# Statistics with Sparrows - many models, matrices, and some magic

Julia Schroeder

17/12/2017

## Linear mixed models – part 2

Before we begin, we clear our workspace. I will never stop writing this. And I hope you will never stop doing this!

```r
rm(list=ls())
setwd("~/Box Sync/Teaching/MagicalStats")
```

We were finally lucky to get permission from the UK Home Office to capture and sample unicorns in the wild. Home Office Offical Dolores Umbridge allowed us to sample them over a period of 20 months. So we spend nearly two years sampling unicorns. It took us a while to figure out how to catch them best, but we finally are very happy to present you with an amazing dataset about these fabulous creatures. We have now not only structural data on bodymass, size, hornlength, but also were able to quantify the amount of glizz! We captured 100 individuals, all of them more than once, and measured all variables on all occations. We were able to score reproductive fitness - apparently unicorns are able to give birth each month, and their litter size varies between 1 and 11. Unfortunately, it was not always possible to score fitness as some females were a bit elusive. However, we were able to extract family ties - while we couldn't tell the exact relationships we could determine which unicorns belonged to which family group. We also did behavioural observations, and counted the amounts of copulations in which the unicorns too part.

Now we have ammassed an amazing dataset and we hope that with this we will be able to shed some light on unicorn biology, as so far nothing much is known about the biology of these elusive creatures. We had a number of hypotheses that we wanted to test:

1) There is sexual dimorphism (trimorphism!) in unicorn body mass, size, and horn length

2) Heavy unicorns, corrected for sex and glizz, have longer horns

We also believe that horns are a signalling quality in unicorns. We are not sure if it's sexually selected, or maybe something very different or odd. Therefore, we think that

3) Unicorns with longer horns are more likely to mate multiply than unicorns with shorter horns and
4) Unicorns with longer horns have a higher fitness

Given that glizz seems to play an important role in unicorn biology, we also postulate that

5) Glizz is an indicator of quality, and more glizz means a unicorn can secure more copulations

6) Therefore, unicorns with more glizz have a higher fitness

7) All of this really suggests that of course, we assume that the more a unicorn mates, the higher its fitness.

