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

Dr Julia Schroeder

Day 1 Re-visiting what we already know

Before we begin, we clear our workspace. Never forget!

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

```
d<-read.table("SparrowSize.txt", header=TRUE)
str(d)
```

```
## 'data.frame':    1770 obs. of  11 variables:
##  $ BirdID     : int  4401 4401 4405 4405 4405 4409 4409 4409 4409 4409 ...
##  $ Cohort     : int  1991 1991 1994 1994 1994 1994 1994 1994 1994 1994 ...
##  $ CaptureDate: Factor w/ 414 levels "01-Aug-06","01-Dec-07",..: 272 18 25
4 41 88 303 174 18 159 164 ...
##  $ CaptureTime: Factor w/ 293 levels "04:00","04:30",..: NA NA NA NA NA NA
NA NA NA NA ...
##  $ Year       : int  2000 2000 2000 2000 2000 2000 2000 2000 2001 2001 ...
##  $ Tarsus     : num  18.9 18.8 19.1 19 19.1 ...
##  $ Bill       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Wing       : num  82 79 77 78 77 76 76 73 79 77 ...
##  $ Mass       : num  29.4 31.6 29.9 31.6 31 ...
##  $ Sex        : int  1 1 0 0 0 1 1 1 1 1 ...
##  $ Sex.1      : Factor w/ 2 levels "female","male": 2 2 1 1 1 2 2 2 2 2 ..
.
```

```
names(d)
```

```
##  [1] "BirdID"      "Cohort"      "CaptureDate" "CaptureTime" "Year"
##  [6] "Tarsus"      "Bill"        "Wing"        "Mass"        "Sex"
## [11] "Sex.1"
```

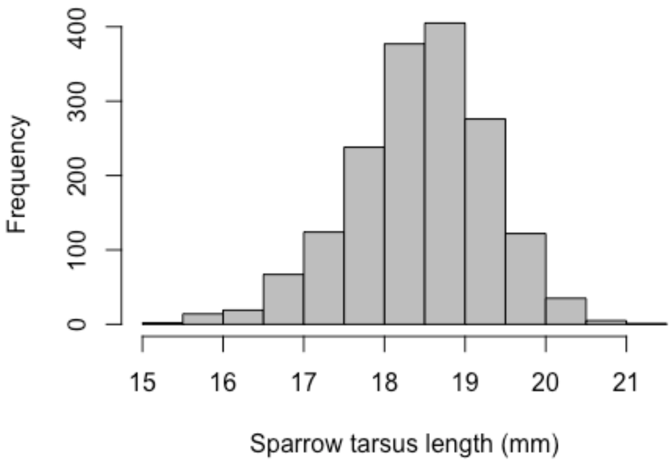
```
head(d)
```

```
##   BirdID Cohort CaptureDate CaptureTime Year Tarsus Bill Wing Mass Sex
## 1   4401   1991   21-Jun-00        <NA> 2000   18.9   NA   82 29.4   1
## 2   4401   1991   02-Oct-00        <NA> 2000   18.8   NA   79 31.6   1
## 3   4405   1994   20-Jun-00        <NA> 2000   19.1   NA   77 29.9   0
## 4   4405   1994   04-Oct-00        <NA> 2000   19.0   NA   78 31.6   0
## 5   4405   1994   07-Oct-00        <NA> 2000   19.1   NA   77 31.0   0
## 6   4409   1994   23-Mar-00        <NA> 2000   18.0   NA   76 28.1   1
##    Sex.1
## 1   male
## 2   male
## 3 female
## 4 female
## 5 female
## 6   male
```

Centrality and spread

Remember, we want to describe distributions. We use simple descriptions in descriptive statistics to do so. The most important descriptors of the centrality are means, and of spread are standard deviation and variance.

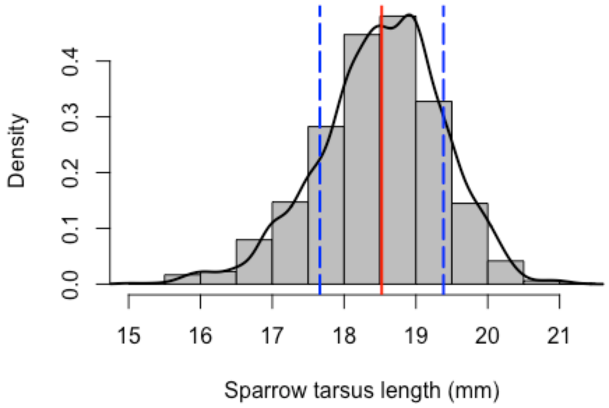
```
hist(d$Tarsus, main="", xlab="Sparrow tarsus length (mm)", col="grey")
```



```
mean(d$Tarsus, na.rm = TRUE)

## [1] 18.52335

var(d$Tarsus, na.rm = TRUE)

## [1] 0.7404059

sd(d$Tarsus, na.rm = TRUE)

## [1] 0.8604684
```

So far, our histograms depict frequencies - that is how often an observation has been seen in the dataset. However, we're now moving to a more professional type of attitude. We know we mostly work with probabilities, for instance, standard deviation tells us the area in which 68.2% of all data points fall would we go out and collect the data again, and again, and again. It's not all about the data we have - it's about what the data we have can tell us about all the potential data that can be collected, about the full population. Therefore now we move on and start plotting densities, like grown ups. Densities are the probablity of the data coming up. We also plot mean and SD:

```
hist(d$Tarsus, main="", xlab="Sparrow tarsus length (mm)", col="grey",
     prob=TRUE) # this argument tells R to plot density instead of frequency,
you can see that on the y-axis
lines(density(d$Tarsus,na.rm=TRUE), # density plot
  lwd = 2)
abline(v = mean(d$Tarsus, na.rm = TRUE), col = "red",lwd = 2)
abline(v = mean(d$Tarsus, na.rm = TRUE)-sd(d$Tarsus, na.rm = TRUE), col = "bl
ue",lwd = 2, lty=5)
abline(v = mean(d$Tarsus, na.rm = TRUE)+sd(d$Tarsus, na.rm = TRUE), col = "bl
ue",lwd = 2, lty=5)
```



Ok, we plotted the probability density curve of sparrow tarsus length. That gives us the likelihood of observation for *any* x (tarsus length). We do that if we don't have observations of *all* tarsus lengths - especially given that tarsus length is an continuous variable. Any point on that curve tells us the probability with which a given sparrow tarsus length can be observed would be measure another sample of sparrows with the same sample size as this one. So, the probability of observing exactly the mean in a sample that size is about 0.5. The red lines gives the mean, and the blue dashed lines the SD.

