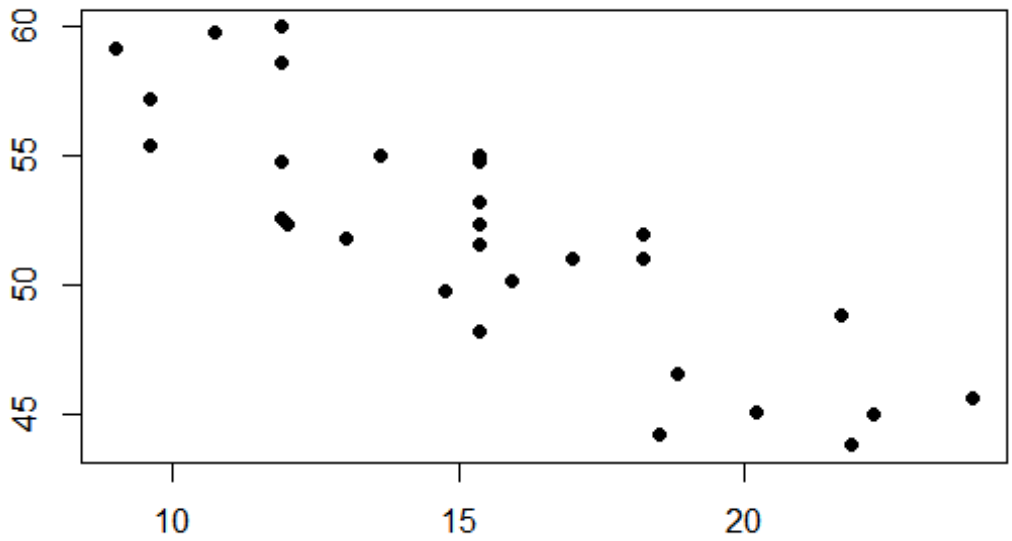
Session 4: Predictive relationships: regression

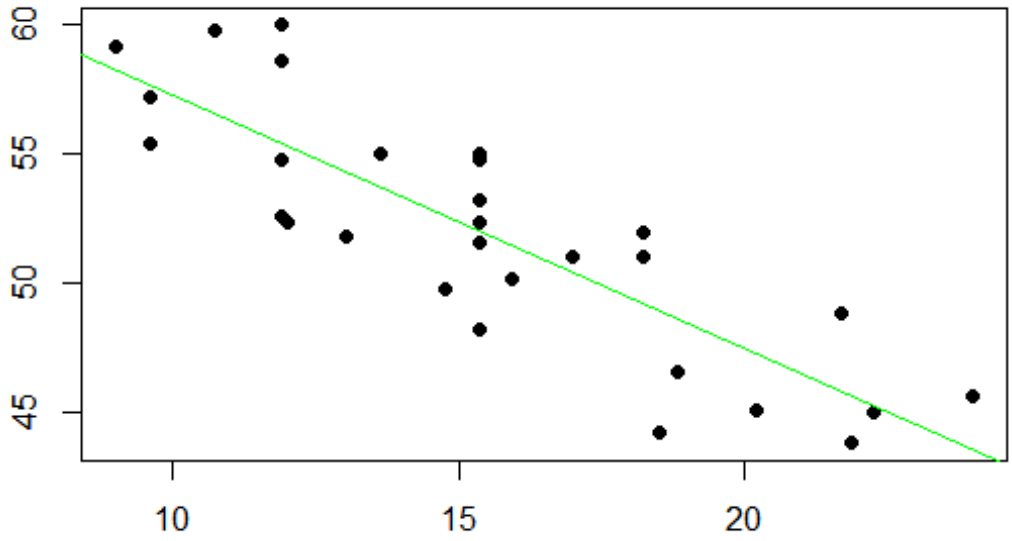
Regression, or linear models with continuous IVs

So far we have looked at linear models with categorical independent variables, or in short ANOVA. However linear models are much more flexible than this and we can also use them with independent variables that are continuous. This is also known as regression and is focused on trying to determine a predictive relationship between your dependent variable and one or more independent variables.

