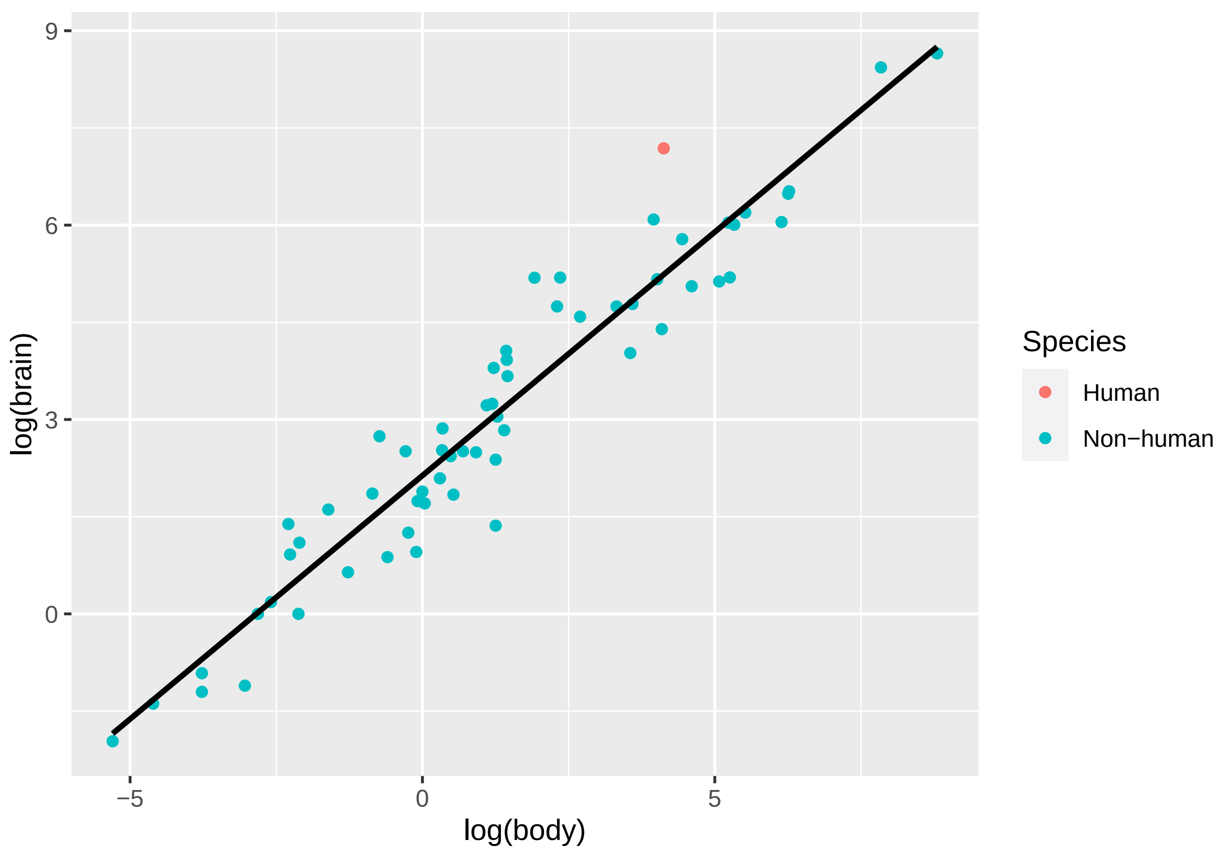
Project 2

olarr@ntnu.no: 10031 ; tizianom@ntnu.no: 10006

Problem 1

a)

```
mammals <- read.table("https://www.math.ntnu.no/~jarlet/statmod/mammals.dat",
    header = T)

Species <- ifelse(1:nrow(mammals) == 32, "Human", "Non-human")

ggplot(mammals, aes(x = log(body), y = log(brain), color = Species)) +
    geom_point() + geom_smooth(formula = y ~ x, method = lm,
    se = F, color = "black")
```



```r
h0 <- glm(log(brain) ~ log(body), data = mammals, family = "gaussian")
h0$coefficients
```

```
## (Intercept)   log(body)
##   2.1347887   0.7516859
```

Here we use the natural log transformation on both body mass and brain size to get a more linear visual representation of their relationship.

b)

```r
# Adding dummy variable to human
mammals$human <- ifelse(mammals$species == "Human", 1, 0)

h1 <- glm(log(brain) ~ log(body) + human, data = mammals, family = "gaussian")
# summary(h1)

h1nonlog <- glm(brain ~ body + human, data = mammals, family = "gaussian")
# summary(h1nonlog)

cat("Human brain size difference is", h1$coefficients[3], "using the log-transformation")
```

```
## Human brain size difference is 2.006907 using the log-transformation
```

```r
cat("Human brain size difference is", h1nonlog$coefficients[3],
    "without the log-transformation")
```

```
## Human brain size difference is 1188.696 without the log-transformation
```

Using the log transformation we see that the difference is approximately 2 when comparing the average human brain size to other species with the same body mass. With no transformation we see that the difference in average human brain size compared to other species is approximately 1189 grams.