Interestingly, the density curve has two peaks. We would not expect that in such a large dataset - the expectation would be that things even out. So why didn't they here? One hypothesis is that there is a slight sexual dimorphism in tarsus length, and females have shorter tarsi than males, and that's where the peaks come up. We can test this hypothesis, and check if the means of them are somewhat different.

```
t.test(d$Tarsus~d$Sex)

##
##  Welch Two Sample t-test
##
## data:  d$Tarsus by d$Sex
## t = -3.7382, df = 1677.4, p-value = 0.0001916
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.23814658 -0.07424028
## sample estimates:
## mean in group 0 mean in group 1
##        18.44317        18.59936

par(mfrow=c(2,1))
hist(d$Tarsus[d$Sex==1], main="", xlab="Male sparrow tarsus length (mm)", col
="grey", prob=TRUE)

lines(density(d$Tarsus[d$Sex==1],na.rm=TRUE), lwd = 2)
abline(v = mean(d$Tarsus[d$Sex==1], na.rm = TRUE), col = "red",lwd = 2)
abline(v = mean(d$Tarsus[d$Sex==1], na.rm = TRUE)-sd(d$Tarsus[d$Sex==1], na.r
m = TRUE), col = "blue",lwd = 2, lty=5)
abline(v = mean(d$Tarsus[d$Sex==1], na.rm = TRUE)+sd(d$Tarsus[d$Sex==1], na.r
m = TRUE), col = "blue",lwd = 2, lty=5)

hist(d$Tarsus[d$Sex==0], main="", xlab="Female sparrow tarsus length (mm)", c
ol="grey", prob=TRUE)
lines(density(d$Tarsus[d$Sex==0],na.rm=TRUE), lwd = 2)
abline(v = mean(d$Tarsus[d$Sex==0], na.rm = TRUE), col = "red",lwd = 2)
abline(v = mean(d$Tarsus[d$Sex==0], na.rm = TRUE)-sd(d$Tarsus[d$Sex==0], na.r
m = TRUE), col = "blue",lwd = 2, lty=5)
abline(v = mean(d$Tarsus[d$Sex==0], na.rm = TRUE)+sd(d$Tarsus[d$Sex==0], na.r
m = TRUE), col = "blue",lwd = 2, lty=5)

dev.off()
```

Interesting! I didn't inclue the plot here, but you can look at it when you make it yourself (sometimes I am evil). Yes, there is a difference in tarsus length, but the two-peak thing still shows up in the female data, yet not in the male data. A possible explanation for this might be that when they are not yet fully molted, males can be mistaken for females, because in the juvenile plumage, both sexes look alike. Another option would be that different observers measure birds differently (but it's odd that it's only females, not males). Or that a measurement tool that was used on more females than males is off.

These are only some of several hypotheses that you could test with the data at hand. When you have some time left over, feel free to explore these hypotheses! But always (best in your R code) explicitly state the hypothesis before you run a test so you can make sure the analysis matches the hypothesis! Also, it will be helpful when you revisit the code later one – you will more often than not be impressed about what code you could write in the past that you in the present do not understand any longer.

Now let's move on from all these interesting biological questions back to theoretical boring statistics. It's **variance** time again, yay!

Variance (because you'll never hear enough of it!)

The variance is a very important concept (as if you wouldn't know this by now, but I like to drive a concept home. Really.).

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

The variance is σ^2, or simply as V. It is the square of the sum of the deviations from the mean, divided by sample size minus one. That's also the square of the standard deviation:

```
var(d$Tarsus,na.rm=TRUE)

## [1] 0.7404059

sd(d$Tarsus,na.rm=TRUE)

## [1] 0.8604684

sd(d$Tarsus,na.rm=TRUE)^2

## [1] 0.7404059

sqrt(var(d$Tarsus,na.rm=TRUE))

## [1] 0.8604684
```

Cool. Some serious descriptive stats, totally underrated. Don’t forget to run those, and look at them, and have a good think about these numbers.

The interesting bit about variances is that they have special mathematical rules. They are *additive* - that means, we can partition variances, which is the next best thing to the internet, really, it is. We will do some serious variance partitioning later this week. Here are some cool variance rulez:

1. **If you sum up two independent variables, then the variance of that summed-up variable is the sum of the two variances for both variables separately:**

$$\sigma^2_{tarsus+wing} = \sigma^2_{tarsus} + \sigma^2_{wing}$$

Let's give this a try with wing and tarsus, but let's remove NA's first (I'm too lazy to type all those na.rm=TRUE all the time...:

```
d1<-subset(d, d$Tarsus!="NA")
d1<-subset(d1, d1$Wing!="NA")
sumz<-var(d1$Tarsus)+var(d1$Wing)
test<-var(d1$Tarsus+d1$Wing)
sumz

## [1] 6.576499

test

## [1] 8.172773
```

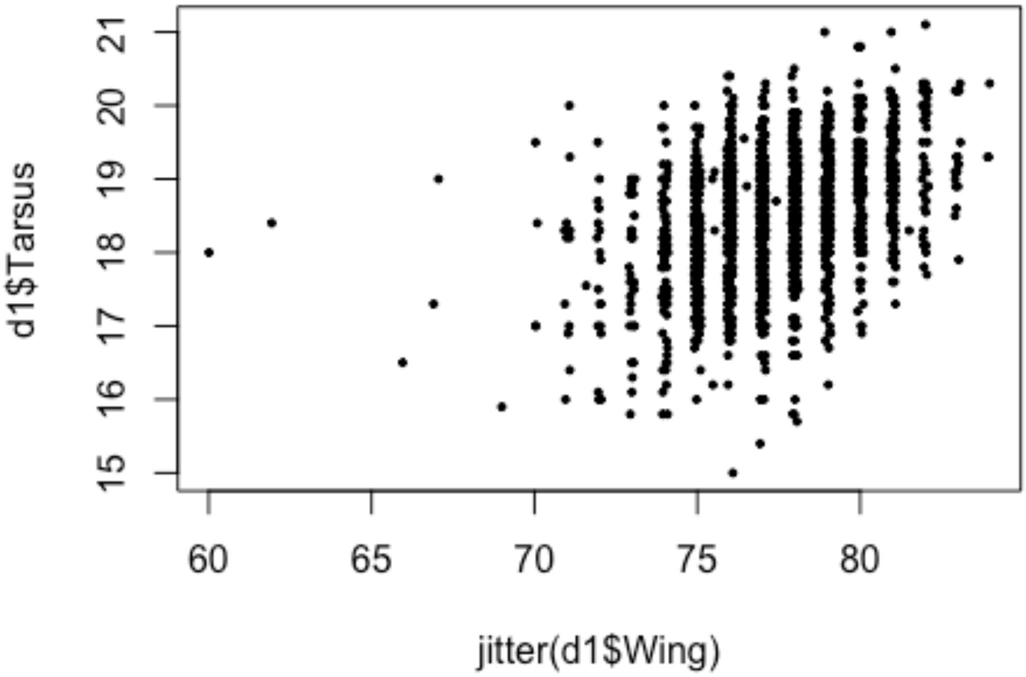
Uuuuups. Not so much. Why doesn't this work? Isn't it frustrating if you get told one thing, and then it doesn't work out in real life. Dang.

Well, but that's not all there is. If you look closely at the rule, then you can see it states that the variables need to be *independent*. Uff. Isn't this the case here? Let's plot wing and tarsus:

```
dev.off()
```

We needed to do this dev.off() thing because we still have the two-panel plot going on. It cancels it and makes it one panel again.

```
plot(jitter(d1$Wing),d1$Tarsus, pch=19, cex=0.4)
```



Errm. Don't look so independent to me. There is a clear relationship between tarsus and wing - the longer the tarsus, the longer the wing, somewhat. So they aren't independent. And clearly, when two variables are not independent, the whole additive rule for variation doesn't work anylonger. The solution is to take the relationship, or *covariation* between both variables also into account. So let's reformulate rule 1 for both, dependent and independent variables.