As we have multiple datapoints for each individual, we need to correct for that pseudoreplication using a random effect, grouping for individual unicorn. We also have family groups, but we do not know if these have an effect, so we need to first check out if we also should correct for those. To do that we can use LMR test. As there are ten different families, we do not want to use family as a fixed factor as to not loose to many degrees of freedom, so family is also a random effect.

```
d<-read.table("DataForMMs.txt", header=T)
str(d)

## 'data.frame':    2000 obs. of  10 variables:
##  $ Individual   : int  2 2 7 2 2 2 2 2 4 4 ...
##  $ DateOfCapture : int  18 23 22 7 47 80 83 89 80 21 ...
##  $ LitterSize   : int  1 1 1 2 2 2 2 2 2 2 ...
##  $ Size         : num  -1.3 -3.052 -1.537 0.582 -1.404 ...
##  $ Hornlength   : num  1.421 1.53 2.075 1.21 0.322 ...
##  $ Bodymass     : num  -1.891 -2.955 0.348 -0.855 -1.781 ...
##  $ Glizz        : num  -1.3 -3.052 -1.537 0.582 -1.404 ...
##  $ SexualActivity: int  3 4 4 3 2 3 3 3 3 4 ...
##  $ Sex          : Factor w/ 3 levels "female","male",..: 1 1 1 1 1 1 1 1
## 1 1 ...
##  $ Family       : Factor w/ 10 levels "Buckwheat","Fluffytail",..: 4 4 4
## 4 4 4 4 4 4 4 ...

max(d$Individual)

## [1] 100
```
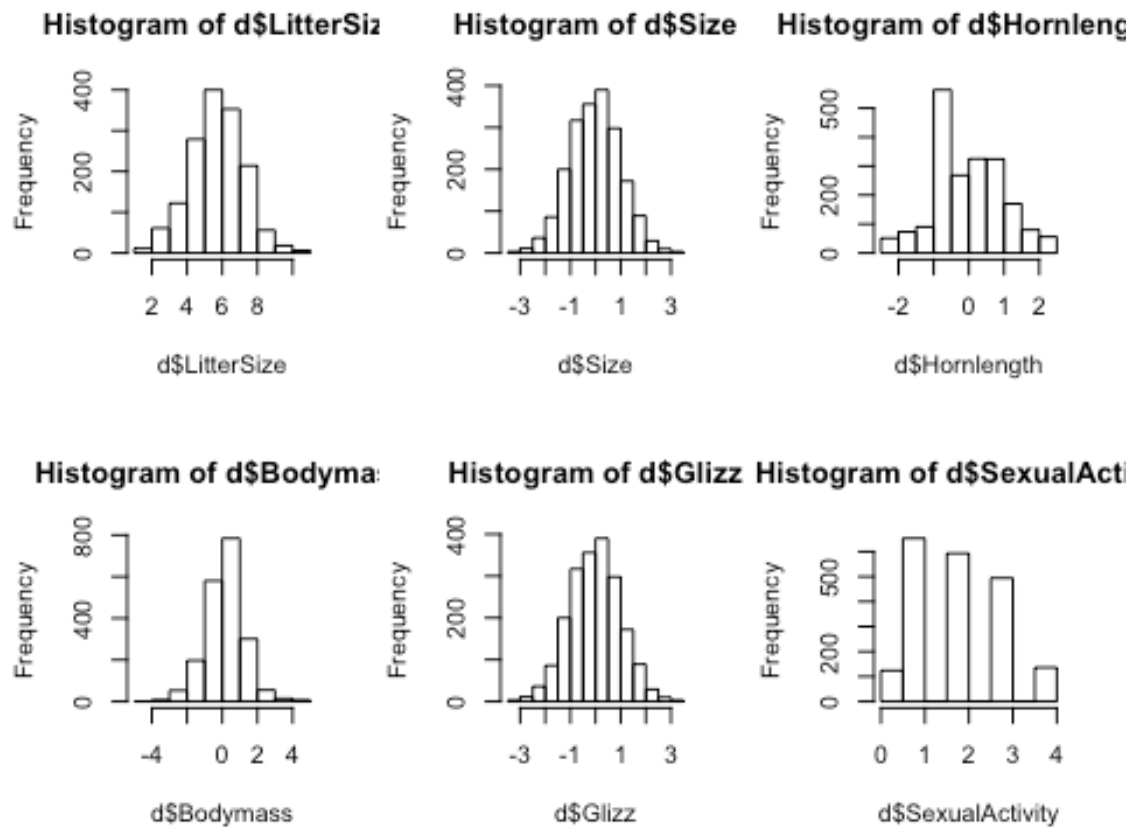
Clearly, while Family is coded as factor already, Individual is coded as a numerical - or continous - variable. So we need to tell R to treat Individual as a factor. We don't have to do that for Family.

```
d$Individual<-as.factor(d$Individual)
names(d)

##  [1] "Individual"    "DateOfCapture"  "LitterSize"     "Size"
##  [5] "Hornlength"    "Bodymass"       "Glizz"          "SexualActivity"
##  [9] "Sex"           "Family"

par(mfrow=c(2,3))
hist(d$LitterSize)
```

```
hist(d$Size)
hist(d$Hornlength)
hist(d$Bodymass)
hist(d$Glizz)
hist(d$SexualActivity)
```


Histogram of d$LitterSiz, Histogram of d$Size, Histogram of d$Hornleng, Histogram of d$Bodyma, Histogram of d$Glizz, Histogram of d$SexualActi
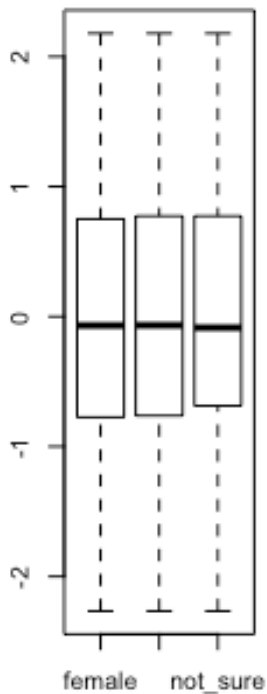
```
dev.off()

## null device
##           1
```

