earthworm_walkthrough_script

Owen

2/13/2017

Introduction and preliminaries

So far, you've experience regression and multiple regression. These are both linear models with explanatory variables that are continuous… they are numbers.

In this series of video I'll cover a few things:

- The usual workflow we use for tackling a quantitative problem.
- Looking at different types of variable in R.
- Transforming a variable to prevent a problem.
- A linear model with categorical explanatory variables.
- A linear model with a categorical and a continuous explanatory variable.
- A usual, degrees of freedom, interpreting the summary table, and making a nice graph.

Apparently badgers eat a lot of earthworms, and some researchers would like to know more about how many they eat. Its a little know fact that a part of the earthworm diet is not digested by the badgers, and this can be retreived from the badger poo. If we can figure out if and how the size of this part is related to worm size, we can derive the mass of earthworms eaten from sampling badger poo. Great!

So, someone collected worms, weighed them, dissected them, and measured the size of the part that doesn't get digested. They did this for worms from three genera.

Lets import that data:

```r
rm(list=ls())
library(tidyverse)
library(ggfortify)
dd <- read_csv("https://raw.githubusercontent.com/opetchey/BIO144/master/3_datasets/earthworm.csv")
```

In the dataset

- Gattung = Genus
- Nummer = individual with Genus identifier
- Gewicht = worm weight
- Fangdatum = Date caught on
- Magenumf = Stomach bit circumference (not digested by badger)

Make variable names consistent capitalisation:

```r
names(dd) <- c("Gattung", "Nummer", "Gewicht", "Fangdatum", "Magenumf")
```

- Look at the different types of variable:
- Response, gewicht, is continuous
- Explanatory magenumf is continuous
- Gattung is categorical, with three levels.

```r
table(dd$Gattung)
```

```
##
##  L  N Oc
## 51 49 43
```

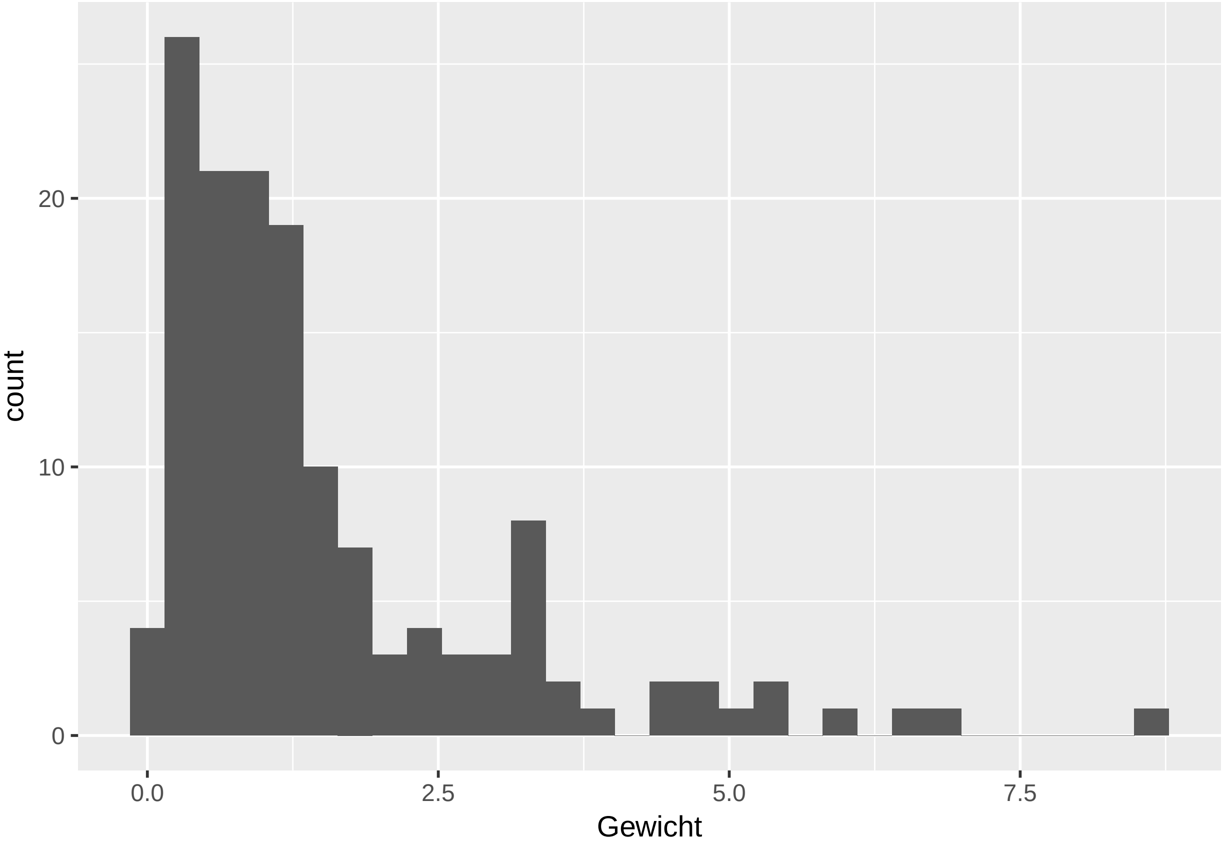
```
unique(dd$Gattung)
```

```
## [1] "Oc" "L"  "N"
```