1. **If you sum up two variables, then the variance of that summed-up variable is the sum of the two variances and *twice their covariance*:**

$$\sigma^2_{tarsus+wing} = \sigma^2_{tarsus} + \sigma^2_{wing} + 2\mathrm{COV}(tarsus, wing)$$

Let's give this a try:

```
cov(d1$Tarsus,d1$Wing)

## [1] 0.798137

sumz<-var(d1$Tarsus)+var(d1$Wing)+2*cov(d1$Tarsus,d1$Wing)
test<-var(d1$Tarsus+d1$Wing)
sumz

## [1] 8.172773

test

## [1] 8.172773
```

Now we're getting somewhere.

So far, it's good to remember this rule. It also holds if both variables are truely independent. Because, if they are, the covariance is zero – that's the definition of independence, after all - and twice zero that is still zero. Huzzah!

Here's rule 2:

2. **When you multiply a variable with a constant its variance equals the variance multiplied with the same constant, but squared:**

$$10^2\sigma^2_{tarsus} = \sigma^2_{tarsus*10}$$

Let's give it a roll:

```
var(d1$Tarsus*10)

## [1] 74.03658

var(d1$Tarsus)*10^2

## [1] 74.03658
```

Cool. One thing that works on the first try.

Now we have learned about covariance, and some cool variance rulez. Commit these to your memory, you will need them often! Also, especially remember the additive rule. A lot of statistical analyses rely on the additive nature of variances, and clearly, if the two variables of interest are not independent, these don't work any longer. That's why we can't have collinear covariates, for instance. If two variables are collinear (that means in stats speak that they are not independent), then we can't use the additive variance rule. That also means that some of our linear models calculations won't be valid, and calculated statistics may not be reliable. And BAM! you have a violation of assumptions. That's why it's important to considere these.

Let's move on the linear models now.

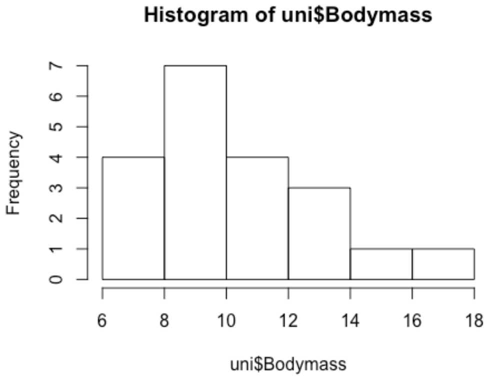
Linear models

Linear models - simple model overview

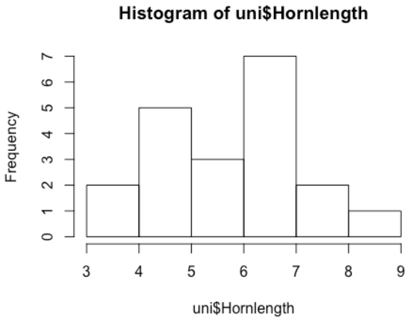
Let's work with a smaller dataset. We'll use unicorns and test the hypothesis that heavier unicorns have larger horns. First we do some descriptive statistics (we can do that now, we're not statistical beginners any longer!), and some happy plotting:

```
uni<-read.table("RUnicorns.txt", header=T)
str(uni)

## 'data.frame':    20 obs. of  8 variables:
##  $ Unicorn    : Factor w/ 20 levels "Alice_Dogface_McDonald",..: 4 7 1 10 1
7 6 5 15 8 18 ...
##  $ Gender     : Factor w/ 3 levels "Female","Male",..: 1 1 1 1 1 1 1 1 2 2
...
##  $ Bodymass   : num  9.5 9.86 10.16 10.37 11.98 ...
##  $ Hornlength: num  6.51 6.19 6.33 5.5 6.34 ...
##  $ Pregnant   : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ Height     : num  3.76 1.56 8.19 1.58 4.29 ...
##  $ Season     : Factor w/ 2 levels "Autumn","Spring": 2 1 2 1 2 1 2 1 1 1 .
..
##  $ Glizz      : int  0 0 0 1 1 1 1 1 0 0 ...

head(uni)

##                      Unicorn Gender  Bodymass Hornlength Pregnant   Height
## 1     Beginda_Friday_McNutt Female  9.500673      6.515        0 3.758080
## 2 Carol_the_Cannon_Richards Female  9.860319      6.190        0 1.558938
## 3    Alice_Dogface_McDonald Female 10.162390      6.330        0 8.190941
## 4               Diane_Gumbo Female 10.365228      5.505        0 1.584386
## 5      Ratline_Slinger_Rose Female 11.983053      6.340        0 4.287208
## 6  Betty_Striker_Boot_Rogue Female 13.199578      6.110        0 1.084253
##   Season Glizz
## 1 Spring     0
## 2 Autumn     0
## 3 Spring     0
## 4 Autumn     1
## 5 Spring     1
## 6 Autumn     1

mean(uni$Bodymass)

## [1] 10.39162

sd(uni$Bodymass)

## [1] 2.786788
```

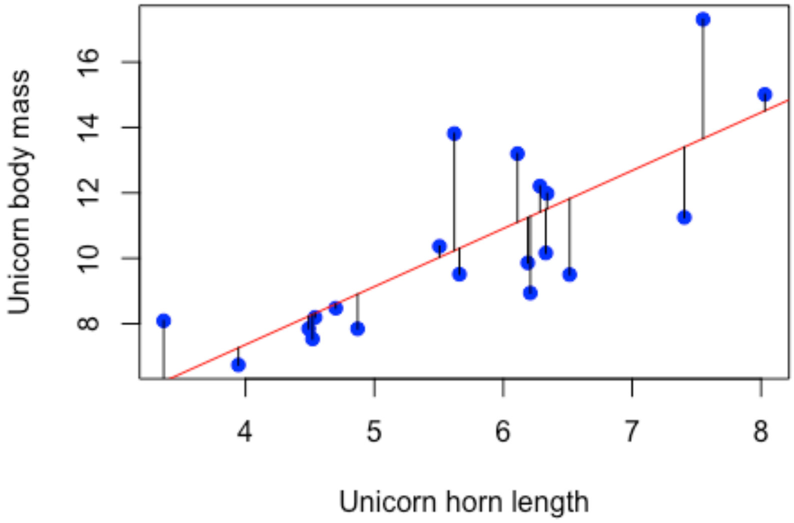
```
var(uni$Bodymass)

## [1] 7.766185

hist(uni$Bodymass)
```



```
mean(uni$Hornlength)

## [1] 5.709

sd(uni$Hornlength)

## [1] 1.229192

var(uni$Hornlength)

## [1] 1.510912

hist(uni$Hornlength)
```



```
plot(uni$Bodymass~uni$Hornlength, pch=19, xlab="Unicorn horn length", ylab="U
nicorn body mass", col="blue")
mod<-lm(uni$Bodymass~uni$Hornlength)
abline(mod, col="red")
res <- signif(residuals(mod), 5)
pre <- predict(mod)
segments(uni$Hornlength, uni$Bodymass, uni$Hornlength, pre, col="black")
```



That's a beautiful plot with a nice relationship between body mass and horn length. From this plot, we would conclude that unicorns with larger horns are fatter, and our hypothesis - that long horns make them dominant so they get preferred access to food, seems to be supported. We can also see that some residuals are larger than others. They seem to be clustered towards the unicorns with larger horns, but it might just seem so. We make a mental note for later, when we check residual plots.