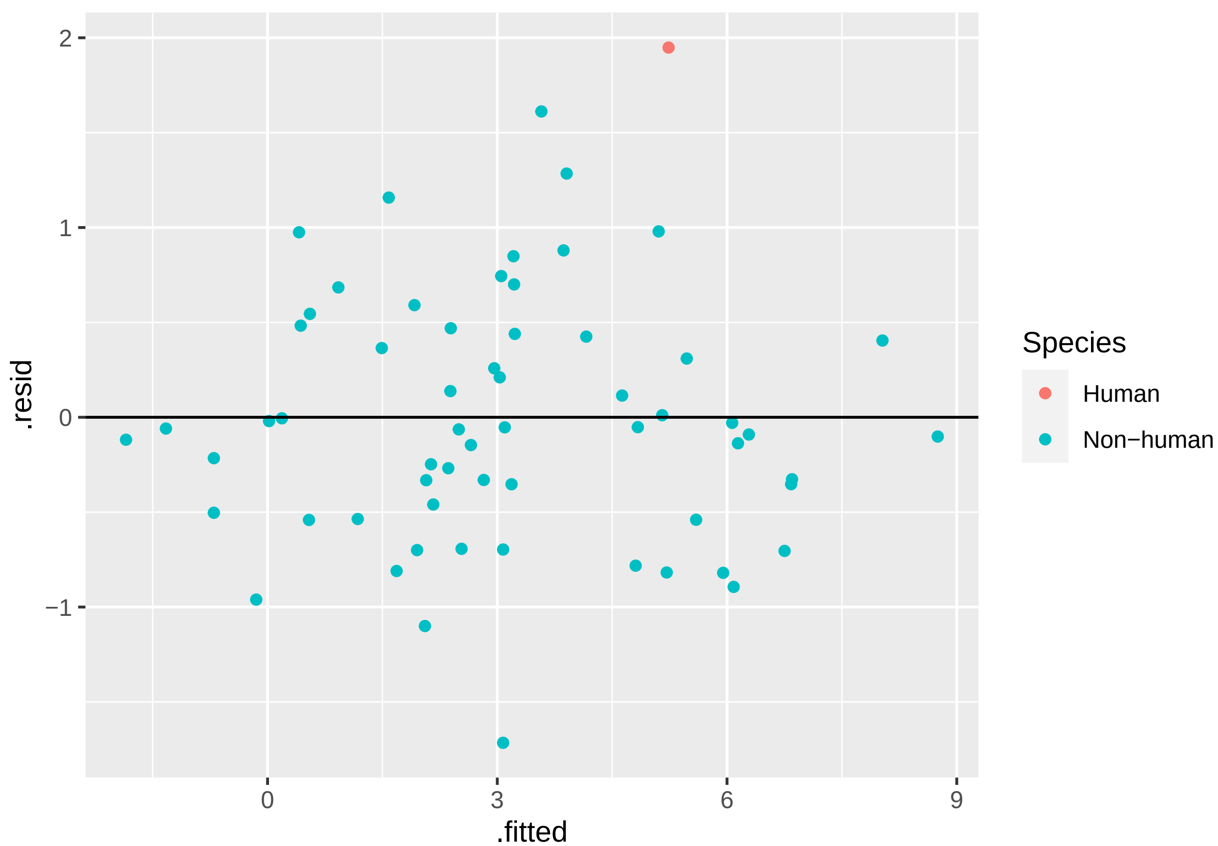
```r
summary(h1)
```

```
##
## Call:
## glm(formula = log(brain) ~ log(body) + human, family = "gaussian",
##     data = mammals)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.68392  -0.46764  -0.02398   0.47237   1.64949
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11500    0.09030  23.421  < 2e-16 ***
## log(body)    0.74228    0.02687  27.622  < 2e-16 ***
## human        2.00691    0.66083   3.037  0.00356 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.4239436)
##
##     Null deviance: 365.111  on 61  degrees of freedom
## Residual deviance:  25.013  on 59  degrees of freedom
## AIC: 127.67
##
## Number of Fisher Scoring iterations: 2
```