Check the distributions of the data. First the weight of the worm:

```
ggplot(dd) +
  geom_histogram(mapping=aes(x=Gewicht))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

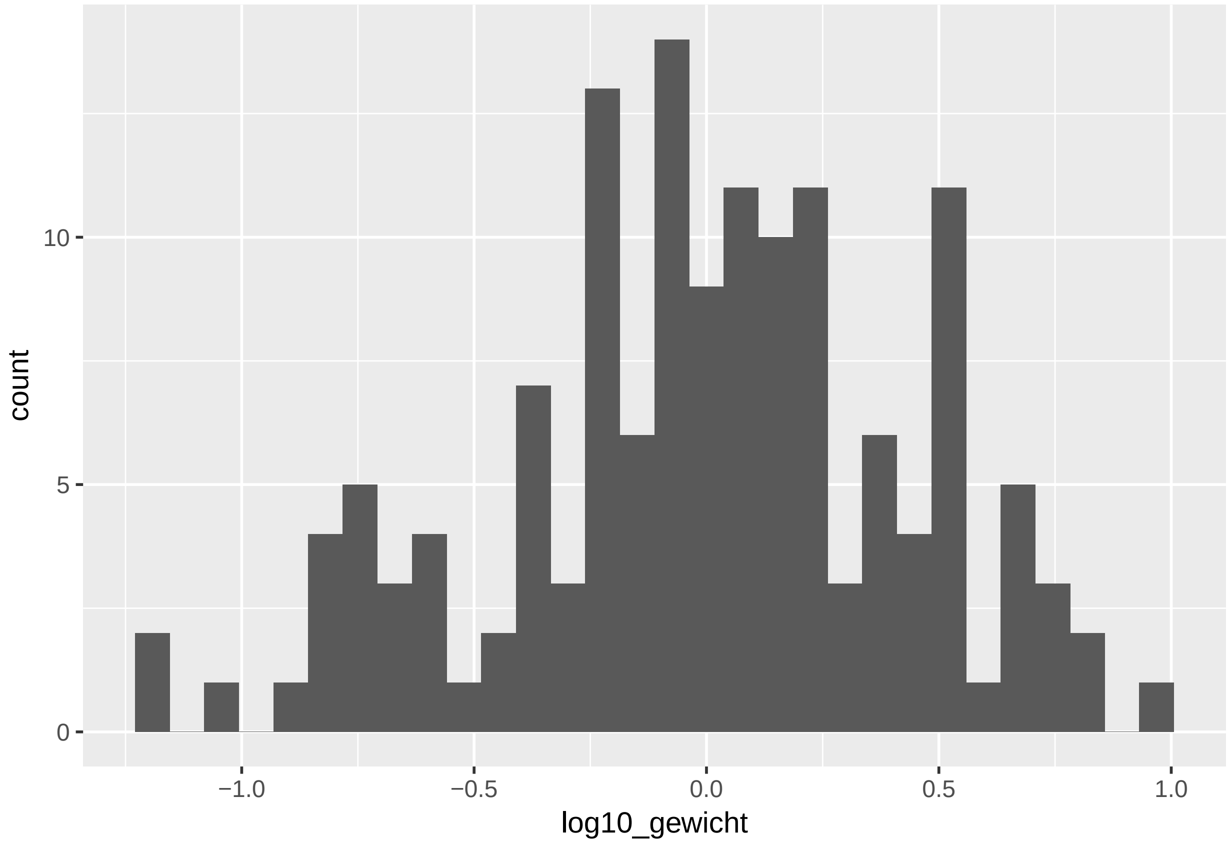


Oops, that is rather skewed. We should try a transformation. Log will probably work:

Note that Steffi in the lecture used natural log, though I find log10 generally preferable, as its easier to back transform in my head.

```
dd <- mutate(dd, log10_gewicht=log10(Gewicht))
ggplot(dd) +
  geom_histogram(mapping=aes(x=log10_gewicht))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