However, are longer-horned unicorns really fatter? Maybe they are simply larger, and that's why they are heavier? Or, they are pregnant, and unicorns with long horns are super attractive and fertile and that's why long-horned unicorns are more likely pregnant, and heavier. Or, unicorns with longer horns have more glizzer and jewels because the longer the horn, the more decorations fit on. And glizz is heavy. Or, horns fall off in spring, and grow back in autumn, when there is also lots to eat, so the heavier unicorns with long horns were all caught in autumn. Or, or, or... the possibilities are endless! Clearly, this is not a good way to test our hypothesis! Really, we want an experiment, where we keep unicorns in captivity, singly, and then put them together and see if the ones with longer horns gain body mass faster. Then we could just as well measure dominance directly. Well, unicorns are very secretive creatures, and the UK home office is very strict and won't allow us to keep them in captivity, so observational data from the wild is all we have. What do we do?

Statistics to the rescue. While statistics can NEVER resolve the quesion of causality (*note: correlation does not prove causality!*), we can do a bit more than a simple correlation. We can add covariates and fixed factors to account for all the potential covarying variables. Then we'll test which variable explains most variation. Let's give this a go. We'll follow our checklist for GLMs that we learned in the lecture today, and hopefully, remember from last year:

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision
11. Model validation
12. Interpretation

We'll plot happily away (please look at the plots in your own computer):

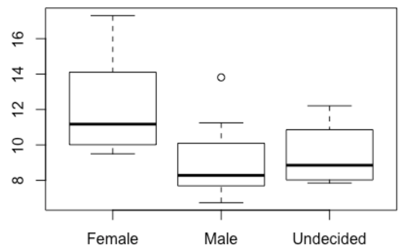
```
hist(uni$Bodymass)

hist(uni$Hornlength)

hist(uni$Height)
```

Some not-so-normal, unruly data, but overall, it's ok-ish. We're not too happy with it, but we'll check the model validation at the end.
No zero-inflation. What about collinearity?

```
cor.test(uni$Hornlength,uni$Height)

## 
##  Pearson's product-moment correlation
## 
## data:  uni$Hornlength and uni$Height
## t = -0.3232, df = 18, p-value = 0.7503
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5016186  0.3793118
## sample estimates:
##        cor 
## -0.0759589
```

Not much, so we can live with that, they seem to be independent. Happy plotting:

```
boxplot(uni$Bodymass~uni$Gender)
```

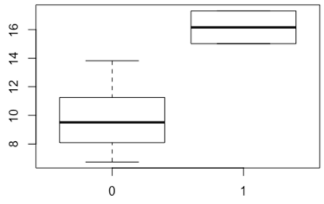


There seems to be something going on. Females are heavier, but also have larger variance. Might be to do with pregnancy? Good that we picked up on that!

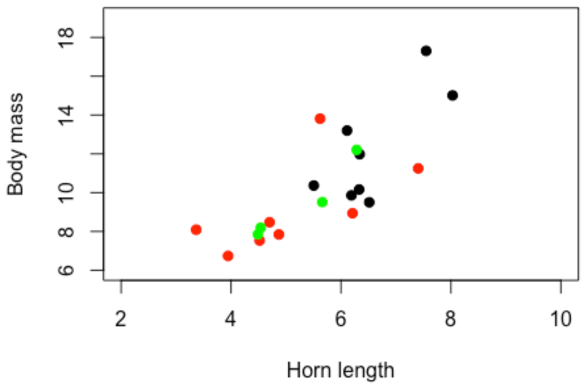
```
par(mfrow=c(2,1))
boxplot(uni$Bodymass~uni$Pregnant)

plot(uni$Hornlength[uni$Pregnant==0],uni$Bodymass[uni$Pregnant==0], pch=19, x
lab="Horn length", ylab="Body mass", xlim=c(2,10), ylim=c(6,19))
points(uni$Hornlength[uni$Pregnant==1],uni$Bodymass[uni$Pregnant==1], pch=19,
col="red")
```

Uhhhh. Clearly, we have some things to account for if we want to better understand why bodymass varies so much in Unicorns...

```
dev.off()

boxplot(uni$Bodymass~uni$Pregnant)
```

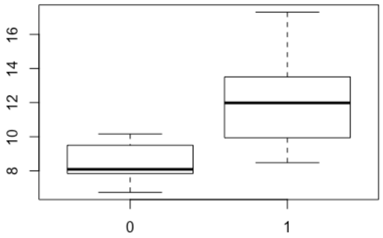


```
plot(uni$Hornlength[uni$Gender=="Female"],uni$Bodymass[uni$Gender=="Female"],
pch=19, xlab="Horn length", ylab="Body mass", xlim=c(2,10), ylim=c(6,19))
points(uni$Hornlength[uni$Gender=="Male"],uni$Bodymass[uni$Gender=="Male"],pc
h=19, col="red")
points(uni$Hornlength[uni$Gender=="Undecided"],uni$Bodymass[uni$Gender=="Unde
cided"],pch=19, col="green")
```

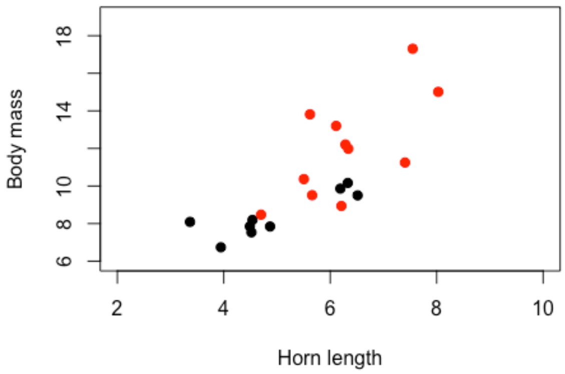


Ok. It looks like there is a sex effect in body mass and horn length, with males being heavier, and having longer horns than females, and indecisive unicorns. This is getting complicated.

```
boxplot(uni$Bodymass~uni$Glizz)
```



```
plot(uni$Hornlength[uni$Glizz==0],uni$Bodymass[uni$Glizz==0], pch=19, xlab="H
orn length", ylab="Body mass", xlim=c(2,10), ylim=c(6,19))
points(uni$Hornlength[uni$Glizz==1],uni$Bodymass[uni$Glizz==1], pch=19, col="
red")
```



Dang. Ok. Now we've used red dots to indicate unicorns with glizz. Unicorns with glizz are heavier and have longer horns than those without. So, now we know, we want gender in, and pregnant or not, and glizz. Height is difficult. We'll run it, but we don't think it's important, so we are prepared to throw it out. So our full model is this:

```
FullModel<-lm(uni$Bodymass~uni$Hornlength+uni$Gender+uni$Pregnant+uni$Glizz)
summary(FullModel)

## 
## Call:
## lm(formula = uni$Bodymass ~ uni$Hornlength + uni$Gender + uni$Pregnant + 
##     uni$Glizz)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.8041 -0.8176 -0.0876  0.4675  3.3779 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)  
## (Intercept)           6.5761     2.5909   2.538   0.0237 *
## uni$Hornlength        0.5189     0.4423   1.173   0.2603  
## uni$GenderMale       -1.1966     0.9011  -1.328   0.2055  
## uni$GenderUndecided  -0.9260     0.9984  -0.928   0.3694  
## uni$Pregnant          3.3983     1.2984   2.617   0.0203 *
## uni$Glizz             2.1406     0.8011   2.672   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.412 on 14 degrees of freedom
## Multiple R-squared:  0.8108, Adjusted R-squared:  0.7433 
## F-statistic:    12 on 5 and 14 DF,  p-value: 0.0001176
```

