MAST30027 - Assignment 2

A first step is to create a response variable `pulled_prosoc` for whether the chimpanzee chose the prosocial option, and to aggregate the data.

```r
library(dplyr)
library(tidyr)
library(faraway)
library(MASS)
data <- read.table("assign2.txt", header=TRUE)
data$pulled_prosoc <- (data$prosoc_left == data$pulled_left)
( agg <- data %>%
        group_by(actor, condition, prosoc_left) %>%
        summarise(pulled_prosoc=sum(prosoc_left == pulled_left), total=n()) %>%
        as.data.frame() )
```

```
##    actor condition prosoc_left pulled_prosoc total
## 1      1         0           0            12    18
## 2      1         0           1             9    18
## 3      1         1           0            13    18
## 4      1         1           1            10    18
## 5      2         0           0             0    18
## 6      2         0           1            18    18
## 7      2         1           0             0    18
## 8      2         1           1            18    18
## 9      3         0           0            13    18
## 10     3         0           1            11    18
## 11     3         1           0            15    18
## 12     3         1           1             6    18
## 13     4         0           0            12    18
## 14     4         0           1             9    18
## 15     4         1           0            16    18
## 16     4         1           1             8    18
## 17     5         0           0            12    18
## 18     5         0           1            10    18
## 19     5         1           0            13    18
## 20     5         1           1             9    18
## 21     6         0           0             4    18
## 22     6         0           1            11    18
## 23     6         1           0             8    18
## 24     6         1           1            11    18
## 25     7         0           0             4    18
## 26     7         0           1            15    18
## 27     7         1           0             1    18
## 28     7         1           1            18    18
```

Note that `condition` is a treatment factor, whereas `prosoc_left` is a blocking factor, as the chimpanzees may have handedness preferences. The equal group sizes along the `total` column shows that this is a completely balanced design.

In fact, it appears that some of the individuals to display handedness preferences. This can be quantified by performing a binomial test for each individual separately. The null hypothesis is that an animal pulls the left lever with a probability of $p = 0.5$. The results are as follows:

```
( n.left <- aggregate(data$pulled_left, list(data$actor), sum)$x )
```

```
## [1] 30 72 25 25 30 46 64
```

```
sapply(n.left, function (x) binom.test(x, 72)$p.value)
```

```
## [1] 1.945052e-01 4.235165e-22 1.277460e-02 1.277460e-02 1.945052e-01
## [6] 2.446091e-02 5.765519e-12
```

In particular, chimpanzees 2 and 7 display very strong handedness bias, almost always choosing the left lever. Moreover, the null hypothesis is rejected for most individuals at a significance level of 5%. This indicates that prosoc_left should be one of the variables of the models to be used.

A straightforward way to proceed is to fit an additive model using binomial regression with a logit link. The proportion to be estimated is pulled_prosoc / total. The results are as follows:

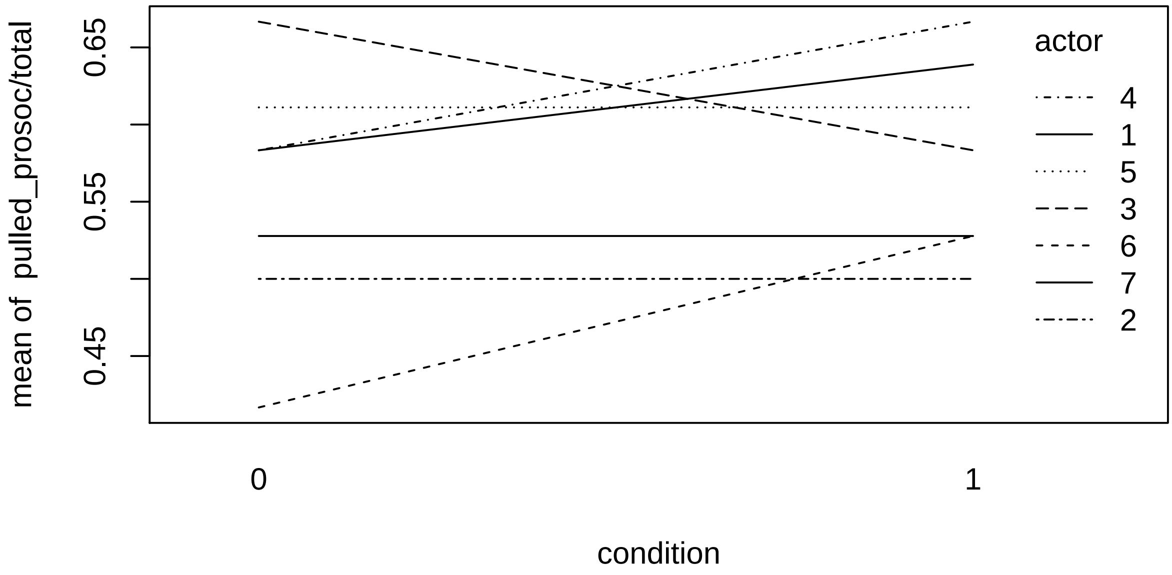
```
agg$actor <- as.factor(agg$actor)
additive <- glm(cbind(pulled_prosoc,total-pulled_prosoc)~actor+condition+prosoc_left,
                family=binomial, agg)
summary(additive)
```