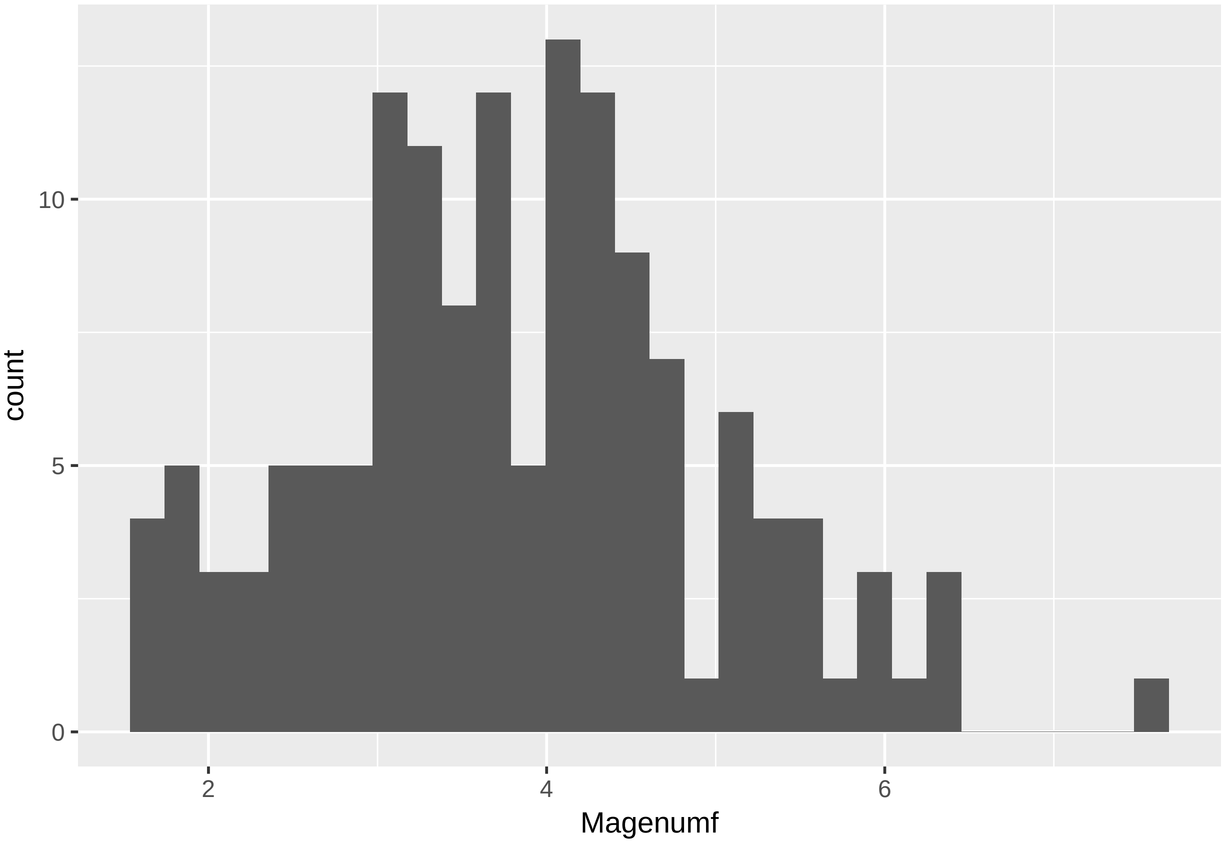


Much better, though still not perfect (never will be!).