Hmm. Difficult. It seems that yes, pregnancy and glizz are important predictors for unicorn body mass. But why did Horn length not come out in the summary statistics? Uhh. And why didn't Gender show up significant? This is clearly confusing. Let's think this through. Only female unicorns can get pregnant (or so we think). That means, the pregant factor is not very useful for any other than female unicorns. Also, only two unicorns are pregnant. Maybe we're better off excluding those? Let's see:

```
u1<-subset(uni, uni$Pregnant==0)
FullModel<-lm(u1$Bodymass~u1$Hornlength+u1$Gender+u1$Glizz)
summary(FullModel)

##
## Call:
## lm(formula = u1$Bodymass ~ u1$Hornlength + u1$Gender + u1$Glizz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8421 -0.6139 -0.0973  0.2461  3.3757
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.2330     2.5395   2.454   0.0290 *
## u1$Hornlength          0.5795     0.4337   1.336   0.2044
## u1$GenderMale         -1.1308     0.8800  -1.285   0.2212
## u1$GenderUndecided    -0.8702     0.9743  -0.893   0.3880
## u1$Glizz               2.0795     0.7824   2.658   0.0197 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.377 on 13 degrees of freedom
## Multiple R-squared:  0.6533, Adjusted R-squared:  0.5466
## F-statistic: 6.123 on 4 and 13 DF,  p-value: 0.005368
```

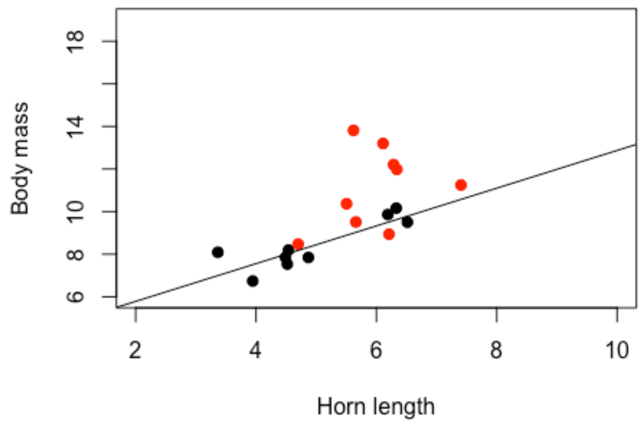
Ok. according to this model, we don't need gender to explain differences in body mass. Ok.

```
ReducedModel<-lm(u1$Bodymass~u1$Hornlength+u1$Glizz)
summary(ReducedModel)

##
## Call:
## lm(formula = u1$Bodymass ~ u1$Hornlength + u1$Glizz)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3470 -0.6972 -0.2123  0.5723  3.0519
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.0106     1.8420   2.177   0.0458 *
## u1$Hornlength     0.8864     0.3589   2.470   0.0260 *
## u1$Glizz          1.7702     0.7369   2.402   0.0297 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.362 on 15 degrees of freedom
## Multiple R-squared:  0.6085, Adjusted R-squared:  0.5563
## F-statistic: 11.66 on 2 and 15 DF,  p-value: 0.0008827
```

```
plot(u1$Hornlength[u1$Glizz==0],u1$Bodymass[u1$Glizz==0], pch=19, xlab="Horn
length", ylab="Body mass", xlim=c(2,10), ylim=c(6,19))
points(u1$Hornlength[u1$Glizz==1],u1$Bodymass[u1$Glizz==1], pch=19, col="red"
)
abline(ReducedModel)
```

```
## Warning in abline(ReducedModel): only using the first two of 3 regression
## coefficients
```

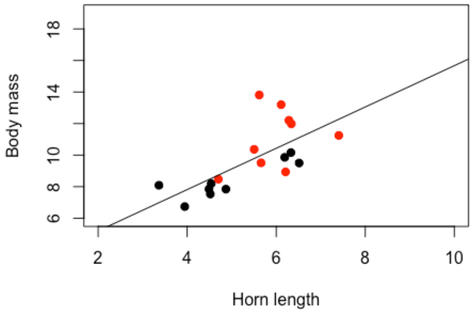


Um... looking at the warning message: The abline is only plotted for the first two of three regression coefficients. That means, it's plotted only for horn length, but that estimate already takes Glizz into account! It makes sense, it's tough to plot a line that takes into account both. Make a mental note for this, as it usually plots the first variable if you have multiple variables in your model. Maybe it would be a better plot if we'd plot it without glizz?

```
ModForPlot<-lm(u1$Bodymass~u1$Hornlength)
summary(ModForPlot)
```

```
##
## Call:
## lm(formula = u1$Bodymass ~ u1$Hornlength)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7718 -0.9944 -0.5421  0.9718  3.8768
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.5774     1.9856   1.298  0.21268
## u1$Hornlength   1.3096     0.3563   3.676  0.00204 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.552 on 16 degrees of freedom
## Multiple R-squared:  0.4578, Adjusted R-squared:  0.4239
## F-statistic: 13.51 on 1 and 16 DF,  p-value: 0.002045
```

```
plot(u1$Hornlength[u1$Glizz==0],u1$Bodymass[u1$Glizz==0], pch=19, xlab="Horn
length", ylab="Body mass", xlim=c(2,10), ylim=c(6,19))
points(u1$Hornlength[u1$Glizz==1],u1$Bodymass[u1$Glizz==1], pch=19, col="red"
)
abline(ModForPlot)
```

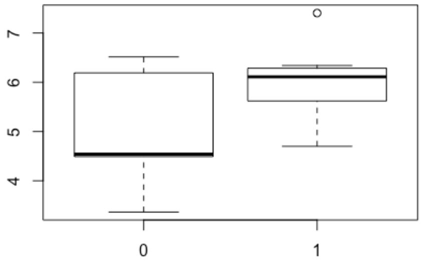


That looks indeed more true, but it does not reflect the biological effects. Also, when you look at the summary statistics for "ReducedModel" and "ModForPlot", you can see a difference in the parameter estimate: In the model we used for plotting the parameter estimate is 0.4 larger than the proper reduced model.

But what does it all mean?

The interpretation is as follows: Some of the variation in bodymass is caused by pregnancy - clearly, unicorns that are pregnant carry with them a lot of weight, and that's picked up by our model. We've excluded them because we are not interested in unicorn reproduction (at this point in time). Then, some of the variation in body mass is caused by some unicorns being really into decorating themselves with glizz. We can account for that statistically. It turns out, quite a lot of variation is explained by glizz. If we look at R square of the ModForPlot, it is 0.46. Roughly explained, this means 46% of the variation is explained by horn length. Now, if we look at R square of the ReducedMod, it shows that in this model, about 61% of the variation is explained by both, horn length and glizz. Thus, we could assume that glizz explaines quite some variance in body mass. However, could we assume it explaines 61-46=15% of the variance? Well, if Horn length and glizz are independent, we could, because of the additive rule for variances (isn't it cool we understand this now?). So, let's check this:

```
boxplot(u1$Hornlength~u1$Glizz)
```



```
t.test(u1$Hornlength~u1$Glizz)

##
##  Welch Two Sample t-test
##
## data:  u1$Hornlength by u1$Glizz
## t = -2.2536, df = 13.924, p-value = 0.04087
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.9673961 -0.0481595
## sample estimates:
## mean in group 0 mean in group 1
##        4.973889        5.981667
```

While it seems that unicorns with longer horns are more likely to wear glizz, it's not the cleanest relationship... It seems that the variances differ between the groups, which is ok. But, hey, the difference is statistically significant, that's something!