Looking at the summary, we see that the human parameter is statistically significant. So human is an outlier.

```r
# Plotting residuals
df <- fortify(h0)

ggplot(df, aes(x = .fitted, y = .resid, color = Species)) + geom_point() +
    geom_hline(yintercept = 0)
```



Looking at the residual plot, we can see that humans are not only an outlier, but it is the biggest outlier in the data.

c)

We know from introductory statistics course that the pivotal quantity is equal to:

$$\frac{y_{n+1}-\hat{\beta}_0-\hat{\beta}_1 x_{n+1}}{s\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}$$

We use this quantity to compute the one-sided (1-α) prediction interval $(-\infty, U)$ for human brain size based on all mammals in the data set except humans.

$$P(\frac{y_{n+1}-\hat{\beta}_0-\hat{\beta}_1 x_{n+1}}{s\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} < t_{\alpha,n-2}) = 1-\alpha$$

Now we can invert the expression inside the parenthesis, solving it for y_{n+1} and finding the required prediction interval

$$y_{n+1}-\hat{\beta}_0-\hat{\beta}_1 x_{n+1} < t_{\alpha,n-2} * s\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$y_{n+1} < \hat{\beta}_0+\hat{\beta}_1 x_{n+1} + t_{\alpha,n-2} * s\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$IC_{y_{n+1}} = (-\infty, \hat{\beta}_0+\hat{\beta}_1 x_{n+1} + t_{\alpha,n-2} * s\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}})$$

We now consider a new model where we included a new parameter β_2 that represents the significance of the human brain (i.e. we multiplied it for a dummy variable z_i equal to 1 if the brain considered is the human one, zero otherwise). Then we compute the profile log-likelihood $l_p(\beta_0, \beta_1) = sup_{\beta_2} l(\beta_0, \beta_1, \beta_2)$ to find the MLE of β_2 given $\hat{\beta}_0$, $\hat{\beta}_1$.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$

$$\hat{\beta}_2 = y_{n+1} - \hat{\beta}_0 - \hat{\beta}_1 x_{n+1},$$

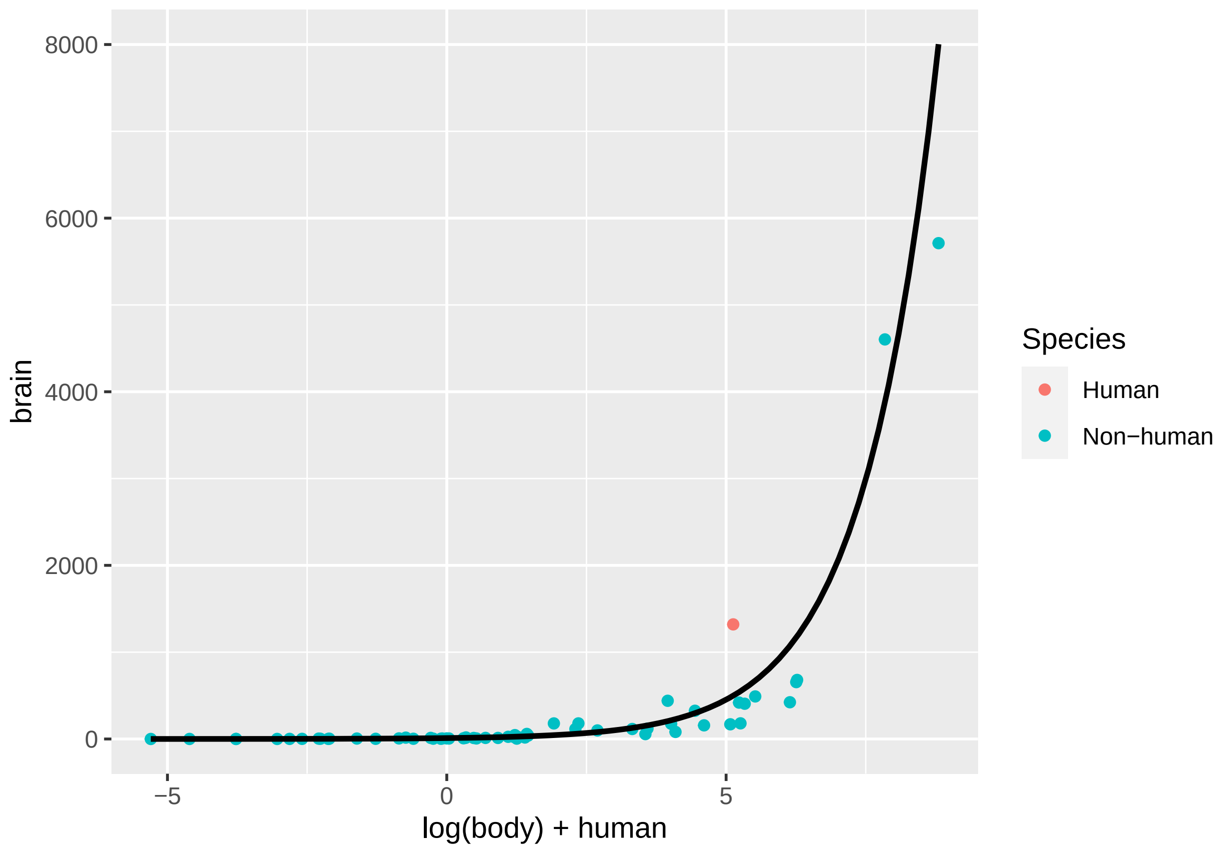
where the subscript in y_{n+1} and x_{n+1} refer to the (n+1)th observation (the human brain size). Applying the "outlier" test involving the test statistic:

$$T = \frac{\hat{\beta}_2}{\sqrt{\widehat{Var\hat{\beta}_2}}}$$

and considering the result for $\hat{\beta}_2$ obtained above we can see the equivalence between the two tests and infer the equivalence of A and B, the events that the $IC_{y_{n+1}}$ does not include the observed human brain size and that the H_0 hypothesis is rejected at significance level α

d)

```r
gglm <- glm(brain ~ log(body) + human, data = mammals, family = Gamma(link = "log"))
# summary(gglm)

ggplot(mammals, aes(x = log(body) + human, y = brain, color = Species)) +
    geom_point() + geom_smooth(formula = y ~ x, method = "glm",
    method.args = list(family = Gamma(link = "log")), se = F,
    color = "black")
```



e)

To test if Kleiber's law applies to this data we test:

$$H_0 : \hat{\beta}_{body} = \frac{3}{4} \quad vs \quad H_1 : \hat{\beta}_{body} \neq \frac{3}{4}$$

```r
h1kleiber <- glm(log(brain) ~ I((3/4) * log(body)) + human, data = mammals,
    family = "gaussian")
# summary(h1kleiber)

lr.test(h1kleiber, h1)
```

```
## $LR
## [1] 0
##
```

```
## $pvalue
## [1] 1
##
## attr(,"class")
## [1] "lrt.test"
```

```
gglmkleiber <- glm(brain ~ human, offset = I((3/4) * log(body)),
    data = mammals, family = Gamma(link = "log"))
# summary(gglmkleiber)

lr.test(gglmkleiber, gglm)
```

```
## $LR
## [1] 0.08078692
##
## $pvalue
## [1] 0.7762338
##
## attr(,"class")
## [1] "lrt.test"
```

For the linear model, we see that there is no reason to reject the null-hypothesis, so this means that it does follow Kleiber's law. Also for the gamma model, we do not reject the null-hypothesis. This indicates that the model does follow Kleiber's law

f)

We need to make the log-likelihoods comparable because the response variables for the gamma glm is not on the log scale.

```
h1nonlogresp <- glm(brain ~ log(body) + human, data = mammals,
    family = "gaussian")

AIC(h1nonlogresp)
```

```
## [1] 1008.526
```

```
AIC(gglm)
```

```
## [1] 523.3768
```

The model with the lowest AIC offer the best fit, so this means that our gamma model is the best model fro our data.

Finding the theoretical skew of the gamma model:

$$
\begin{aligned}
M_{ln(Y)}(t) &= E[e^{t \;\; ln(Y)}] \\
&= E[Y^t] \\
&= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} Y^{t+\alpha} e^{-\lambda y}\, dy \\
&= \frac{\lambda^\alpha \Gamma(t+\alpha)}{\lambda^{t+\alpha}\Gamma(\alpha)} \int_0^\infty \frac{\lambda^{t+\alpha}}{\Gamma(t+\alpha)} Y^{t+\alpha} e^{-\lambda y}\, dy \\
&= \frac{\lambda^\alpha \Gamma(t+\alpha)}{\lambda^{t+\alpha}\Gamma(\alpha)} * 1 \\
&= \lambda^{-t} \frac{\Gamma(t+\alpha)}{\Gamma(\alpha)}
\end{aligned}
$$

$$
\begin{aligned}
K_{ln(Y)}(t) &= ln(M_{ln(Y)}(t)) \\
&= -tln(\lambda) + ln(\Gamma(t+\alpha)) - ln(\Gamma(\alpha)) \\
&= [Derivating \;\; two \;\; times \;\; or \;\; more] \\
&= ln(\Gamma(t+\alpha))
\end{aligned}
$$

$$
\begin{aligned}
Skew(X) &= \frac{E[(X - E[X])^3]}{(Var[X])^{\frac{3}{2}}} \\
&= \frac{k_3}{k_2^{\frac{3}{2}}}
\end{aligned}
$$

$$
\begin{aligned}
k_3 &= K^{(3)}_{ln(Y)}(0) \\
&= \frac{d^3}{dt^3}(ln(\Gamma(t+\alpha)))\Big|_{t=0}
\end{aligned}
$$

$$
\begin{aligned}
k_2 &= K^{(2)}_{ln(Y)}(0) \\
&= \frac{d^2}{dt^2}(ln(\Gamma(t+\alpha)))\Big|_{t=0}
\end{aligned}
$$