Regression is closely related to correlation. Both relate to whether there is some kind of relationship between two sets of continuous variables (which would normally be represented as a scatterplot).



The key difference is that a correlation does not imply any *causal or predictive relationship* between the two variables being considered - it just assesses the extent to which there is an *association* between them (strength and direction). Regression on the other hand establishes a predictive relationship between the independent variable or predictor (which is always on the x-axis) and the dependent variable or response (which is always on the y-axis). This means that for a any given value of your independent variable, you can predict a value of your dependent variable. The predictive relationship is represented by a line fitted through the points.



The line does not fit precisely to the points, as they are not in an exact straight line. The distances from each point to the line in a vertical direction are the **residual** values. The line that is fitted minimises the sum of the squared residual values.

Mongoose mania

Download the `mongoose.csv` file from Moodle and import it into an object called `mongoose`.

We will also be making use of the `car` package again, so load this up

```
library(car)
```

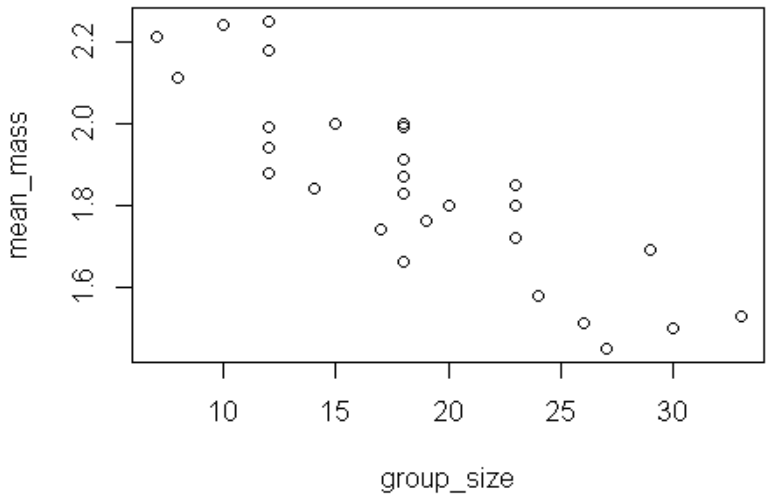
These data came from a study of the factors influencing the body masses of adult banded mongooses (*Mungos mungo*). The data are mean body mass (kg) of mongooses in a social group, the group size (no. of individuals), the area of the territory (hectares) and the area of refuse dumps within each territory (hectares). Mongooses are known to supplement their food supply by raiding refuse dumps for suitable items of food.

Review the data using `str(mongoose)`. All the columns are continuous - we have no categories as regression is about relationships between continuous values.

To start with we are going to focus on the relationship between average adult body mass and group size.

Regression analysis is achieved through the use of the `lm` command as for ANOVA. As it is a parametric technique it makes certain assumptions. The first of these is that the relationship between your predictor and response is linear. We can assess this by plotting a scatterplot of the data.

```
plot(mean_mass~group_size, data=mongoose)
```



Just from the look of the plot there would seem to be evidence of a negative relationship between mean body mass and group size, and the relationship certainly looks fairly linear, so we can go ahead. If it is not then you should consider transforming your dependent variable to see if this makes the relationship linear.

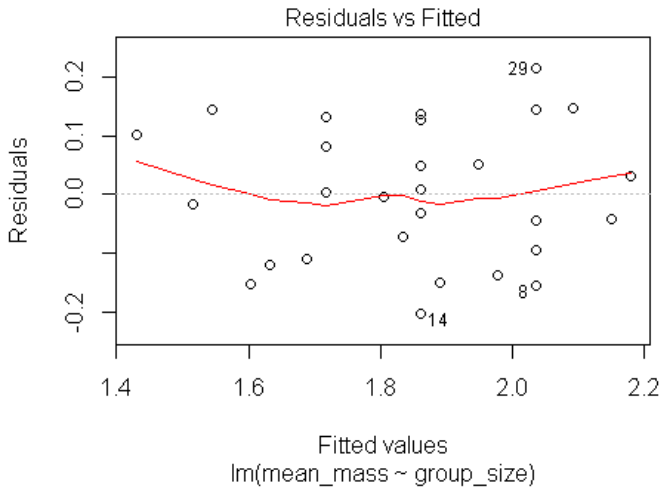
The code for the regression is very similar to that for an ANOVA.

```
m1 <- lm(mean_mass~group_size, data=mongoose)
```

