

2.5%	Fitted Value (pair)	97.5%
0.4802891	0.4880952	0.4959071
0.4802891	0.4880952	0.4959071
0.6393226	0.6468254	0.6542565
0.4802891	0.4880952	0.4959071
0.6393226	0.6468254	0.6542565
0.6393226	0.6468254	0.6542565

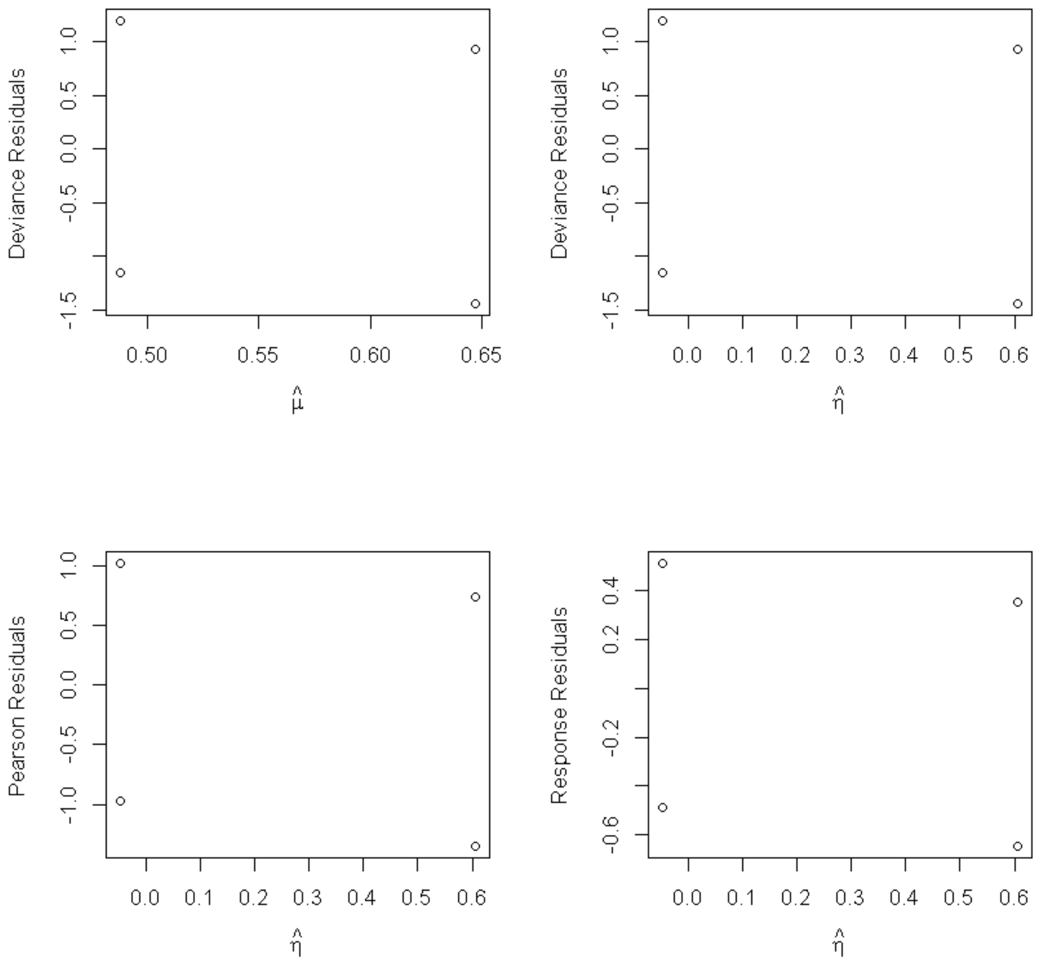
In [24]:

```
CI3logit = data.frame(F(CIpair[,1]), F(CIpair[,2]), F(CIpair[,3]))
# Rename for clarity
names(CI3logit)[names(CI3logit)=="F.CIpair...1.."] = "2.5%"
names(CI3logit)[names(CI3logit)=="F.CIpair...2.."] = "Fitted Value (pair)"
names(CI3logit)[names(CI3logit)=="F.CIpair...3.."] = "97.5%"

head(CI3logit)
```

2.5%	Fitted Value (pair)	97.5%
0.4802891	0.4880952	0.4959071
0.4802891	0.4880952	0.4959071
0.6393226	0.6468254	0.6542565
0.4802891	0.4880952	0.4959071
0.4802891	0.4880952	0.4959071
0.4802891	0.4880952	0.4959071

In [25]:

```
# Residual Plots
par(mfrow=c(2,2))
plot(residuals(bm.step) ~ predict(bm.step, type="response"), xlab=expression(hat(mu)),
ylab="Deviance Residuals")
plot(residuals(bm.step) ~ predict(bm.step, type="link"), xlab=expression(hat(eta)), yla
b="Deviance Residuals")
plot(residuals(bm.step, type="pearson") ~ predict(bm.step, type="link"),
     xlab=expression(hat(eta)), ylab="Pearson Residuals")
plot(residuals(bm.step, type="response") ~ predict(bm.step, type="link"),
     xlab=expression(hat(eta)), ylab="Response Residuals")
```



There seems to be no trend at all

Let's check for overdispersion - we've assumed $\hat{\phi} = 1$ so far...