Ok, so both variables are not fully independent, so we cannot simply assume that the difference in R squares is also the difference in variance explained by each variable. However, we are confident now to say that unicorns with longer horns are heaviers. Also, unicorns body mass is affected by glizz, where if they wear glizz they are heavier. There is also some collinearity between horn lenght and glizz, such that unicorns with longer horns have more glizz, or maybe, unicorns with more glizz grow longer horns. It's unclear. Clearly, we want to explore the glizz stuff a bit more closely. One might also come up with the hypothesis that more glizz is heavier than just a bit of glizz. Maybe wearing heavy glizz makes them grow stronger, and heavier! Lots of interesting hypotheses for future research! If you're fast with this, happy analyzing of this data – there are lots of effects waiting to be found!

Now back from exciting unicorn biology to boring statistics. We should at least check the model for violations of assumptions. From looking at plots and summary statistics we already know that some assumptions have been slightly violated - the assumption that horn length and glizz are independent. However, we are somewhat confident that we've interpreted this correctly, and are a bit cautious with the exact numbers. Let's see what else we'll find out.

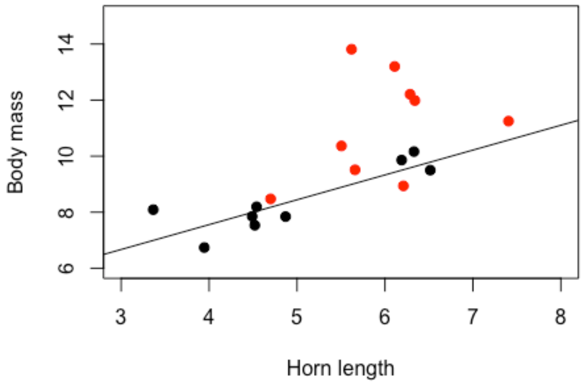
```
plot(ReducedModel)
```

Points 14 is quite the odd one out. As are 6 and 12. Let's see who they are:

```
View(u1)
```

There seems to be nothing unusual with *Betty Striker Boot Rogue*, *Ian the Daggar*, and *Ambrose Buoy Christopher*. So we don't exclude them, we just assume this is due to natural variation. Also, we have a sample size on the lower end of things, so that means there will be some variability. It means that we should be somewhat cautious with our inferences, and maybe not use the exact numbers we got for prediction. In a paper or thesis, we'd probably mention that our sample size and thus statistical power is a bit on the low side, and thus the estimates should be considered with caution, although, on the whole, they are ok. We'd still say that we interpret these results as good enough to conclude that both, glizz and horn length are positively associated with body mass. We'd also report the association between glizz and horn length. We'd provide the plot where we indicate glizz with colored dots, and the actual regression line from our reduced model (not the beautiful one).

```
plot(u1$Hornlength[u1$Glizz==0],u1$Bodymass[u1$Glizz==0], pch=19, xlab="Horn
length", ylab="Body mass", xlim=c(3,8), ylim=c(6,15))
points(u1$Hornlength[u1$Glizz==1],u1$Bodymass[u1$Glizz==1], pch=19, col="red"
)
abline(ReducedModel)

## Warning in abline(ReducedModel): only using the first two of 3 regression
## coefficients
```



We'd clearly explain in the plot legend what the red and what the black dots are, and where the regression line comes from - that is from the model that we'd also present in a table, and that the regression line is from the estimate for horn length from this table. We'd then, in the discussion, outline that glizz seems to introduce quite some variation in body mass, and that representing it as simple "have glizz or not" a bimodal category is probably not representing the true nature of this trait. Apart from the fact that unicorns with longer horns are heavier, we'd also discuss that it might be useful to somehow empirically estimate the mass of the glizz, or measure it in any other continuous way. This is important to better account for it when we want to know the relationship between body mass and horn length. We'd also discuss that there also seems to be an association between horn length and glizz, and that this might indicate some sort of interesting phenomenon that needs to investigated further.

Linear models - interpretation of interactions - two-level fixed factor and continuous variable

Now we will continue practicing our skills interpreting the results from a linear model. The new dataset is creatively names "data.txt". It contains fictional data about species richness of arthropods in grasslands. Some of these grasslands have super high diversity, up to 60 species, while others are very low, with only one species. Half of the sampled grasslands were farmed with conventional measures, while the other half was farmed following standards for an "organic" certificate. The question was to determine the effect of fertilizer on species richness. To test the hypothesis that increasing amounts of fertilizer lead to lower species diversity, varying amounts of fertilizer have been applied on both types of grasslands. We will test this hypothesis.

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

dat<-read.table("data.txt", header=TRUE)
head(dat)

##   species_richness fertilizer       method
## 1                1      48.20 conventional
## 2                5      43.80 conventional
## 3                8      44.23 conventional
## 4                8      51.14 conventional
## 5                9      37.89 conventional
## 6                9      47.47 conventional

str(dat)

## 'data.frame':    100 obs. of  3 variables:
##  $ species_richness: int  1 5 8 8 9 9 9 10 12 13 ...
##  $ fertilizer      : num  48.2 43.8 44.2 51.1 37.9 ...
##  $ method          : Factor w/ 2 levels "conventional",..: 1 1 1 1 1 1 1 1
1 1 ...
```

Now, because we know that organic grassland likely has higher species diversity, we add it as a fixed factor to the model. Also, we assume that the amount of fertilizer affects conventional grassland and organic grassland differently, - specifically, we expect that fertilizers might affect organic grassland less (at least in the first year, which is what is measured here) than conventional ones. *I just made this up. I assumed you could tell in the previous example.* Therefore, we add an interaction to the model.

```
fullmodel<-(lm(species_richness~fertilizer*method,data=dat))
summary(fullmodel)