Assume nothing…

Before we go any further, we need to check the remaining assumptions. In common with ANOVA, regression also makes the assumption that the variance is equal, although this time rather than between groups (as we have no groups) it assumes that the variance in the dependent variable (or response - body mass) is equal across the range of the values of the independent variable (or predictor - group size). We can check this using the `plot` command again.

```
plot(m1, which =1)
```

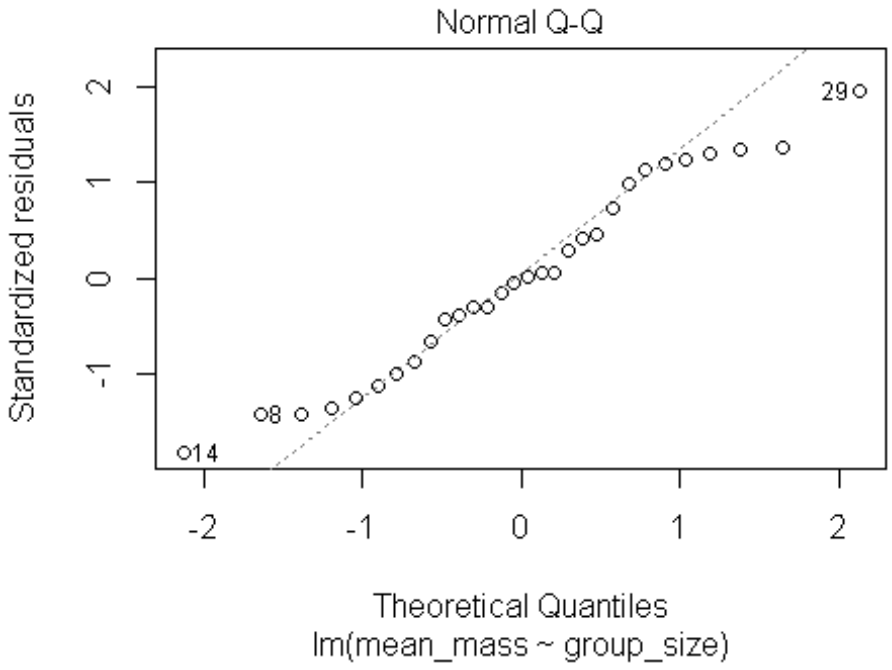


So as before, what we are looking for is any evidence of a consistent change in the range of values from left to right. The most common issue is the 'wedge' shape, with increasing variability from left to right. If you have a wedge or other shape evident in the plot then consider transformation of your dependent variable again.

In this case there is no evidence of any problem. Unlike ANOVA, there is no test (like the Levene test) that you can use to check your interpretation - it is down to examining the plot.

The final assumption is that the residuals are normally distributed. We can check this in exactly the same way as for ANOVA.

```
plot(m1, which = 2)
```




```
shapiro.test(m1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m1$residuals
## W = 0.96365, p-value = 0.3826
```

Again in this case, there is no evidence of any issue - the plot looks OK, although the residual values at either end look like they are a bit away from the line. The outcome of the Shapiro-Wilk test however indicates that the residuals are not significantly different from normal. Therefore we can move on to interpeting the output.

Relatable?