In [26]:

```
n = dim(df)[1]
p = 2 # number of fitted params
phi.hat = sum(residuals(bm.step, type="pearson")^2) / (n - p)
phi.hat
```

1.00398406374501

Nice, our dispersion parameter is pretty much 1, so the model is good.

Contingency Table - Testing for Independence

In [27]:

```
# The first most interesting attribute to look at is the handedness, which is tested fo
r during control experiment
t1 = table(single$pulled_left, single$actor)
summary(t1)
```

```
Number of cases in table: 252 
Number of factors: 2 
Test for independence of all factors:
	Chisq = 49.14, df = 6, p-value = 6.99e-09
```

It is very significant, so there is a dependency between the handedness of each chimp. This is expected since it is natural to have a preferred handedness (humans are more right handed).

In [28]:

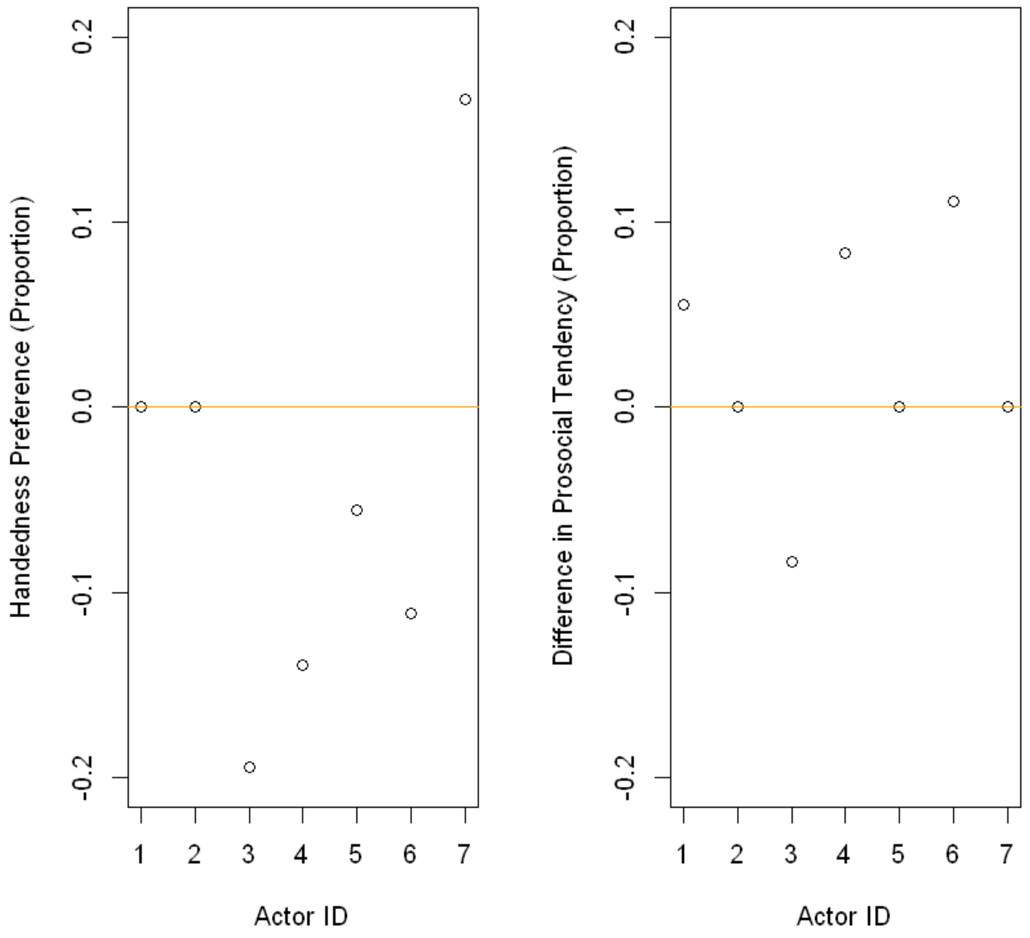
```
# We can also see if some chimps may be more prosocial by nature
t2 = table(df$is_prosoc, df$actor)
summary(t2)
```

```
Number of cases in table: 504 
Number of factors: 2 
Test for independence of all factors:
	Chisq = 7.518, df = 6, p-value = 0.2756
```

So during experiments, we observe that chimps are acting independently regardless of the prosocial option