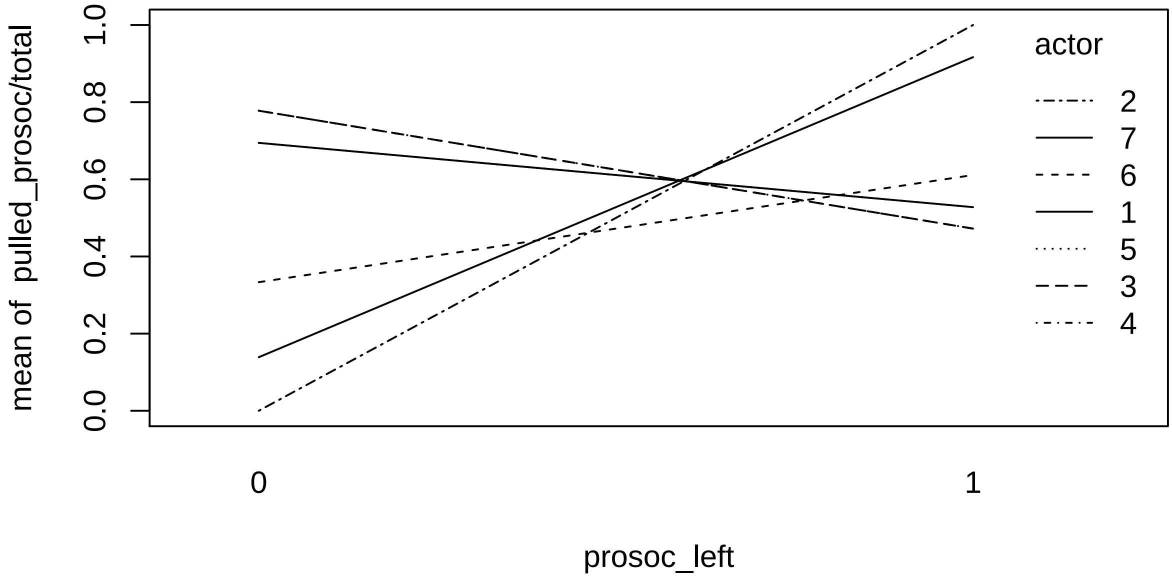
```
## 
## Call:
## glm(formula = cbind(pulled_prosoc, total - pulled_prosoc) ~ actor + 
##     condition + prosoc_left, family = binomial, data = agg)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.4996  -1.7085   0.3726   1.5595   4.4996  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  8.227e-02  2.750e-01   0.299 0.764827    
## actor2      -4.646e-01  3.423e-01  -1.357 0.174750    
## actor3       6.041e-02  3.476e-01   0.174 0.862037    
## actor4       6.041e-02  3.476e-01   0.174 0.862037    
## actor5      -2.252e-15  3.464e-01   0.000 1.000000    
## actor6      -5.790e-01  3.426e-01  -1.690 0.091070 .  
## actor7      -3.502e-01  3.426e-01  -1.022 0.306625    
## condition    1.011e-01  1.837e-01   0.551 0.581850    
## prosoc_left  6.635e-01  1.840e-01   3.606 0.000311 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 197.21  on 27  degrees of freedom
## Residual deviance: 176.20  on 19  degrees of freedom
## AIC: 266.3
## 
## Number of Fisher Scoring iterations: 4
```

Notice that the deviance of 176.20 is very large compared to its corresponding 19 degrees of freedom. This suggests that the additive model is inadequate. One possible reason is because that some of the variables interact with each other. There are three possible sets of pairwise interactions, as there are three explanatory variables in agg, namely actor, condition, and prosoc_left. No interaction between condition and prosoc_left will be explored, since there should be no interaction between the treatment factor and the blocking factor by design.