To get the detail of the outcome, and assess whether there is a significant relationship between the two variables, we use the `summary` command.

```
summary(m1)
```

```
##
## Call:
## lm(formula = mean_mass ~ group_size, data = mongoose)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.202549 -0.090261 -0.001493  0.095866  0.214211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.382268   0.063039  37.790  < 2e-16 ***
## group_size  -0.028873   0.003235  -8.926 1.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1138 on 28 degrees of freedom
## Multiple R-squared:   0.74,  Adjusted R-squared:  0.7307
## F-statistic: 79.68 on 1 and 28 DF,  p-value: 1.11e-09
```

There is quite a lot going on here, so we will unpack this information bit by bit.

Start at the bottom: the information here gives the overall significance of the regression i.e. is there a statistically significant relationship between your dependent variable (`mean_mass`) and your independent variable (`group_size`). The p-value here is very much less than 0.05, so you have evidence of a significant relationship. The other values (F and df) are also needed for reporting, see below.

The next line up indicates what proportion of the variation in `mean_mass` is explained by `group_size` (this can vary between 0 - none of the variation and 1 - all of the variation). This is called the R-squared or R^2 value). There are two values given but **we always use the `Adjusted R-squared' value**. The closer this is to 1, the stronger the relationship. In this case it is 0.731, which indicates quite a strong relationship.

For reporting, you would write this up as 'there was a significant negative relationship between mean body mass of banded mongooses and the size of the group they were living in (linear regression, $F = 79.68$, $df = 1, 28$, $p < 0.001$, adjusted $R^2 = 0.731$)'.

Lining it up

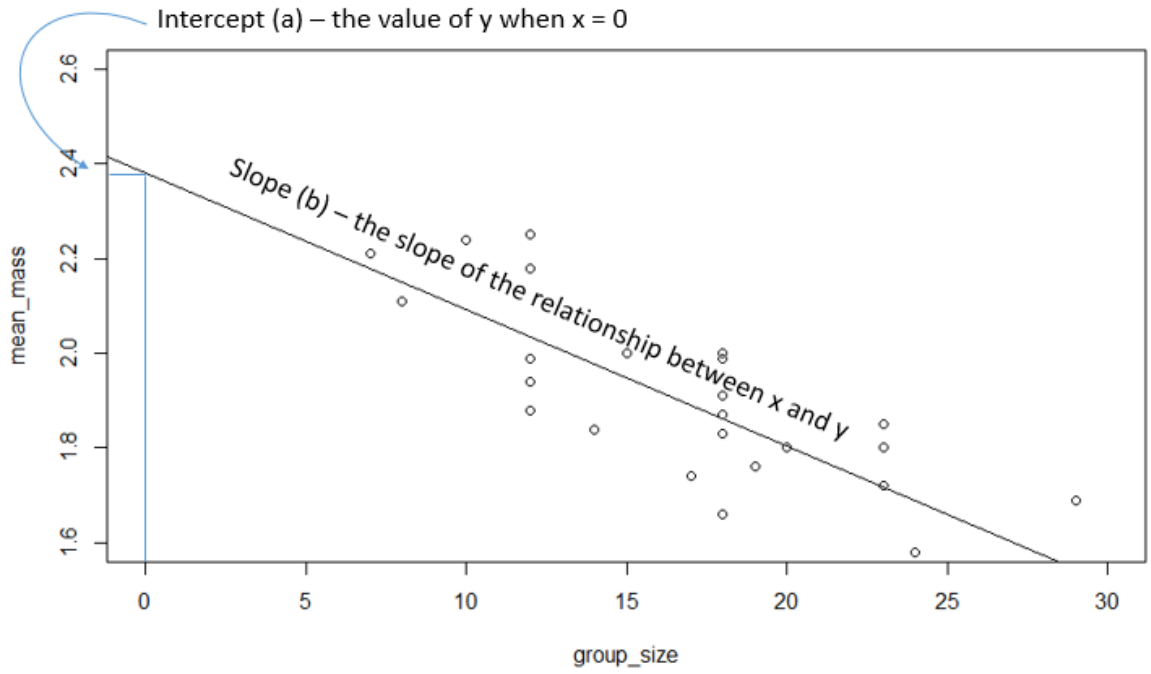
The next information that you need to focus on is the `Coefficients` part. The values in the 'Estimates' column give you the information required to fit the line described by the regression to the data.