## Hypothesis 1

Ok. It looks like we have a good idea about the variables. The first hypothesis is: 1) There is sexual dimorphism (trimorphism!) in unicorn body mass, size, and horn length. Let's do some plotting:

```
par(mfrow=c(1,3))
boxplot(d$Bodymass~d$Sex)
boxplot(d$Size~d$Sex)
boxplot(d$Size~d$Sex)
```
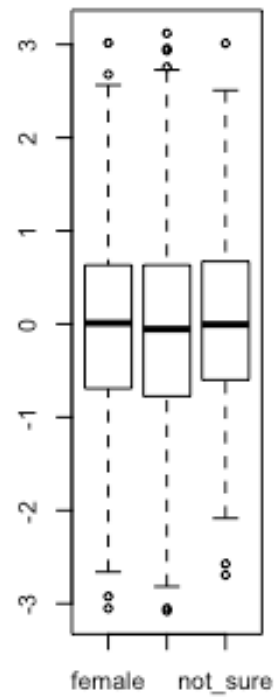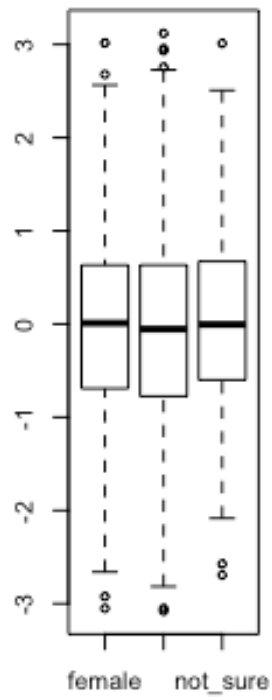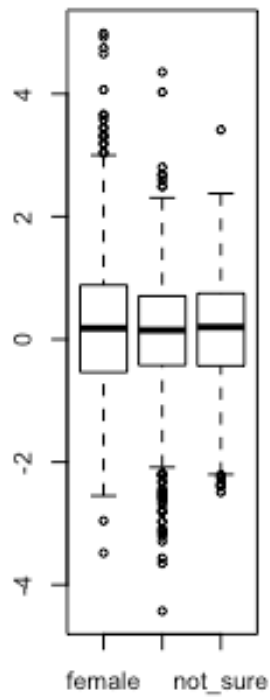
```r
boxplot(d$Hornlength~d$Sex)
```

It doesn't look like there is some. But we need to run a proper analysis, just looking at pretty pictures does not make a statistically convincing argument.

As we know that many individuals have been captured more than once, there is pseudoreplication in the dataset. That means a simple t-test won't cut it (apart from the fact that we have three sexes, and t-tests only do it for two levels). Neither will a simple linear model be sufficient, because of the pseudoreplication, and as we have 100 individuals, we can't use them as fixed factors. We need to run a linear mixed model and put individual as random effect on the intercept to account for pseudoreplication.

To do that, I will introduce the package lme4, and the function lmer.

We have 20 observations per unicorn! And 100 unicorns!

```
## Loading required package: lme4

## Loading required package: Matrix

require(lme4)
#this is a package that we can use for lmms. MCMCglmm does the same, but this
one is ok for easier stuff.

h1m1<-lmer(Bodymass~Sex+(1|Individual), data=d)
summary(h1m1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Bodymass ~ Sex + (1 | Individual)
##    Data: d
##
## REML criterion at convergence: 5294
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.1911 -0.5216  0.0006  0.5050  5.3696
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  Individual (Intercept) 0.4634   0.6807
##  Residual               0.7232   0.8504
## Number of obs: 2000, groups:  Individual, 100
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.20859    0.10655   1.958
## Sexmale     -0.10126    0.15247  -0.664
## Sexnot_sure -0.07435    0.21688  -0.343
##
## Correlation of Fixed Effects:
##             (Intr) Sexmal
## Sexmale     -0.699
## Sexnot_sure -0.491  0.343
```