The shape parameter α is found by using the dispersion parameter in the summary of "gglm". $\alpha = \frac{1}{dispersion}$.

```
summary(gglm)
```

```
##
## Call:
## glm(formula = brain ~ log(body) + human, family = Gamma(link = "log"),
##     data = mammals)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4464  -0.6099  -0.2276   0.2725   1.8835
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.32733    0.10298  22.601   <2e-16 ***
```

```
## log(body)     0.74193    0.03064  24.212   <2e-16 ***
## human         1.79601    0.75356   2.383   0.0204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.5512612)
##
##     Null deviance: 310.710  on 61  degrees of freedom
## Residual deviance:  25.849  on 59  degrees of freedom
## AIC: 523.38
##
## Number of Fisher Scoring iterations: 5
```

```r
dispersion <- 0.5512612
alpha <- 1/dispersion

k3 <- psigamma(alpha, 2)
k2 <- psigamma(alpha, 1)

cat("Skew is", k3/((k2)^(1.5)))
```

```
## Skew is -0.8244107
```

Sample skew formula:

$$SSkew = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2)^{\frac{3}{2}}}$$

```r
resid <- as.vector(h0$residuals)

men <- mean(resid)

teller <- 0

for (i in 1:62) {
    teller <- teller + (resid[i] - men)^3
}

teller <- (1/62) * teller

nevner <- 0

for (i in 1:62) {
    nevner <- nevner + (resid[i] - men)^2
}

nevner <- ((1/(62 - 1)) * nevner)^(3/2)

cat("Sample Skew from model in a) is", teller/nevner)
```

```
## Sample Skew from model in a) is 0.3957011
```

Problem 2

We want to test if there is an advantage starting as white. Test:

$$\begin{aligned} H_0: &\quad \beta_{1,white} - \beta_{1,black} \leq 0 \\ &\quad vs \\ H_1: &\quad \beta_{1,white} - \beta_{1,black} > 0 \end{aligned}$$

Using the Wald test, we reject the null hypothesis if our Chi square is larger than $X^2_{16,\ 0.05} = 26.296$.

```r
chessdata <- read_excel("Norway Chess 2020_2021.xlsx")

# Alt. hypothesis
mod1 <- vglm(y ~ factor(white) + factor(black), family = cumulative(parallel = T,
    link = "logitlink"), data = chessdata)
# summary(mod1)

# Null hypothesis
mod0 <- vglm(y ~ 1, family = cumulative(parallel = T, link = "logitlink"),
    data = chessdata)
# summary(mod0)

waldtest(mod0, mod1)
```

```
## Wald test
##
## Model 1: y ~ 1
## Model 2: y ~ factor(white) + factor(black)
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    170
## 2    154 16 27.543    0.03582 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that our Chi square is larger than the limit, so we reject the null-hypothesis. This means that overall, there is evidence of an advantage starting as white.

We see that Carlsen, Firouzja and Rapport are strong while playing as white, and that Duda and Tari are strong while playing as black. Using only these players:

```r
summary(mod1)
```

```
##
## Call:
## vglm(formula = y ~ factor(white) + factor(black), family = cumulative(parallel = T,
##     link = "logitlink"), data = chessdata)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                 -3.7050     1.2046  -3.076  0.00210 **
## (Intercept):2                 -1.6902     1.1450  -1.476  0.13992
## factor(white)carlsen           2.8407     1.0182   2.790  0.00527 **
```