And the stomach part circumference variable?

```
ggplot(dd) +
  geom_histogram(mapping=aes(x=Magenumf))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Looks fine.

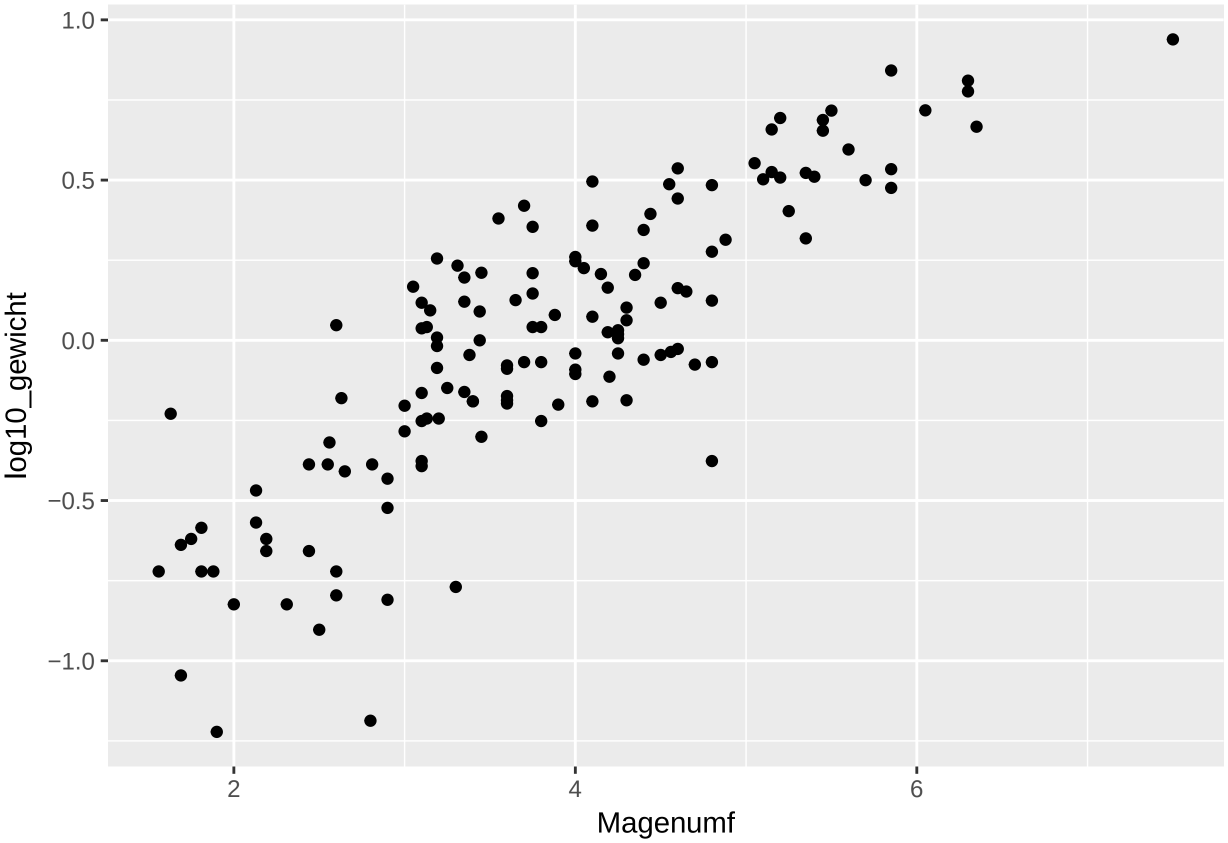
Regression, again…

Now lets look for a relationship between circumference and weight, imagine we didn't know the worms belonged to different genera.

Is there a relationship?

Degrees of freedom…

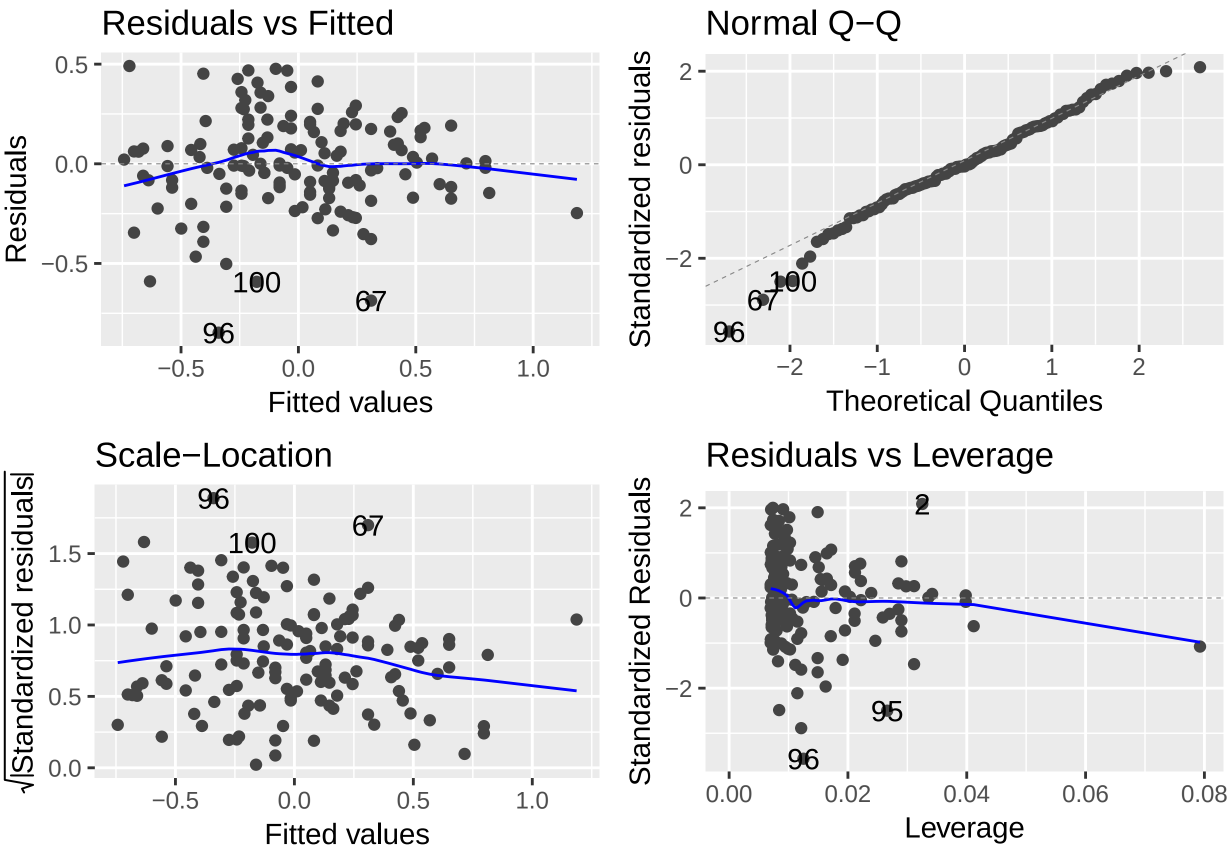
```
ggplot(dd, mapping=aes(x=Magenumf, y=log10_gewicht)) +
  geom_point()
```



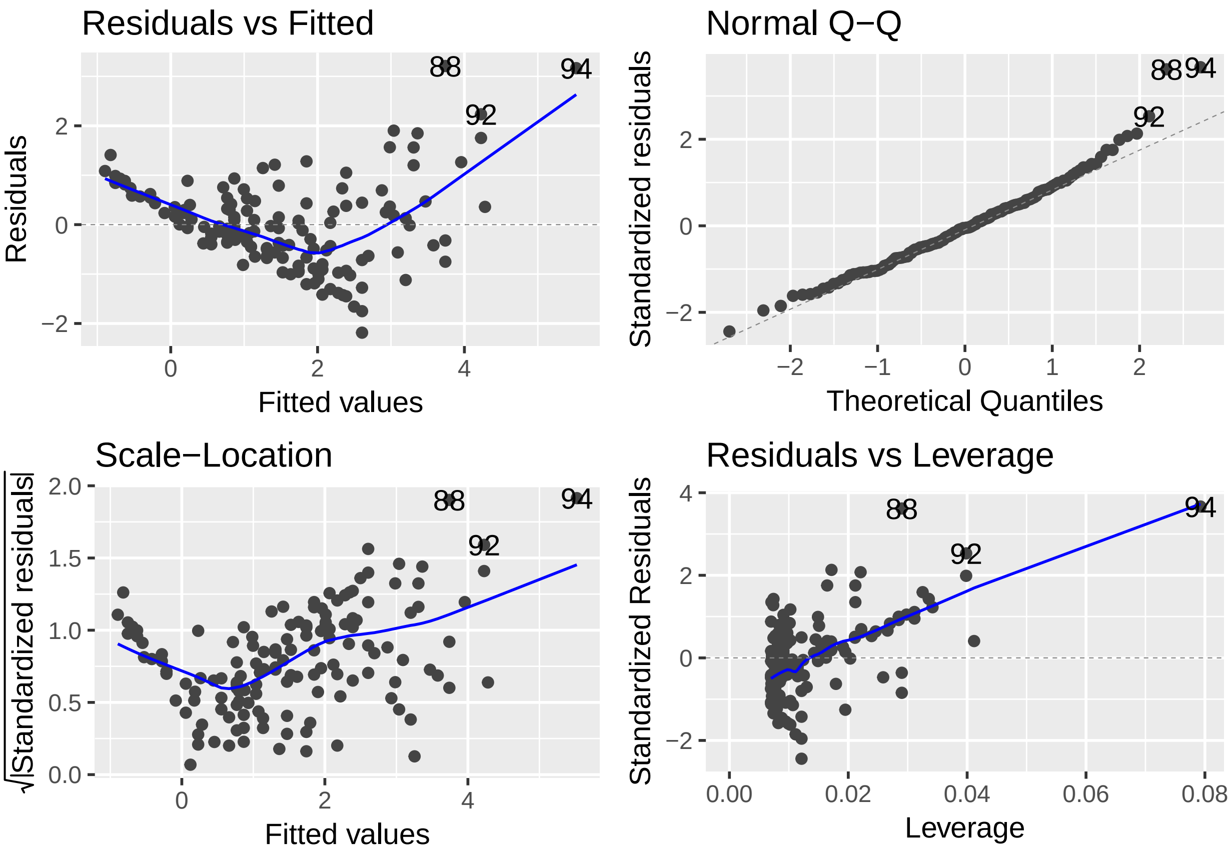
Guess the slope… 1.25 / 4 = 0.3125

And the regression model.