Ok, the summary statistics confirm that actually, there is no effect of sex on body mass.
Lmer does not provide us with any p-values (there is a big controversy about whether or
not to provide p-calues for REML estimated MMLs, - google *lmer, p-values and all that, doug
bates* and read the blog posts on the stat mailing list from the ethz from Doug Bates).
However, we can look at the parameter estimate and the corresponding se (Sexmale -0.10,
SE 0.15). The SE is actually *larger* than the parameter estimate. Also, the t-value is WAY
below 2. As we know from the beginner's course, that means it is nowhere near statistical
significance. We could now think well, maybe variation within and betwen families does
cloud the picture. Maybe.. so we'll add the family as random factor:

```
h1m2<-lmer(Bodymass~Sex+(1|Individual)+(1|Family), data=d)
summary(h1m2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Bodymass ~ Sex + (1 | Individual) + (1 | Family)
##    Data: d
##
## REML criterion at convergence: 5261
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.1818 -0.5169  0.0012  0.5046  5.3105
##
```

```
## Random effects:
##  Groups      Name          Variance Std.Dev.
##  Individual (Intercept) 0.2615   0.5114
##  Family     (Intercept) 0.2427   0.4927
##  Residual               0.7232   0.8504
## Number of obs: 2000, groups:  Individual, 100; Family, 10
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) 0.1428628  0.2166078   0.660
## Sexmale     0.0003037  0.2517078   0.001
## Sexnot_sure 0.0904081  0.3567560   0.253
##
## Correlation of Fixed Effects:
##             (Intr) Sexmal
## Sexmale     -0.619
## Sexnot_sure -0.508  0.568
```

Oh! It seems like family explains nearly as much variance as individual. However, the SD is also quite large. Let's use a likelihood ratio test to see if adding this extra parameter to the model improves the model:

```
require(lmtest)

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

lrtest(h1m1,h1m2)

## Likelihood ratio test
##
## Model 1: Bodymass ~ Sex + (1 | Individual)
## Model 2: Bodymass ~ Sex + (1 | Individual) + (1 | Family)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -2647.0
## 2   6 -2630.5  1 33.062  8.925e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What the model tells us here is that adding an extra parameter (model 2 in the R output) improves the fit statistically significantly. The LogLikelihood of the reduced model without *family* is -2647, and the LogLikelihood of the model with *family* as random effect is -2630,

thus 17 units higher. The chisquare test supports this, with a super low p-value. Thus, when running body mass, we'd better add family!

Interestingly, however, the summary statistics of the new model (h1m2) also show that a lot of variation is between individuals. However, the standard deviation of the random effects is huge. This indicates that the difference between both, or the ratio, is probably relatively uncertain, so we shouldn't build a big theory on these exact numbers. Yet, we have to include individual because of the pseudo-replication issue - this way we make sure the fixed effect estimate is not biased.

So, a method and result section about hypothesis 1 would read like this:

Methods: To test the hypotheses that body mass of unicorns is sexually trimorphic, we build a linear mixed model with body mass as response variable. We modelled sex as three-level fixed factor (with female as reference). Each unicorn was measured 20 times, and to account for this pseudoreplication, we added individual identity as a random effect on the intercept. Family group explained a large amount of variation and was added as a random effect on the intercept to account for the nested structure of the data.

Results: We found no sexual trimorphism in body mass, as the mean values for body mass did not differ from each other. The mean for females was 0.14±0.27SE (in z-score SD units), and male and sex_notsure did not statistically differ from this. The random factor individual explained 28% of the variation, and the family group structure explaied 26% of the total variation in body mass. We kept family group in the final model because a likelihood-ratio test testing this against a model without family revealed that the reduced model explained variance statisctically significantly less well (chi-square=33.06, df=1 p<0.001).

Then I would present a table with the results from the mixed model:

| Variable | Estimate | Precision | |
|---|---|---|---|
| Fixed effects | b | SE | t value |
| Intercept (female) | 0.14 | 0.22 | 0.66 |
| Sex (male) | 0.00 | 0.25 | 0.00 |
| Sex (not sure) | 0.09 | 0.37 | 0.25 |
| | | | |
| Random effects | Variance | Standard Deviation | |
| Individual | 0.26 | 0.51 | |
| Family | 0.24 | 0.49 | |
| Residual | 0.72 | 0.85 | |

**Table 1: Linear mixed-effects model with body mass of unicorns as response variable.**

You could also calculate the <mark>within-individual repeatability</mark> of body mass, like we did for observer repeatability in the stuffed sparrow dataset. Here, for the sake of clarity, I would not add the Family group, as we are solely interested in whether individuals keep their body mass steady over time.