The following is an interaction plot between the chimpanzee individuals and the presence of another chimpanzee, and an interaction plot between the chimpanzee and whether the prosocial option corresponds to the left lever:

```r
with(agg, interaction.plot(condition, actor, pulled_prosoc / total))
with(agg, interaction.plot(prosoc_left, actor, pulled_prosoc / total))
```





For both plots, the gradients are different, so a refined model will include these pairwise interactions. Note that in the first plot, the gradients (change in mean of pulled_prosoc due to condition) seem relatively small, with actors 2 and 7 displaying no change in pulled_prosoc when condition varies. On the other hand, in the second plot, the gradients (change in mean of pulled_prosoc due to prosoc_left) is relatively large for actors 2 and 7. This agrees with the aforementioned observation that chimpanzees 2 and 7 are heavily biased towards a handedness preference.

An interaction model is fitted as follows:

```
full <- glm(cbind(pulled_prosoc,total-pulled_prosoc)~actor*(condition+prosoc_left),
            family=binomial, agg)
summary(full)
```

```
##
## Call:
## glm(formula = cbind(pulled_prosoc, total - pulled_prosoc) ~ actor *
##     (condition + prosoc_left), family = binomial, data = agg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1437  -0.5223   0.0000   0.4938   1.7699
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         7.032e-01  4.315e-01   1.630 0.103197
## actor2             -2.236e+01  6.233e+03  -0.004 0.997138
## actor3              7.588e-01  6.560e-01   1.157 0.247357
## actor4              3.621e-01  6.331e-01   0.572 0.567349
## actor5              1.177e-01  6.143e-01   0.192 0.847988
## actor6             -1.649e+00  6.204e-01  -2.659 0.007844 **
## actor7             -2.528e+00  7.485e-01  -3.377 0.000732 ***
## condition           2.412e-01  4.918e-01   0.490 0.623882
## prosoc_left        -7.122e-01  4.932e-01  -1.444 0.148760
## actor2:condition   -2.412e-01  7.197e+03   0.000 0.999973
## actor3:condition   -6.378e-01  7.133e-01  -0.894 0.371184
## actor4:condition    1.555e-01  7.133e-01   0.218 0.827399
## actor5:condition   -2.412e-01  6.947e-01  -0.347 0.728504
## actor6:condition    2.450e-01  6.985e-01   0.351 0.725771
## actor7:condition   -2.412e-01  8.993e-01  -0.268 0.788583
## actor2:prosoc_left  4.402e+01  7.197e+03   0.006 0.995120
## actor3:prosoc_left -6.638e-01  7.201e-01  -0.922 0.356627
## actor4:prosoc_left -6.638e-01  7.201e-01  -0.922 0.356627
## actor5:prosoc_left  2.425e-03  6.969e-01   0.003 0.997224
## actor6:prosoc_left  1.874e+00  6.997e-01   2.678 0.007407 **
## actor7:prosoc_left  4.935e+00  9.161e-01   5.387 7.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 197.206  on 27  degrees of freedom
## Residual deviance:  13.017  on  7  degrees of freedom
## AIC: 127.12
##
## Number of Fisher Scoring iterations: 18
```

The residual deviance of 13.017 is now more comparable to the corresponding 7 degrees of freedom.

To test whether condition is a significant variable in this model, one can fit a reduced model with condition removed. The results are:

```
no.cond <- glm(cbind(pulled_prosoc,total-pulled_prosoc)~actor*prosoc_left,
               family=binomial, agg)
summary(no.cond)
```

```
##
## Call:
## glm(formula = cbind(pulled_prosoc, total - pulled_prosoc) ~ actor *
##     prosoc_left, family = binomial, data = agg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1946  -0.2538   0.0000   0.2581   1.7699
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          8.210e-01  3.618e-01   2.269  0.02326 *
## actor2              -2.247e+01  5.089e+03  -0.004  0.99648
## actor3               4.318e-01  5.400e-01   0.800  0.42396
## actor4               4.318e-01  5.400e-01   0.800  0.42396
## actor5               2.331e-15  5.117e-01   0.000  1.00000
## actor6              -1.514e+00  5.059e-01  -2.993  0.00276 **
## actor7              -2.646e+00  6.026e-01  -4.390 1.13e-05 ***
## prosoc_left         -7.098e-01  4.923e-01  -1.442  0.14939
## actor2:prosoc_left   4.402e+01  7.197e+03   0.006  0.99512
## actor3:prosoc_left  -6.542e-01  7.173e-01  -0.912  0.36173
## actor4:prosoc_left  -6.542e-01  7.173e-01  -0.912  0.36173
## actor5:prosoc_left  -2.601e-15  6.962e-01   0.000  1.00000
## actor6:prosoc_left   1.855e+00  6.959e-01   2.666  0.00769 **
## actor7:prosoc_left   4.932e+00  9.156e-01   5.387 7.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 197.206  on 27  degrees of freedom
## Residual deviance:  15.414  on 14  degrees of freedom
## AIC: 115.52
##
## Number of Fisher Scoring iterations: 18
```