```
## factor(white)caruana          1.5106     1.0307   1.466  0.14276
## factor(white)duda             1.1577     1.1217   1.032  0.30201
## factor(white)firouzja         2.5130     0.9859   2.549  0.01081 *
## factor(white)karjakin         1.5831     1.0763   1.471  0.14132
## factor(white)nepomniachtchi   1.7223     1.0581   1.628  0.10357
## factor(white)rapport          2.2371     1.1118   2.012  0.04421 *
## factor(white)tari            -0.3166     0.9607  -0.330  0.74177
## factor(black)carlsen          0.5068     0.9800   0.517  0.60505
## factor(black)caruana          0.8383     1.1120   0.754  0.45095
## factor(black)duda             3.2543     1.2349   2.635  0.00840 **
## factor(black)firouzja         1.3004     0.9981   1.303  0.19262
## factor(black)karjakin         1.6405     1.1359   1.444  0.14869
## factor(black)nepomniachtchi   1.3218     1.0954   1.207  0.22758
## factor(black)rapport          1.7440     1.1656   1.496  0.13460
## factor(black)tari             2.6556     1.0585   2.509  0.01211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 150.6603 on 154 degrees of freedom
##
## Log-likelihood: -75.3302 on 154 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##       factor(white)carlsen        factor(white)caruana
##                  17.128474                    4.529632
##          factor(white)duda       factor(white)firouzja
##                   3.182687                   12.341323
##      factor(white)karjakin factor(white)nepomniachtchi
##                   4.870107                    5.597393
##       factor(white)rapport           factor(white)tari
##                   9.365907                    0.728658
##       factor(black)carlsen        factor(black)caruana
##                   1.659975                    2.312367
##          factor(black)duda       factor(black)firouzja
##                  25.901302                    3.670910
##      factor(black)karjakin factor(black)nepomniachtchi
##                   5.157646                    3.750124
##       factor(black)rapport           factor(black)tari
##                   5.720235                   14.233730
```

```
aronian_white <- as.factor(chessdata$white == "aronian")
carlsen_white <- as.factor(chessdata$white == "carlsen")
caruana_white <- as.factor(chessdata$white == "caruana")
duda_white <- as.factor(chessdata$white == "duda")
firouzja_white <- as.factor(chessdata$white == "firouzja")
karjakin_white <- as.factor(chessdata$white == "karjakin")
nepomniachtchi_white <- as.factor(chessdata$white == "nepomniachtchi")
```

```r
rapport_white <- as.factor(chessdata$white == "rapport")
tari_white <- as.factor(chessdata$white == "tari")

aronian_black <- as.factor(chessdata$black == "aronian")
carlsen_black <- as.factor(chessdata$black == "carlsen")
caruana_black <- as.factor(chessdata$black == "caruana")
duda_black <- as.factor(chessdata$black == "duda")
firouzja_black <- as.factor(chessdata$black == "firouzja")
karjakin_black <- as.factor(chessdata$black == "karjakin")
nepomniachtchi_black <- as.factor(chessdata$black == "nepomniachtchi")
rapport_black <- as.factor(chessdata$black == "rapport")
tari_black <- as.factor(chessdata$black == "tari")

mod11 <- vglm(y ~ carlsen_white + firouzja_white + rapport_white +
    duda_black + tari_black, family = cumulative(parallel = T,
    link = "logitlink"), data = chessdata)
# summary(mod11)

waldtest(mod0, mod11)
```

```
## Wald test
##
## Model 1: y ~ 1
## Model 2: y ~ carlsen_white + firouzja_white + rapport_white + duda_black +
##     tari_black
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    170
## 2    165  5 18.652   0.002231 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
AIC(mod0)
```