```
mR<-lmer(Bodymass~1+(1|Individual), data=d)
summary(mR)

## Linear mixed model fit by REML ['lmerMod']
## Formula: Bodymass ~ 1 + (1 | Individual)
##    Data: d
##
## REML criterion at convergence: 5291.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.1967 -0.5193  0.0012  0.5048  5.3650
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  Individual (Intercept) 0.4556   0.6750
##  Residual                0.7232   0.8504
## Number of obs: 2000, groups:  Individual, 100
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.15565    0.07013    2.22

repeatability<-(0.56/(0.46+0.72))
repeatability

## [1] 0.4745763
```

<mark>The repeatability of unicorn body mass is 47%,</mark> we consider that an ok repeatability. Anything <mark>above 50</mark> for a behaviour would be <mark>high</mark>, for morphological traits you'd expect some 70-90%. Anything below 20 would be low for most but the most variable behaviours.

You can now run the same analyses for size and horn length!

## Hypothesis 2

2)   Heavy unicorns, <mark>corrected for</mark> sex and glizz, have longer horns.

Ummm. It seems that we do not need to correct for sex, we found that out when we tested for H1. But what about glizz? And is there any relationship between bodymass and hornlength? As usualy, we first happily plot away:

```
plot(Bodymass~Hornlength,data=d, pch=19, cex=0.5)
```
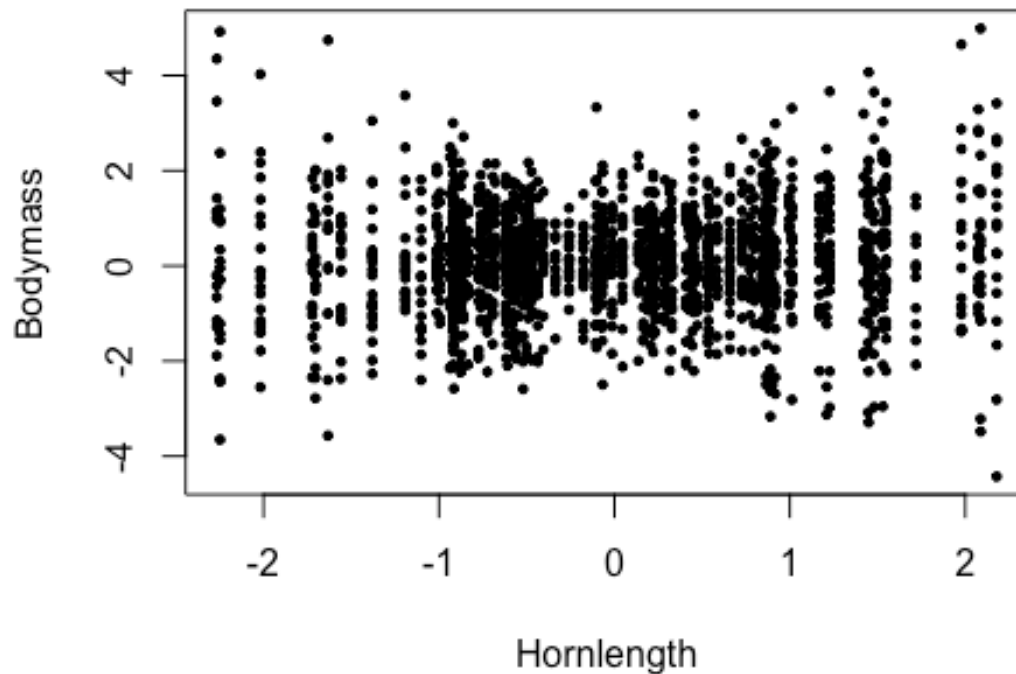
Uhhh.
This looks a bit odd, to be honest. There seems to be a higher variance in body mass in very short horned and very long horned unicorns, compared with averagely horned ones.