A likelihood ratio test between the full model with interaction and the reduced model, where the null hypothesis is that the condition is not significant for modelling the response variable pulled_prosoc. This gives:

```
anova(no.cond, full, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(pulled_prosoc, total - pulled_prosoc) ~ actor * prosoc_left
## Model 2: cbind(pulled_prosoc, total - pulled_prosoc) ~ actor * (condition +
##     prosoc_left)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        14     15.415
## 2         7     13.017  7   2.3977   0.9346
```

The LRT returns a high p-value of 0.9346, which suggests that `condition`, i.e. the presence of another chimpanzee has no effect on the prosocial behaviour of the test chimpanzees. This conclusion is not too surprising as the interaction plots showed large gradients for the `prosoc_left` plot but small gradients for the `condition` plot.

A check for overdispersion for the reduced model, the dispersion parameter can be estimated by dividing the Pearson's χ^2-statistic by the corresponding number of degrees of freedom:

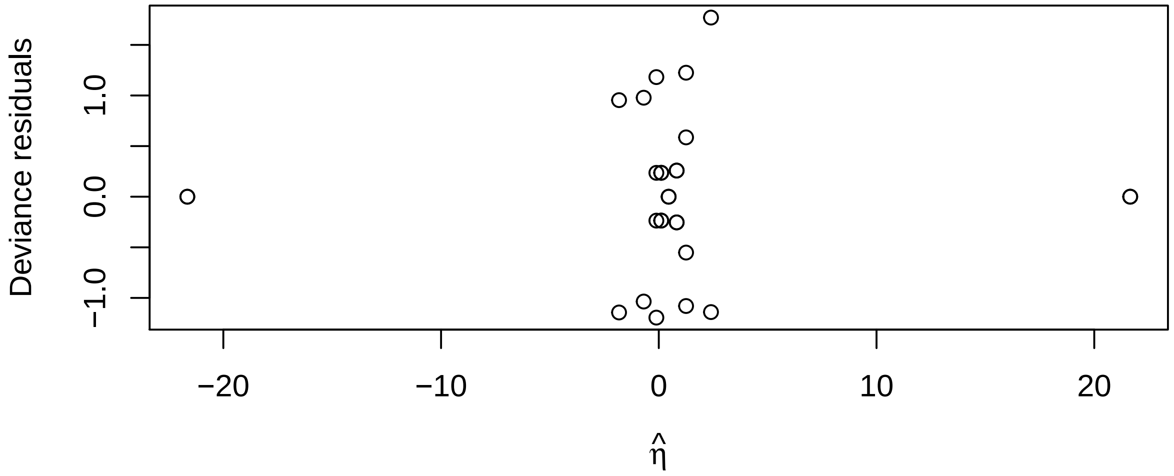
```
sum(residuals(no.cond, type="pearson") ^ 2) / 14
```