```
m1 <- lm(log10_gewicht ~ Magenumf, data=dd)
autoplot(m1)
```



```
autoplot(lm(Gewicht ~ Magenumf, data=dd))
```



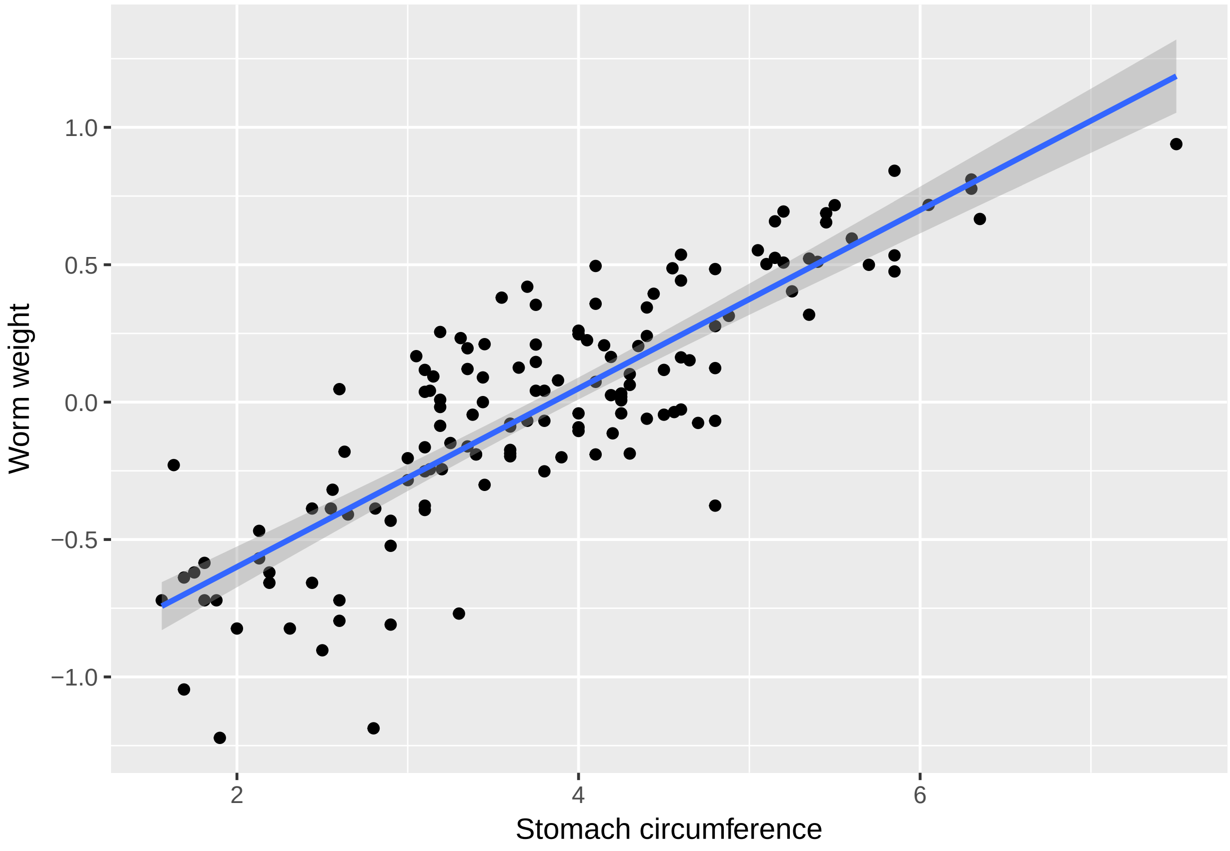
```
summary(m1)
```