In [29]:

```
par(mfrow=c(1,2))

l = (prop.table(table(single$actor, single$pulled_left), 1) - prop.table(table(pair$act
or, pair$pulled_left), 1))[,1]
plot(l, ylim=c(-.2,.2), xlab="Actor ID",
     ylab="Handedness Preference (Proportion)")
lines(x=c(0,8), y=c(0,0), col="orange")

# If the value is 0, then there is either no preference (50/50) or complete bias (100/
0) to a side
# Negative difference indicates a tendency to pull the left lever compared to the contr
ol,
# and a positive indicates a tendency to pull the right lever compared to the control

diff = (prop.table(table(single$actor, single$is_prosoc), 1) - prop.table(table(pair$ac
tor, pair$is_prosoc), 1))[,1]
plot(diff, ylim=c(-.2,.2), xlab="Actor ID",
     ylab="Difference in Prosocial Tendency (Proportion)")
lines(x=c(0,8), y=c(0,0), col="orange")

# A 0 value inidcates no change in their original choice (i.e their actions are more in
diciative of just pulling the
# lever and not choosing the prosocial option). From observations, there is only a very
minor difference between
# choosing the prosocial option.
```



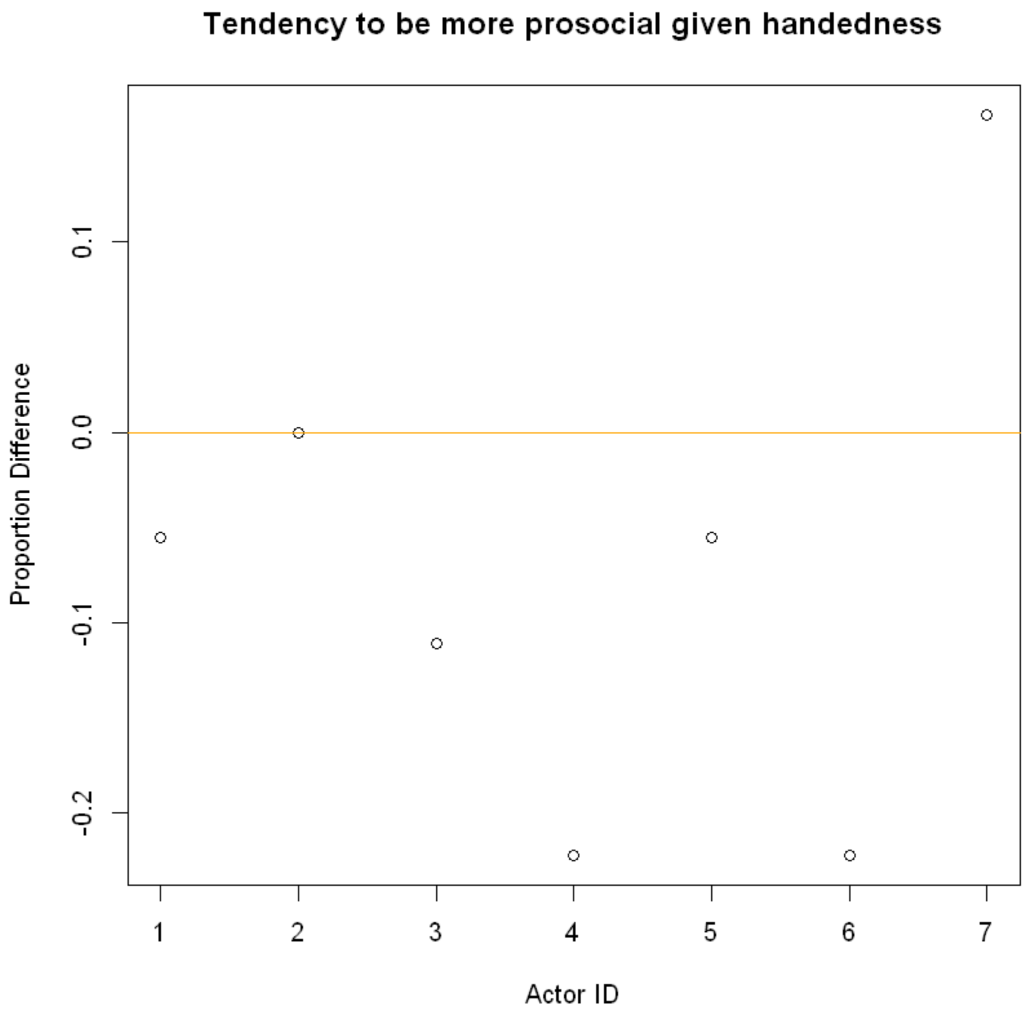
```
In [30]:
```

```r
cbind(l, diff)
```

```
          l        diff
 0.00000000  0.05555556
 0.00000000  0.00000000
-0.19444444 -0.08333333
-0.13888889  0.08333333
-0.05555556  0.00000000
-0.11111111  0.11111111
 0.16666667  0.00000000
```

In [31]:

```
l = (prop.table(table(single$actor, single$pulled_left), 1) - prop.table(table(pair$act
or, pair$pulled_left), 1))[,1]
diff = (prop.table(table(single$actor, single$is_prosoc), 1) - prop.table(table(pair$ac
tor, pair$is_prosoc), 1))[,1]
plot(l - diff, xlab="Actor ID", ylab="Proportion Difference", main="Tendency to be more
prosocial given handedness")
lines(x=c(0,8), y=c(0,0), col="orange")
```


Tendency to be more prosocial given handedness
Proportion Difference
Actor ID