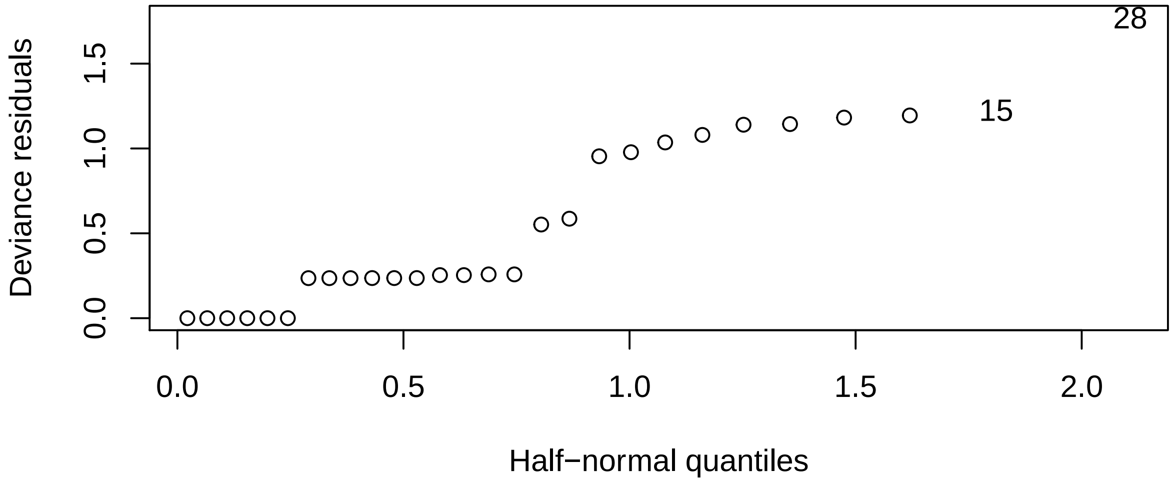
```
## [1] 0.9971355
```

Since the dispersion parameter estimate of 0.9971355 is very close to 1, it is safe to conclude that this model does not suffer from overdispersion.

Diagnostic plots:

```
res <- residuals(no.cond)
eta.hat <- predict(no.cond,type="link")
plot(res ~ eta.hat, xlab=expression(hat(eta)), ylab="Deviance residuals")
halfnorm(res, ylab="Deviance residuals")
```





In the first plot, there are two extreme fitted values. Upon inspection of the model coefficients, these values correspond to chimpanzee 2, who always pulls the left lever. Thus, the probability of chimpanzee 2 selecting the prosocial option when it is placed on the left is estimated to be almost 1, explaining the extreme fitted values. There is no visible trend in the centre of the plot. Neither is there anything alarming in the quantile plot, though the deviance residuals do not seem to follow a normal distribution.

The LRT suggested that the reduced model explains the data better than the model with an interaction term between actor and condition. As one last check, a model involving condition is built, but without interaction (though actor and prosoc_left are still allowed to interact as different chimpanzees have different degrees of handedness):

```
incl.cond <- glm(cbind(pulled_prosoc,total-pulled_prosoc)~actor*prosoc_left+condition,
               family=binomial, agg)
summary(incl.cond)
```

```
## 
## Call:
## glm(formula = cbind(pulled_prosoc, total - pulled_prosoc) ~ actor * 
##     prosoc_left + condition, family = binomial, data = agg)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.3386  -0.3818   0.0000   0.3818   1.7108  
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         7.531e-01  3.767e-01   1.999  0.04561 *  
## actor2             -2.248e+01  5.084e+03  -0.004  0.99647    
## actor3              4.322e-01  5.403e-01   0.800  0.42375    
## actor4              4.322e-01  5.403e-01   0.800  0.42375    
## actor5              0.000e+00  5.119e-01   0.000  1.00000    
## actor6             -1.516e+00  5.062e-01  -2.995  0.00275 ** 
## actor7             -2.648e+00  6.029e-01  -4.392 1.12e-05 ***
## prosoc_left        -7.105e-01  4.926e-01  -1.443  0.14916    
## condition           1.377e-01  2.143e-01   0.642  0.52070    
## actor2:prosoc_left  4.402e+01  7.190e+03   0.006  0.99512    
## actor3:prosoc_left -6.549e-01  7.177e-01  -0.913  0.36150    
## actor4:prosoc_left -6.549e-01  7.177e-01  -0.913  0.36150    
## actor5:prosoc_left -1.301e-16  6.966e-01   0.000  1.00000    
## actor6:prosoc_left  1.857e+00  6.963e-01   2.667  0.00765 ** 
## actor7:prosoc_left  4.937e+00  9.160e-01   5.390 7.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 197.206  on 27  degrees of freedom
## Residual deviance:  15.002  on 13  degrees of freedom
## AIC: 117.1
## 
## Number of Fisher Scoring iterations: 18
```

To check whether condition plays any significance in this model, a 95% confidence interval for its parameter estimate is created:

```
confint(incl.cond, "condition", level=0.95)
```

```
##       2.5 %     97.5 %
## -0.2822506  0.5589879
```

Since 0 is inside the confidence interval, there is insufficient evidence to conclude that the presence of another chimpanzee has an effect on the choice of the prosocial option.

The final model is the reduced model `no.cond`, which uses the variables `actor` and `prosoc_left` with interaction, and has the lowest AIC of all considered models. As a summary, the prosocial behaviour of chimpanzees is not affected by the presence of another chimpanzee, instead it is evident that the chimpanzees each have different degrees of handedness.