```
h2m1<-lmer(Bodymass~Hornlength+Glizz+(1|Individual)+(1|Family), data=d)
summary(h2m1)

## Linear mixed model fit by REML ['lmerMod']
## Formula: Bodymass ~ Hornlength + Glizz + (1 | Individual) + (1 | Family)
##    Data: d
##
## REML criterion at convergence: 5263.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3204 -0.5167 -0.0024  0.5093  5.4691
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  Individual (Intercept) 0.2574   0.5074
##  Family     (Intercept) 0.2169   0.4658
##  Residual               0.7210   0.8491
## Number of obs: 2000, groups:  Individual, 100; Family, 10
##
## Fixed effects:
```

```
##              Estimate Std. Error t value
## (Intercept)  0.15665     0.15694   0.998
## Hornlength   0.05523     0.01972   2.801
## Glizz        0.01519     0.01946   0.781
##
## Correlation of Fixed Effects:
##            (Intr) Hrnlng
## Hornlength 0.001
## Glizz      0.004  0.050
```

```r
h2m2<-lmer(Bodymass~Hornlength+(1|Individual)+(1|Family), data=d)
summary(h2m2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Bodymass ~ Hornlength + (1 | Individual) + (1 | Family)
##    Data: d
##
## REML criterion at convergence: 5257.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3229 -0.5165 -0.0026  0.5064  5.4642
##
## Random effects:
##  Groups     Name        Variance Std.Dev.
##  Individual (Intercept) 0.2577   0.5076
##  Family     (Intercept) 0.2169   0.4657
##  Residual               0.7208   0.8490
## Number of obs: 2000, groups:  Individual, 100; Family, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.15619    0.15692   0.995
## Hornlength   0.05447    0.01969   2.766
##
## Correlation of Fixed Effects:
##            (Intr)
## Hornlength 0.001
```

Ok, there is a somewhat postive association between hornlength and bodymass, but the plot is misleading, tbh. Let's check the residuals out:

```r
plot(h2m2)
```
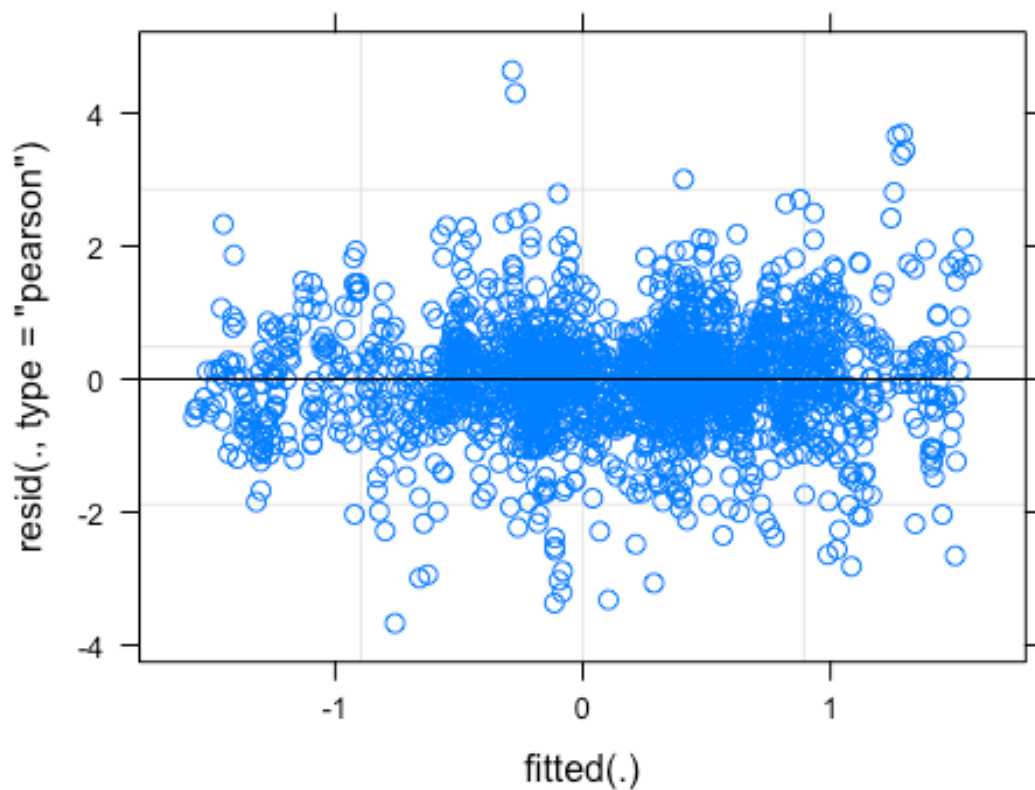
Ok, that doesn't look too bad. Clearly, the difference in variance is captured by the grouping, which is why the residuals look ok-ish. We'll stop here, but in the real world, we'd investiage further to find out why variance is so low for average-horned unicorns.

In a paper, we'd acknowlede that there is a positive association, but we'd mention that the effect size is very, very small with 0.05, which is tiny, and probably not biologically meaningful.
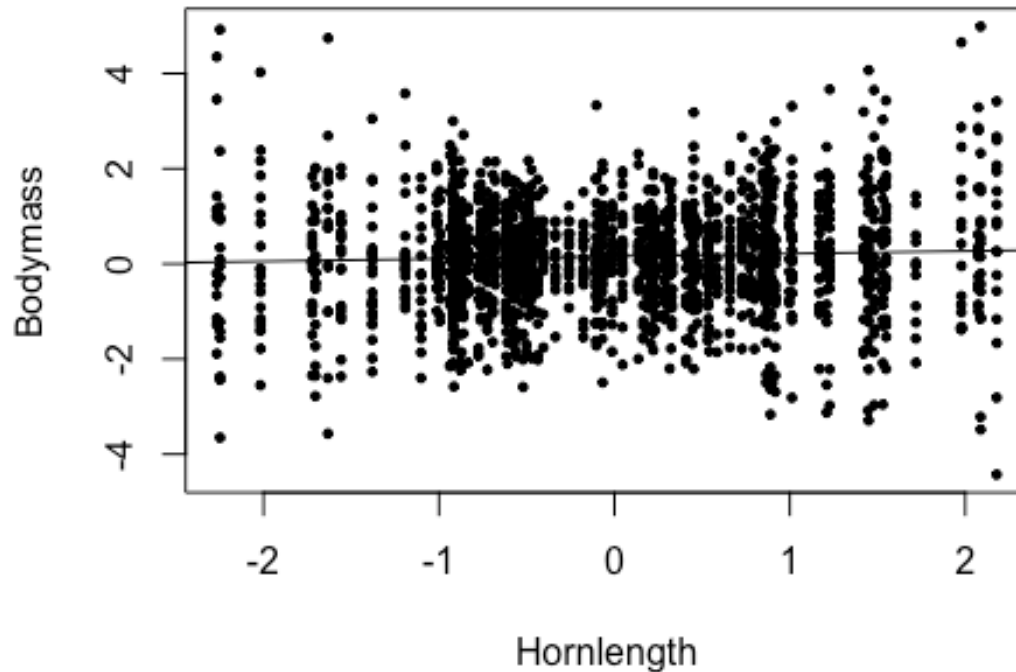
Don't forget to

```
dev.off()

## null device
##           1
```

before you plot again!

```
plot(Bodymass~Hornlength,data=d, pch=19, cex=0.5)
abline(0.16,0.054)
```

Write the methods and results section! Run a model for, and calculate, repeatability.

Go ahead and test the following hypotheses. Do not forget to go through the steps outlined for linear models - you should check all of the steps here, too, especially the collinearity. It often helps to better understand the data when you do that.

## Hypothesis 3

We also believe that horns are a signalling quality in unicorns. We are not sure if it's sexually selected, or maybe something very different or odd. Therefore, we think that 3) Unicorns with longer horns are more likely to mate more often than unicorns with shorter horns, and horn length is repeatable.

## Hypothesis 4

4)    Unicorns with longer horns have a higher fitness

Given that glizz seems to play an important role in unicorn biology, we also postulate that

## Hypothesis 5

5)    Glizz is an indicator of quality, and more glizz means a unicorn can secure more copulations, and glizz is repeatable.

## Hypothesis 6

6)  Therefore, unicorns with more glizz have a higher fitness

## Hypothesis 7

7)  All of this really suggests that of course, we assume that the more a unicorn mates, the higher its fitness.