##
## Call:
## lm(formula = species_richness ~ fertilizer * method, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0076  -5.9471  -0.6839   5.1785  18.7971
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                46.25725    2.03928  22.683  < 2e-16 ***
## fertilizer                 -0.77690    0.06946 -11.185  < 2e-16 ***
## methodorganic               1.19196    2.87430   0.415    0.679
## fertilizer:methodorganic    0.80883    0.09800   8.253 8.26e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.14 on 96 degrees of freedom
## Multiple R-squared:  0.7889, Adjusted R-squared:  0.7823
## F-statistic: 119.6 on 3 and 96 DF,  p-value: < 2.2e-16
```

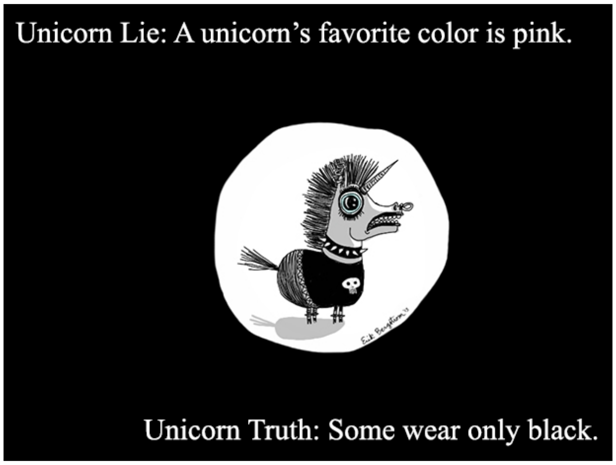
Wow. The interaction is indeed statistically significant. Now, let's try to figure out what it means. First, we can interpret the parameter estimate for the fertilizer in conventional grassland. We can interpret it directly because the reference level of the fixed factor (method) is conventional (the level that is *not* mentioned here). So, per unit fertilizer more applied, we find 0.78 fewer species in conventionally managed grassland. That was easy. Now on to the other terms. On average, there are 1.19 more species in organic grassland (parameter estimate of the fixed factor for the level organic). The interaction tells us that in organic grassland, per unit fertilizer, there is 0.81 more species, *in addition* to the other effect. Now, wait, what does that mean, in addition? Let's examine the formula and plug in values:

$$y_i = b_{intercept} + b_{fertlizer} x_i + b_{method} x_i + b_{method} b_{fertlizer} x_i + \varepsilon_i$$

Ok, it's clear what we plug in for fertilizer - the values in the column of the fertilizer. But what do we plug in for method? We can hardly calculate with words. What happens internally when you run a model with a categorical variable is that it gets recoded. Remember how we ran models with both, sex coded as female and male, and as 0 and 1? R does that for you, automatically, without you even noticing. R recodes the level that comes first in alphanumerical order as zero, and the next as 1. Thus, in our example, conventional gets a 0, and organic a 1. Then plugging in shows that when we calculate expected values for conventional methods, we only have to calculate the first to parts of the equation, simply because $b_{method} x_i + b_{method} b_{fertlizer} x_i$ is zero when we plug in zero for method. How convenient! Also, when we plug in 1 for method, this part comes into play. But it is not all that is, it is *added* to the first bit. So, to interpret an interaction term for another level, you have to *add* its parameter estimate to the slope of the main effect. In our case, that's -

0.78+0.81, which is 0.03. So, really, the slope for organic grassland is 0.03. So with every unit of fertilizer more, the species richness increases with 0.03. So the effect is really positive! And, is it statistically significant? The answer is - no. This is quite unintuitive. If you look at the standard error of the interaction, it is 0.1 (rounded). So, the effect size (0.03) is much smaller than the standart error! So, really, we cannot conclude that there is a statistically significant effect in the organic grassland.

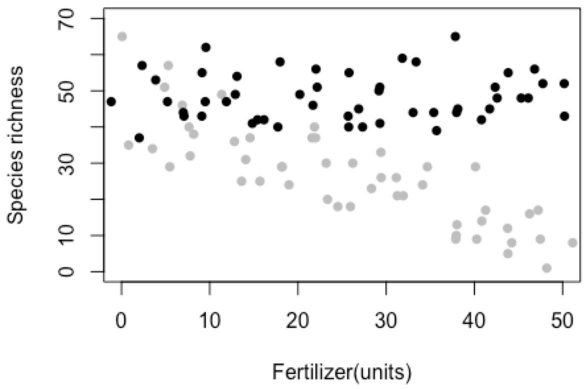
But, but - wait! Why is the interaction then statistically significant? Is it all a lie?


Unicorn Lie: A unicorn's favorite color is pink.
Unicorn Truth: Some wear only black.

Of course, not. The tricky thing is to interpret it correctly. The fact that the interaction is statistically significant only tells us that there is a statistically significant *difference* between the slope of the conventional and that of the organic grassland. This difference is statistically significantly different from zero - meaning the slope of the organic grassland differs from the slope of the conventional grassland. It does *not* tell us whether one or the other is statistically significantly different from zero. And that's the trick here.

So, the conclusion is that there is a statistically significantly negative effect of increasing amounts of fertilizer on conventional grassland, where they lose just about under a species per unit fertilizer applied. We also know that on average, organic grasslands have more species than conventional ones, - on average about 1 species more. We also know that the relationship between fertilizer and species richness is different in organic grassland, and its effect is clearly less strong - we loose fewer species for each unit fertilizer applied, about 0.8 species fewer than in conventional grassland. That ends up in a zero-sum thing, really, where we think that fertilizer has no effect on species richness in organic grassland. We can plot this easily, and then we also see that the increase in species richness with fertilizer applied in organic grassland (black dots) is really not different from zero:

```
plot(dat$species_richness[dat$method=="conventional"]~dat$fertilizer[dat$meth
od=="conventional"], pch=16, xlim=c(0,50),ylim=c(0,70), col="grey", ylab="Spe
cies richness", xlab="Fertilizer(units)")
points(dat$fertilizer[dat$method=="organic"],dat$species_richness[dat$method=
="organic"], pch=16, col="black")
```



Clearly, interactions are tricky. They are easier when you only have two levels in a fixed factor in the interaction. One rule is to always write out your model equation, and plug in your variables. That usually helps clarify things up a bit. Always keep the error size in mind. If you interact two continuous variables, things can get even a bit more complicated.

Let's quickly check the plots (spoiler: they are ok).

```
plot(fullmodel)
```

Linear models - interpretation of interactions - three-level fixed factor and continuous variable

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

For this one, we'll further explore the biology of unicorns. But we'll use a different dataset: Three-way-Unicorn.

```
d<-read.table("Three-way-Unicorn.txt", header=TRUE)
str(d)

## 'data.frame':    150 obs. of  3 variables:
##  $ Gender    : Factor w/ 3 levels "female","male",..: 2 2 2 2 2 2 2 2 2 2
...
##  $ HornLength: num  8.11 13.92 3.81 10.58 6.56 ...
##  $ Bodymass  : num  90.1 84.7 96.7 89.2 90.2 ...

names(d)
```

```
## [1] "Gender"     "HornLength" "Bodymass"

head(d)

##   Gender HornLength Bodymass
## 1   male   8.111044 90.06962
## 2   male  13.915633 84.67157
## 3   male   3.811436 96.68255
## 4   male  10.578858 89.24179
## 5   male   6.558336 90.19147
## 6   male  10.011351 87.52648
```

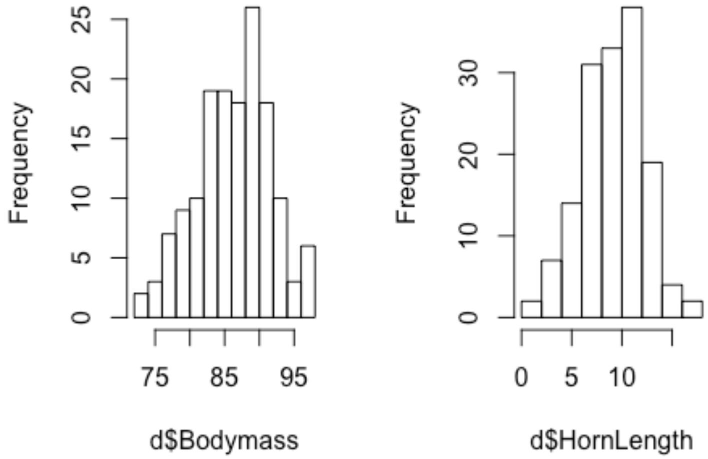
Ok. Here we try to explore whether the relationship between horn length and bodymass is different between the genders. This time we'll turn our hypothesis around: We suggest that only unicorns who are fat are able to grow a long horn. Thus, this time, hornlength is our response variable, and body mass the explanatory variable. See, how the exact way we formulate the hypothesis changes which variable is the response and which is the explanatory variable? It is important to note this, - sometimes students put the variables into the wrong spots for a given hypothesis. Make sure you get this right every time! The analysis must match the hypothesis. Of course, if you don't have a specific hypothesis, you can't find the right analysis. So, if you ever find yourself struggling with coming up with the right analysis, go back to your hypothesis. If you do not have a hypothesis, the go get one fast, because that's all what science is about!

Back to unicorns. Since we know that there are some sex-specific effects going on (from the literature or our earlier analysis), we expect the effect to be sex-specific. Since unicorns have three genders, it is an analysis with a three-level factor, and a continuous explanatory variable.

This is clearly a different population, judging by the data. Basic descriptive statistics first:

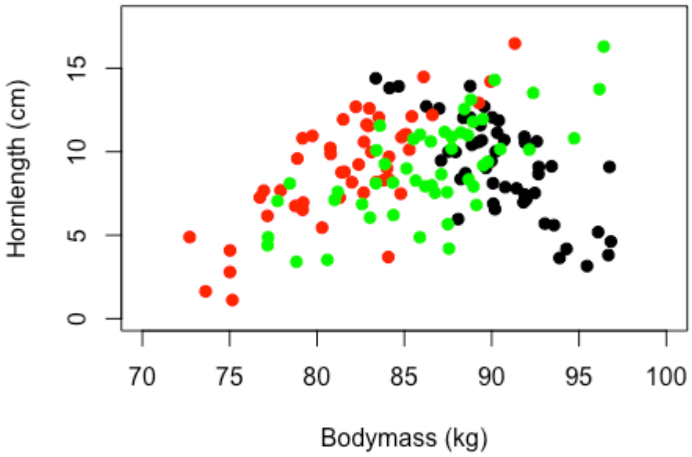
```
mean(d$Bodymass)