```
## [1] 192.0937
```

```r
AIC(mod1)
```

```
## [1] 186.6603
```

```r
AIC(mod11)
```

```
## [1] 178.7269
```

The AIC is lower for this reduced model, so it may fit the data better. Then we have $X^2_{6,\ 0.05} = 12.592$. Our observed Chi square is larger, so there is still evidence that there is an advantage playing as white with the strongest players from each color.

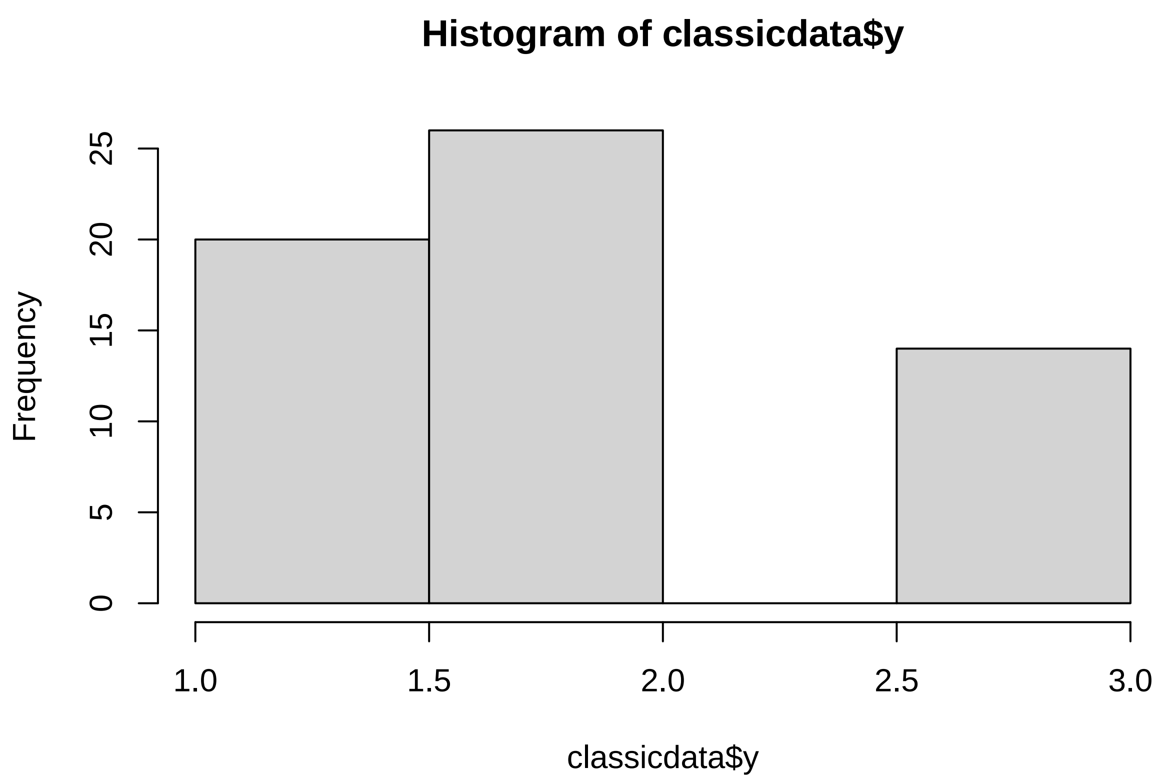
```r
# making only classic data
classicdata <- subset(chessdata, type == "classic")
```

```r
# model from the classic data
classicmod1 <- vglm(y ~ factor(white) + factor(black), family = cumulative(parallel = T,
    link = "logitlink"), data = classicdata)
# summary(classicmod1)

classicmod0 <- vglm(y ~ 1, family = cumulative(parallel = T,
    link = "logitlink"), data = classicdata)
# summary(classicmod0)

waldtest(classicmod0, classicmod1)
```

```
## Wald test
##
## Model 1: y ~ 1
## Model 2: y ~ factor(white) + factor(black)
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    118
## 2    102 16 23.649    0.09745 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using only classic matches, we see that there is no significant advantage to start as white. This looks weird, but can be explained by looking at the histogram

```r
hist(classicdata$y)
```



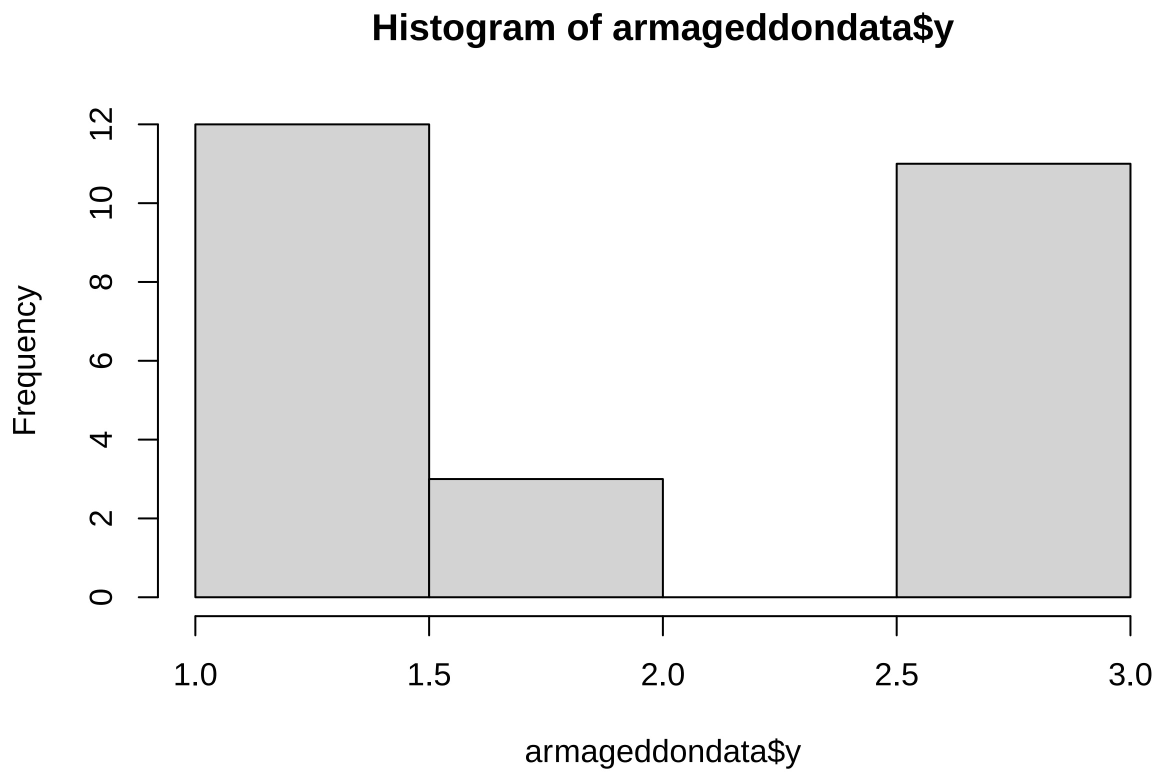
Here we can see that the majority of matches end in a draw.

```r
# making only armageddon data
armageddondata <- subset(chessdata, type == "armageddon")

# model from the armageddon data
armageddonmod1 <- vglm(y ~ factor(white) + factor(black), family = cumulative(parallel = T,
    link = "logitlink"), data = armageddondata)
# summary(armageddonmod1)

armageddonmod0 <- vglm(y ~ 1, family = cumulative(parallel = T,
    link = "logitlink"), data = armageddondata)
# summary(armageddonmod0)

waldtest(armageddonmod0, armageddonmod1)
```

```
## Wald test
##
## Model 1: y ~ 1
## Model 2: y ~ factor(white) + factor(black)
##   Res.Df Df  Chisq Pr(>Chisq)
## 1     50
## 2     34 16 1.9771          1
```