The general equation of a linear regression line is as follows:

$y = a + bx$

where

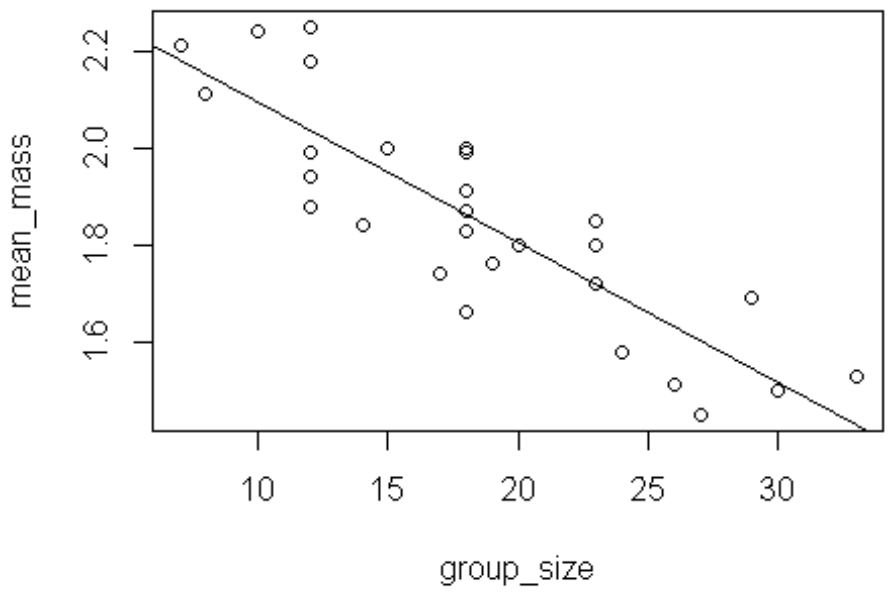
- y is your dependent variable or response
- a is the intercept on the y-axis, the value of y where x = 0
- b is the slope of the relationship between y and x
- x is your independent variable or predictor



So here, the intercept is 2.38 (consistent with the fitted line on the plot above) and the slope of the line is -0.029. The magnitude of this value indicates the change in mean body mass that would result from a change in group size of one individual. The negative value indicates that the line is sloping down (so an **increase** in group size results in a **decrease** in mean body mass)

To replicate the plot given above, use the following code:

```
plot(mean_mass~group_size, data=mongoose)
abline(m1)
```



The axis ranges will be different from the previous plot as by default R limits the axes to the range of the data values. You can change this by using `xlim` and `ylim` e.g. `plot(mean_mass~group-size, data=mongoose, xlim=c(0,30))`. If you want to change the colour of the fitted line, you can do so along these lines `abline(m1, col="Green")`.

The p-values in the Coefficients table (`Pr(>|t|)`) indicate the significance of the different parts of the equation (intercept and slope). We will come back to these shortly.