```
##
## Call:
## lm(formula = log10_gewicht ~ Magenumf, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84723 -0.12446 -0.00855  0.16581  0.49062
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.24904    0.06960  -17.95   <2e-16 ***
## Magenumf     0.32471    0.01746   18.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2392 on 141 degrees of freedom
## Multiple R-squared:  0.7104, Adjusted R-squared:  0.7083
## F-statistic: 345.8 on 1 and 141 DF,  p-value: < 2.2e-16
```

DF good. About 72% variance explained. V. significant. Slope close to our guess.

A nice graph:

```
ggplot(dd, mapping=aes(x=Magenumf, y=log10_gewicht)) +
  geom_point() +
  stat_smooth(method="lm") +
  xlab("Stomach circumference") +
  ylab("Worm weight")
```

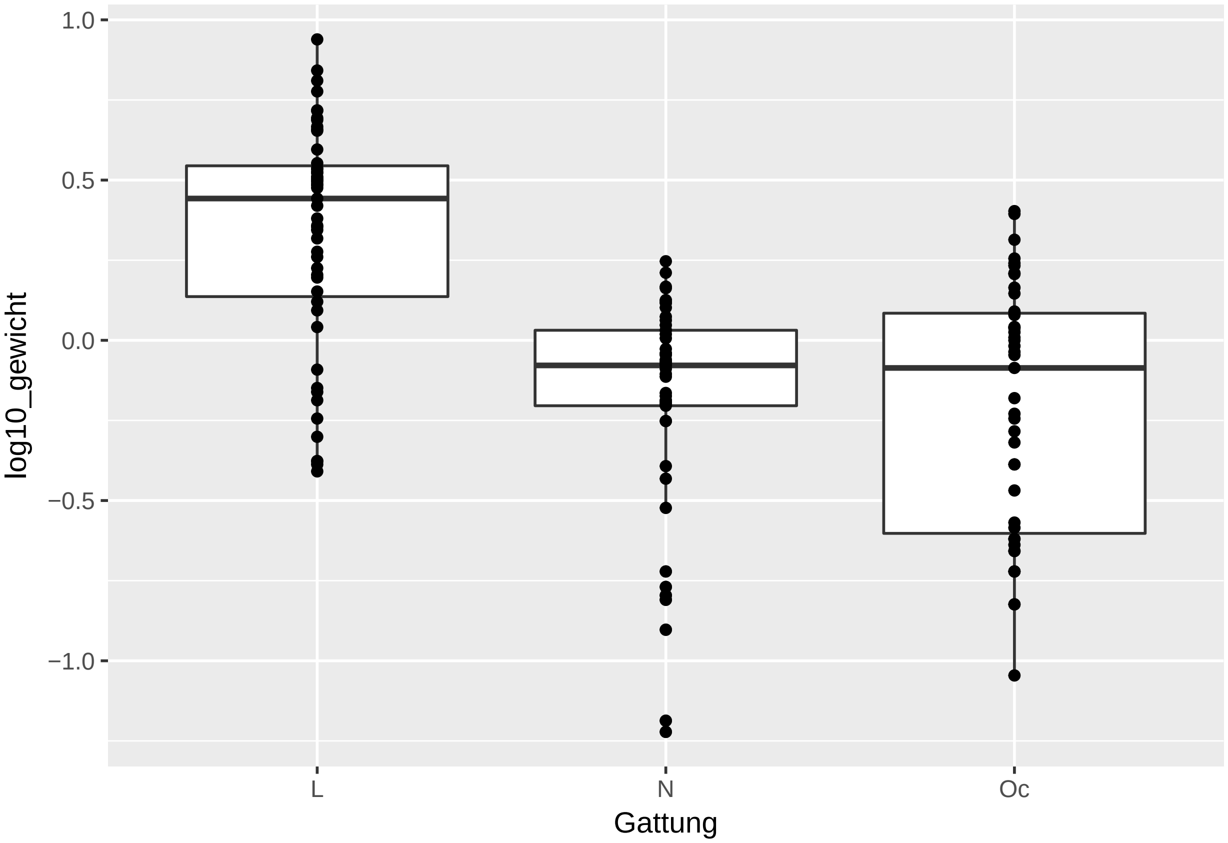


Do different genera have different weights?

Lets ignore the stomach part circumference, and simply ask if the weights differ among the genera.

First a graph:

```
ggplot(dd, mapping=aes(x=Gattung, y=log10_gewicht)) +
  geom_boxplot() +
  geom_point()
```



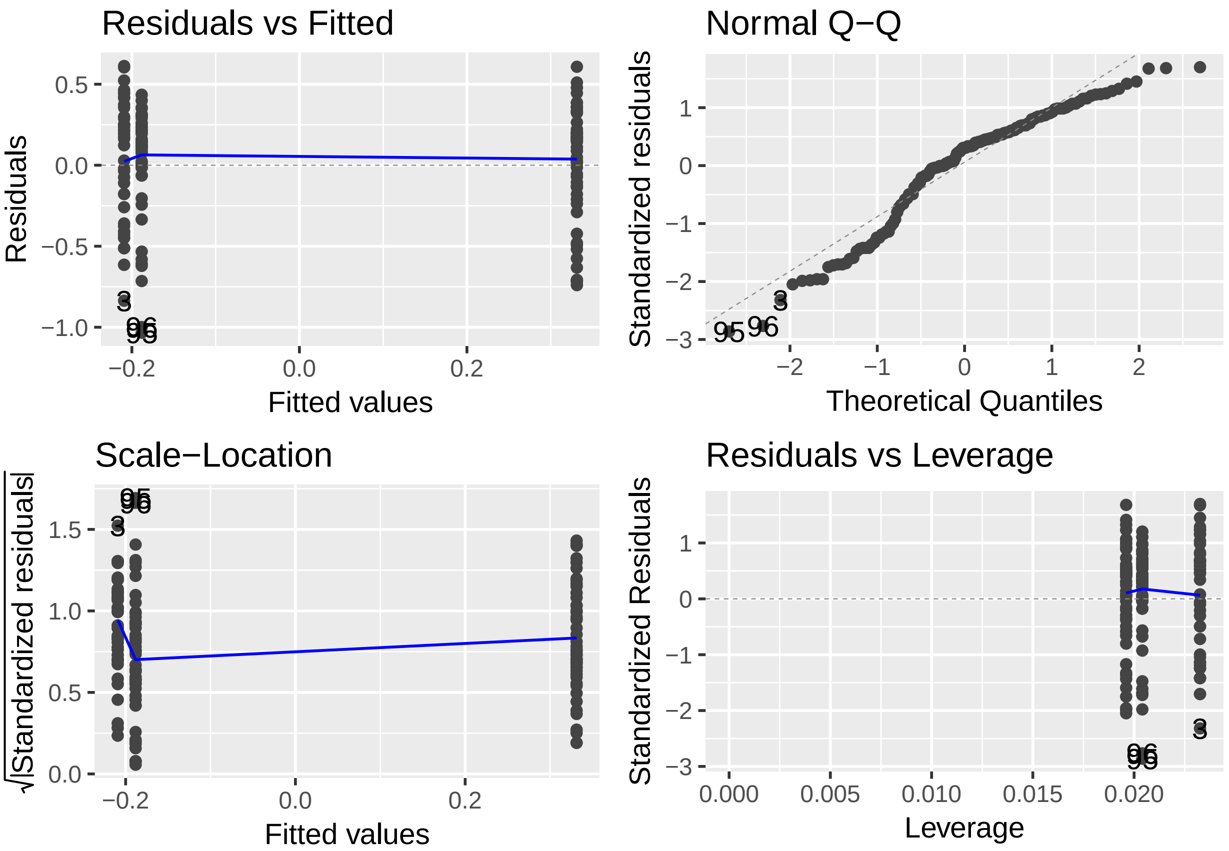
Yes, they do! Guess: * L = 0.4 * N = 0.25 * Oc = -0.1

And note that: * N - L = -0.15 * L - Oc = -0.5

Degrees of freedom. We just guessed three things. This means that the linear model will need to estimate three things: a mean for each of the genera. As before df is number of data points minus number of parameters estimated, so we will have 143 - 3 df for error.

Now the model.

```r
m2 <- lm(log10_gewicht ~ Gattung, data=dd)
autoplot(m2)
```



And the summary table.

```
summary(m2)
```

```
##
## Call:
## lm(formula = log10_gewicht ~ Gattung, data = dd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0336 -0.2075  0.1112  0.2507  0.6124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.33128    0.05109   6.484 1.43e-09 ***
## GattungN    -0.51953    0.07299  -7.118 5.22e-11 ***
## GattungOc   -0.54056    0.07554  -7.156 4.27e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3649 on 140 degrees of freedom
## Multiple R-squared:  0.3306, Adjusted R-squared:  0.321
## F-statistic: 34.57 on 2 and 140 DF,  p-value: 6.293e-13
```

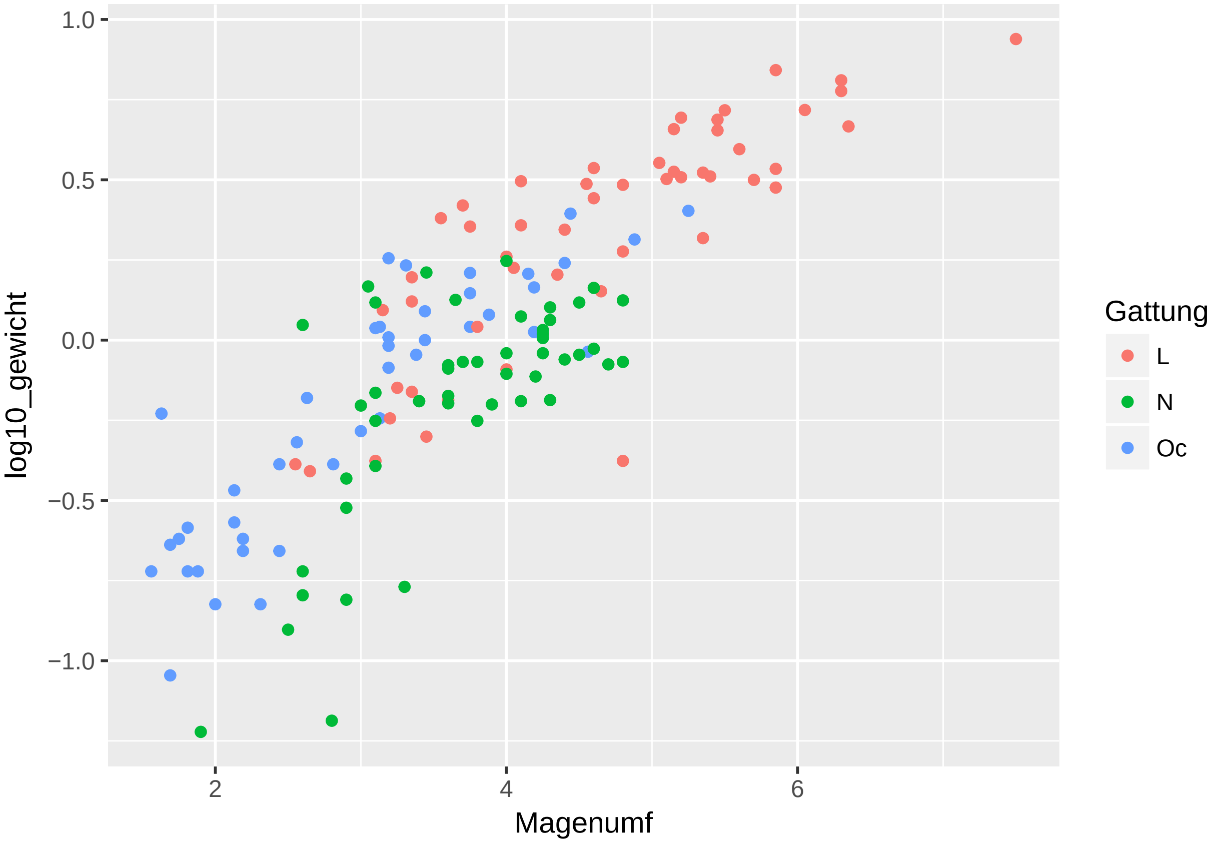
DF good. About quarter of variability explained. Worms species differ in their weights.

How did `lm` know to not do regression? We gave it a categorical variable, so it did what is appropriate for a linear model with a categorical explanatory variable. Super!

Lets give it both!

First a graph:

```
ggplot(dd, mapping=aes(x=Magenumf, y=log10_gewicht, colour=Gattung)) +
  geom_point()
```



Looks like the three species might share the same relationship between circumference and weight... the different colours seem to fall all on one line. At leat there isn't a really obvious difference.

There are now two models we could do that include both:

```
m3 <- lm(log10_gewicht ~ Gattung + Magenumf, data=dd)
```

and

```
m4 <- lm(log10_gewicht ~ Gattung * Magenumf, data=dd)
```

I've done the model diagnostics, and both look fine, by the way.

Lets first see what the first model is, by looking at the summary table.

```
summary(m3)
```

```
##
## Call:
## lm(formula = log10_gewicht ~ Gattung + Magenumf, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75956 -0.11976  0.00954  0.12383  0.56835
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.10117    0.09618 -11.449  < 2e-16 ***
## GattungN    -0.22372    0.04781  -4.679 6.76e-06 ***
## GattungOc   -0.03940    0.05555  -0.709    0.479
## Magenumf     0.30916    0.01967  15.719  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2197 on 139 degrees of freedom
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.7538
## F-statistic: 145.9 on 3 and 139 DF,  p-value: < 2.2e-16
```

We have four parameters. The intercept, which we know if for the L genera. So this is the intercept of the relationship for species L. And we have the other two contrasts for the intercepts. And we have a slope.

That is, we have the four parameters that were estimated in models m1 and m2, but here in one model. Note that the values are a bit different. This we expect.

Hence we see 139 degrees of freedom: four things are estimated.

What question are we asking? After taking into account difference in stomach circumference, do the genera differ in weight?

I.e. this question *show the picture*.

What about m4?

Look first at this figure. The question we're asking here, is are there different slopes for the different species. Does the data more look like this, or like this.

Put another way, we're asking if the relationship with stomach circumference **depends on / differs according to** the genera.

We're asking if there is evidence of an interaction between genera and stomach.

How many straight lines are there here? Three, each with its own intercept and slope. So we have need to estimate six things. So we expect to have 143 - 6 = 137 degrees of freedom for error.

Recall that we used and * between the two variables.

```r
summary(m4)
```

```
##
## Call:
## lm(formula = log10_gewicht ~ Gattung * Magenumf, data = dd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75318 -0.12834  0.01742  0.12268  0.59732
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.92394    0.13402  -6.894 1.82e-10 ***
## GattungN           -0.49990    0.21454  -2.330   0.0213 *
## GattungOc          -0.33921    0.17228  -1.969   0.0510 .
## Magenumf            0.27091    0.02816   9.620  < 2e-16 ***
## GattungN:Magenumf   0.06516    0.05289   1.232   0.2200
## GattungOc:Magenumf  0.07894    0.04430   1.782   0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2185 on 137 degrees of freedom
## Multiple R-squared:  0.7652, Adjusted R-squared:  0.7566
## F-statistic: 89.29 on 5 and 137 DF,  p-value: < 2.2e-16
```

Six rows and 137 df. Great.

We have the intercept, and the two levels of categorical, as before. We have the slope for GattungL, and then the difference between the slope of L and N, and the difference between the slope of L and Oc.

The colon (:) is saying these are interaction terms.

r-squared for circumference along was about the same as with model with interactions. Little to be gained beyond considering circumference.

Recap and summary

Model specification in R

Main effects only + Main effect and interaction *

Four models, one with only a continuous expl var, one with only a categorical, and two with both. With the difference between them being if an interaction was included.

Biologically, what does this mean?