Using only armageddon matches, we see that there is certainly no significant advantage to start as white.

```r
hist(armageddondata$y)
```



Looking at the histogram, one would think there would still be an advantage. We noticed that the vglm function would not give a strength for "Aronian". Also the summary gives us a lot of NAs, but we could not find the reason why.

```
summary(armageddonmod1)
```

```
##
## Call:
## vglm(formula = y ~ factor(white) + factor(black), family = cumulative(parallel = T,
##     link = "logitlink"), data = armageddondata)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                -30.590    126.628  -0.242    0.809
## (Intercept):2                -28.100    126.608      NA       NA
## factor(white)carlsen          47.832    161.140      NA       NA
## factor(white)caruana          28.934    155.448   0.186    0.852
## factor(white)duda             15.308    295.939   0.052    0.959
## factor(white)firouzja         38.675    138.835      NA       NA
## factor(white)karjakin         38.675    138.869   0.278    0.781
## factor(white)nepomniachtchi   37.780    138.853   0.272    0.786
## factor(white)rapport          36.886    138.850   0.266    0.791
## factor(white)tari             20.017    113.102      NA       NA
## factor(black)carlsen          -9.330     57.007  -0.164    0.870
## factor(black)caruana           2.360    193.980      NA       NA
## factor(black)duda             19.338     98.232      NA       NA
```

```
## factor(black)firouzja         11.016    140.517      NA       NA
## factor(black)karjakin        -19.883    119.498      NA       NA
## factor(black)nepomniachtchi   -8.435     57.050  -0.148    0.882
## factor(black)rapport          -7.541     57.047  -0.132    0.895
## factor(black)tari             10.556    124.393      NA       NA
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 12.799 on 34 degrees of freedom
##
## Log-likelihood: -6.3995 on 34 degrees of freedom
##
## Number of Fisher scoring iterations: 18
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', 'factor(white)carlsen', 'factor(white)firouzja', 'factor(white)tari', 'factor(black
##
##
## Exponentiated coefficients:
##       factor(white)carlsen       factor(white)caruana
##               5.931086e+20               3.679690e+12
##          factor(white)duda      factor(white)firouzja
##               4.446835e+06               6.254800e+16
##      factor(white)karjakin factor(white)nepomniachtchi
##               6.254731e+16               2.557516e+16
##       factor(white)rapport          factor(white)tari
##               1.045753e+16               4.933753e+08
##       factor(black)carlsen       factor(black)caruana
##               8.874663e-05               1.059575e+01
##          factor(black)duda      factor(black)firouzja
##               2.503771e+08               6.084397e+04
##      factor(black)karjakin factor(black)nepomniachtchi
##               2.316860e-09               2.170466e-04
##       factor(black)rapport          factor(black)tari
##               5.308166e-04               3.839200e+04
```