## [1] 86.22465

sd(d$Bodymass)

## [1] 5.299923

var(d$Bodymass)

## [1] 28.08919

par(mfrow=c(1,2))
hist(d$Bodymass, main="")
mean(d$HornLength)

## [1] 9.061447

sd(d$HornLength)

## [1] 2.997955
```

```
var(d$HornLength)

## [1] 8.987736

hist(d$HornLength, main="")
```



Now, we first plot the data to see what to expect. Especially when you want to consider interactions it is often worthwhile plotting first, to wrap your brain around it. Remember, we wanted to know if unicorns with a healthy (erm...) body mass are able to grow long horns. Therefore, our response variable is hornlength (assuming it *responds* to how heavy a unicorn is). Bodymass is our explanatory covariate, and gender (male, female, not sure) our three-level fixed factor. Let's give it a go!

```
## null device
##           1

dev.off()

plot(d$HornLength[d$Gender=="male"]~d$Bodymass[d$Gender=="male"], xlim=c(70,1
00),ylim=c(0,18), pch=19, xlab="Bodymass (kg)", ylab="Hornlength (cm)")
points(d$Bodymass[d$Gender=="female"],d$HornLength[d$Gender=="female"], col="
red", pch=19)
points(d$Bodymass[d$Gender=="not_sure"],d$HornLength[d$Gender=="not_sure"], c
ol="green", pch=19)
```



Ok. Looks like we may slightly different means (mostly in bodymass), and different slope in males, than in both females and undecided unicorns. Let's check it out!

```
mod<-lm(HornLength~Gender*Bodymass, data=d)
summary(mod)

##
## Call:
## lm(formula = HornLength ~ Gender * Bodymass, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7861 -1.3918 -0.0359  1.5635  4.2091
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -42.31404    6.01283  -7.037 7.36e-11 ***
## Gendermale              114.49733   10.49957  10.905  < 2e-16 ***
## Gendernot_sure           12.80513    8.28494   1.546   0.1244
## Bodymass                  0.62787    0.07348   8.544 1.70e-14 ***
## Gendermale:Bodymass      -1.32356    0.12008 -11.022  < 2e-16 ***
## Gendernot_sure:Bodymass  -0.18180    0.09871  -1.842   0.0676 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.056 on 144 degrees of freedom
## Multiple R-squared:  0.5453, Adjusted R-squared:  0.5295
## F-statistic: 34.53 on 5 and 144 DF,  p-value: < 2.2e-16
```

Uhh, lots of nicely significant data. That's what we like. Let's see what this means...

First of all, an overview. The intercept is confusing, and statistically significantly different from zero, but since we didn't standardize our data, it doesn't tell us much. Female seems to be the reference category, as male and not sure are explicitly mentioned. The interaction between male and bodymass is statistically significant, that means, the difference between this one and the slope for females (bodymass, 0.63) is statistically significantly different from zero. The interaction for the third gender is not statistically significant, although close.

Now let's look at females. For females, we can calculate the hornlength by the formula:

Hornlength of females = -42.31 + 0.63*Bodymass

Given that bodymass ranges between 70 and 100, this seems to make quite a lot of sense. Clearly, the slope for females is positive, which is what we'd expect from our plot (red dots). With an increase of 1kg, hornlength increases by 0.63 cm.

For males, we would ignore the stuff for not_sure unicorn. And then the formula would be as follows.

Hornlength of males = -42.31 + 114.50 + 0.63 * Bodymass - 1.32 * Bodymass

which translates because, you know, ALGEBRA, into

Hornlength of males = 72.19 + (0.63 - 1.32) * Bodymass

and then because we can use R as a calculator

Hornlength of males = 72.19 **- 0.69** * Bodymass

That's a negative slope! Cool! The SE of the interaction (0.12), even twice the SE (0.24) is much smaller (absolute value) than the slope of -0.69. So we can even say that the slope of males is statistically significantly different from zero, and negative. Cool. Not only is it different from the positive slope of the females, it's also statistically significant. So with an increase of 1kg in body mass, horn length actually decreases by 0.69cm. This is very different from the last example! Can we say something about the difference in average horn length between the genders? Since Gendermale is so super statistically significant? Looking at the equation with the plugged in parameter estimates, clearly, this is not very helpful at all. It feels a bit like an intercept, doesn't it? Spoiler: it is. If you add the values up, you get exactly that - an intercept for that part of the data. This should be clear from the equations. If you don't understand this, don't despair. Just revisit this concept again and again, and try out different things to understand it better. Repetition is helpful here. It's a complex issue! Important is: the rule to **never interpret a main effect in the presence of an interaction**. Remember that rule. But don't obey it without thinking, because, for females, we *did* interpret the main effect (bodymass). So there's that. Use your brain.

Let's move on to the absolutely fabulous indecisive unicorns. What about those? Let's plug in their numbers. Now we ignore the male stuff:

Hornlength of not_sure = -42.31 + 12.81 + 0.63 * Bodymass - 0.18 * Bodymass

And, again, we use the magical powers of algebraic knowledge that you all hopefully retained since 8th grade:

Hornlength of not_sure = -29.5 + 0.45 * Bodymass

And that also makes sense. This slope was not statistically significantly different from the female slope. The reason for this is if you take the difference of both slopes (0.63-0.45 = 0.18) it is just not larger than twice the SE (0.2). However, this slope is statistically significantly different from zero - because 0.45 is much larger than twice the SE - 0.2. So, even if this effect size was significant in the model, when we do our interpretation correctly, we can still confirm or reject a hypothesis. And that's super cool. At least I think so.

If you are interested how the three-level factor is treated inside R, it is converted into three dummy variables, which are 0 if the data is not of the respective group, and 1 if they are. Females are always 1, because they are the reference group. You can look this up in a book if you’re interested.