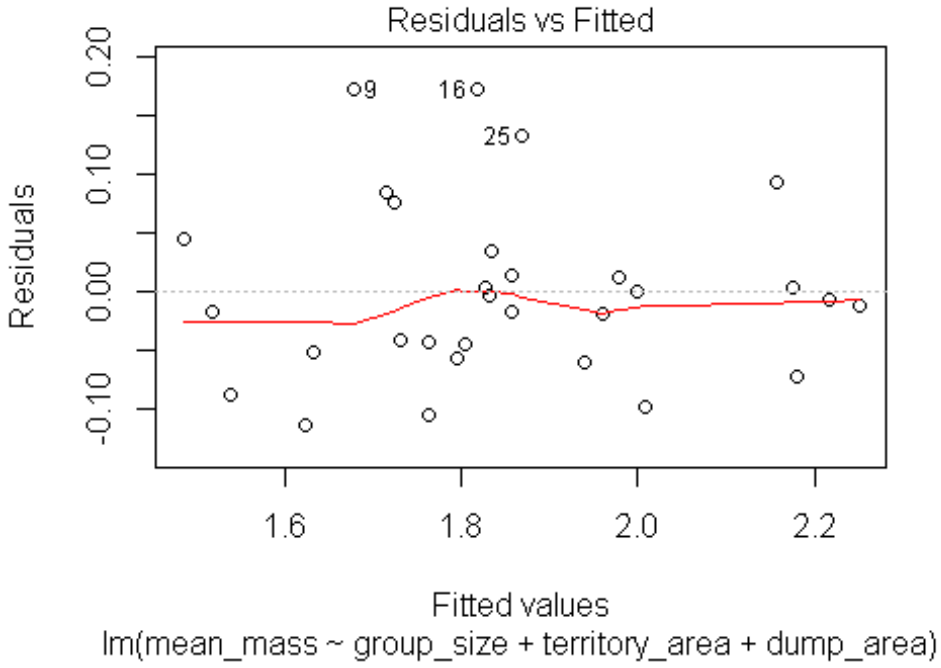
Predictors in plural

In the example above, we just assessed the relationship between a response (mean body mass) and a single predictor (group size). Life is rarely this simple unfortunately! In most cases a particular response variable will be influenced by many different things. We can therefore extend the approach above to consider multiple predictors. This is known as multiple linear regression.

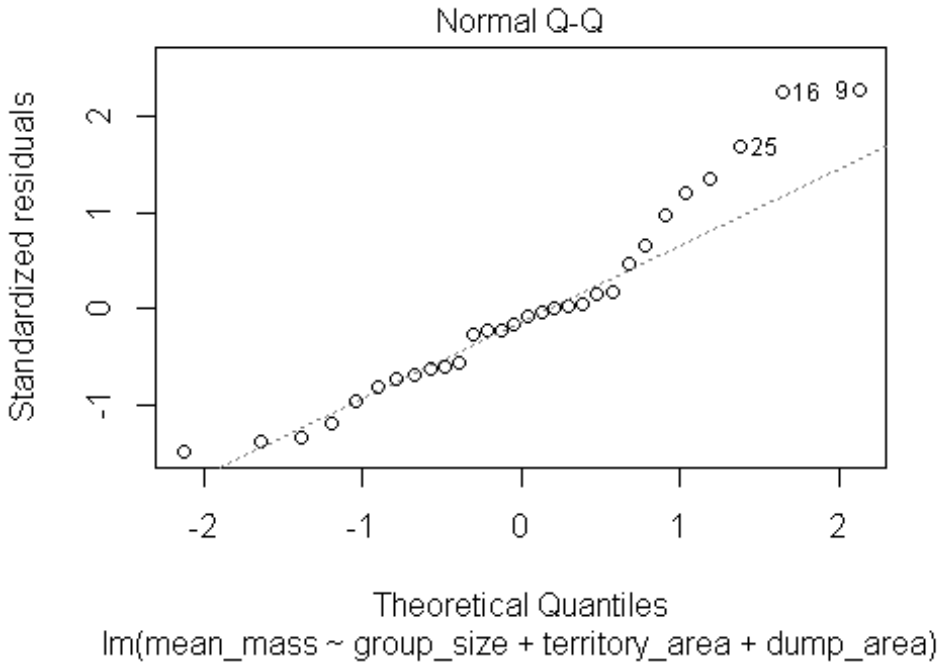
In the `mongoose` object there were a number of other columns which we have not considered so far. We will now go back and reanalyse the data, only this time considering the effects of group size, territory area and area of refuse dumps on mean body mass.

The code for this is a simple extension of that which we used above. First you should do scatterplots of each predictor against the mean mass measurements. These give some information on their own as to the likely results, but we need to evaluate them formally.

```
# repeat the analysis, but include multiple predictors this time
m2 <- lm(mean_mass~group_size+territory_area+dump_area, data=mongoose)
plot(m2, which = 1)
```




```
plot(m2, which = 2)
```




```
shapiro.test(m2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.93588, p-value = 0.07047
```

All looking good again, so we can take a look at the outcome

```
summary(m2)
```

```
##
## Call:
## lm(formula = mean_mass ~ group_size + territory_area + dump_area,
##     data = mongoose)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.112395 -0.049528 -0.008631  0.029691  0.171500
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.1230892  0.0847666  25.046  < 2e-16 ***
## group_size     -0.0267066  0.0022852 -11.687 7.51e-12 ***
## territory_area -0.0003356  0.0010706  -0.313    0.756
## dump_area       0.2957010  0.0534845   5.529 8.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0792 on 26 degrees of freedom
## Multiple R-squared:  0.883,  Adjusted R-squared:  0.8695
## F-statistic:  65.4 on 3 and 26 DF,  p-value: 3.047e-12
```

The output is similar to that shown before, but now we have more rows in the `Coefficients` part of the table, reflecting the fact that we have three predictors. The overall regression (bottom line) is still significant, and the R^2 value has increased slightly.

An inflated opinion

Before we go any further, there is an other consideration to think about. If you include predictors i.e. independent variables, that are highly correlated with each other, the outcome of regression is difficult to interpret and unreliable as it is not possible to separate out the individual influence of the IVs. A high correlation between predictors goes by the wonderful name of **multicollinarity**.

In order to assess this, we calculate the equally well-named **Variance Inflation Factors (VIF)**. These assess the extent to which multicollinearity is a problem and can be calculated as follows.

```
# make sure car is loaded first!
vif(m2)
```

```
##     group_size territory_area      dump_area
##       1.029849       1.020147       1.050243
```

A value above 5 indicates that there is a problem, and above 10 a severe problem. If you have values that indicate there is an issue then a pragmatic strategy is as follows. For each variable that seems to be a problem, run a separate simple regression with just that independent

variable included and only use the independent variable with the highest R^2 value in the subsequent multiple regression.

In this case all the values are low, so there does not seem to be any issues. Now we can look in more detail at the different variables, specifically at the significance of each independent variable included.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.1230892  0.0847666  25.046  < 2e-16 ***
group_size     -0.0267066  0.0022852 -11.687 7.51e-12 ***
territory_area -0.0003356  0.0010706  -0.313    0.756
dump_area       0.2957010  0.0534845   5.529 8.37e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0792 on 26 degrees of freedom
Multiple R-squared:  0.883,  Adjusted R-squared:  0.8695
F-statistic:  65.4 on 3 and 26 DF,  p-value: 3.047e-12
```

In this case the `territory_area` variable highlighted above is not significant (p = 0.756). What this means is that this variable does not explain a significant part of the variation in mean body mass. We should therefore remove it from the regression as it is not important. The other two variables (group_size and dump_area) are significant, so we keep these in.

Rerun the regression (storing the output in `m3`) including only group_size and dump_area as independent variables and doing appropriate checks on assumptions etc. Both of the remaining variables are significant, so there is no need to remove any more. If you had multiple predictors in your regression that were not significant, you would remove them one at a time starting with the least significant (i.e. the one with the *highest* p-value) until only significant predictors are left in the regression.

At the end of all this we have a highly significant regression model for predicting mean body mass of banded mongooses from group size and refuse dump area. Based on the slope values in the summary table, you should be able to see that group size has a negative effect on mean body mass and refuse dump area a positive effect.

For reporting this could be summarised as: there was a significant relationship between mean body mass of banded mongooses and group size and area of refuse dumps (multiple linear regression: $F = 101.4$, $df = 2, 27$, $p < 0.001$, adjusted $R^2 = 0.874$).

You would also report the details of the slope coefficients and intercept (plus their standard errors and significance) in a separate table, to indicate the direction of the relationships.

Make sure that you have written sufficient comments in your script to be able to go back to it in future sessions - you will need to be able to interpret your commands for other examples and for the tests in due course.
