Multiple Regression

Bob O'Hara

February 19, 2019

This week: Multiple Regression

We will look at

- explaining our dependent variable with more than one explanatory variable
- how to fit these models in R
- what a design matrix is (this will be helpful later)
- how to fit a polynomial model

More Monsters

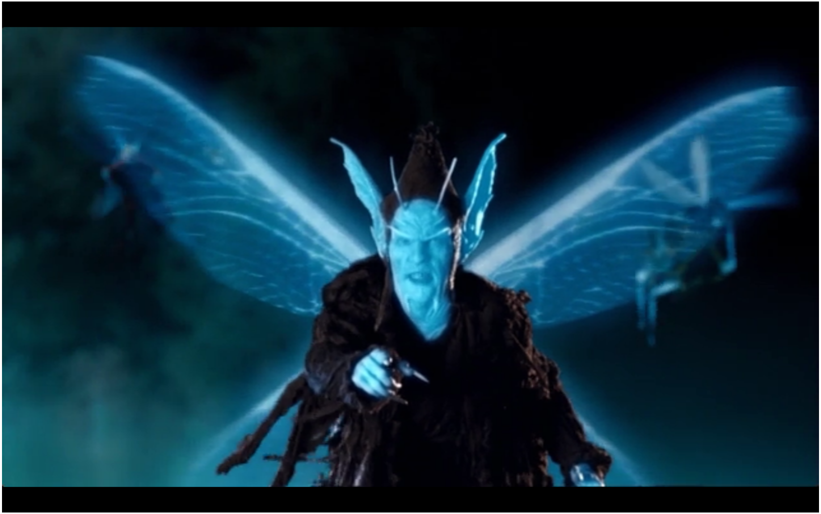


Figure 1:

More Monsters

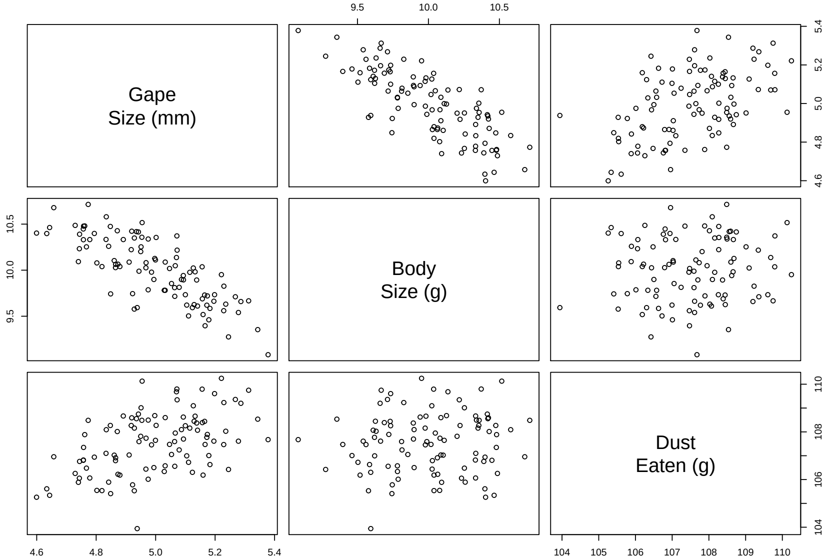
In the cellar of the museum in Frankfurt we had a population of Schey.

These are small creatures that lurk in the dark and eat ancient dust and stale cobwebs.

Some of us wanted to know more about them, and whether they could be trained to clean the museum collections.

We caught 100 and measured the amount of dust they could eat in 5 mins, and wanted to explain that by their body size, their gape size (i.e. how large their mouths are).

The Data


Gape
Size (mm)
Body
Size (g)
Dust
Eaten (g)
9.5
10.0
10.5
5.4
5.2
5.0
4.8
4.6
10.5
10.0
9.5
110
108
106
104
4.6
4.8
5.0
5.2
5.4
104
105
106
107
108
109
110

Simple regression

```r
Dir <- "https://www.math.ntnu.no/emner/ST2304/2019v/"
File1 <- "Week7/ScheyData.csv"
Schey <- read.csv(paste0(Dir, File1))
plot(Schey, labels=c("Gape\nSize (mm)", "Body\nSize (g)",
                     "Dust\nEaten (g)"))
```

Your Task

- regress dust eaten against gape size
- regress dust eaten against body size

For both of these, note the model parameters (in particular the gape and body size effects), and how much variation the model explains.

What if we have $>$1 predictor?

We often want to look at the effects of several variables together

- they may all have some effect
- we might be doing an experiment where factors interact
- we might want to model one variable as a polynomial

The model

This is our model for simple regression

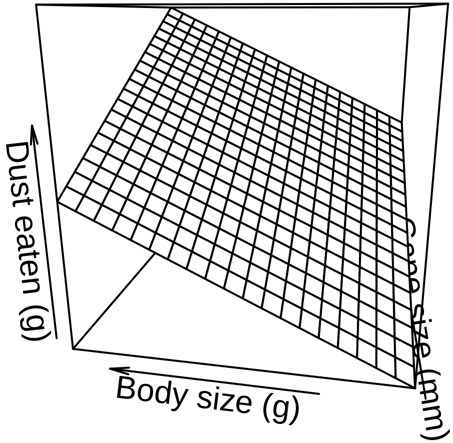
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

How can we extend it to more than one variable?

The obvious model

$$E(y_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}$$

This is a plane

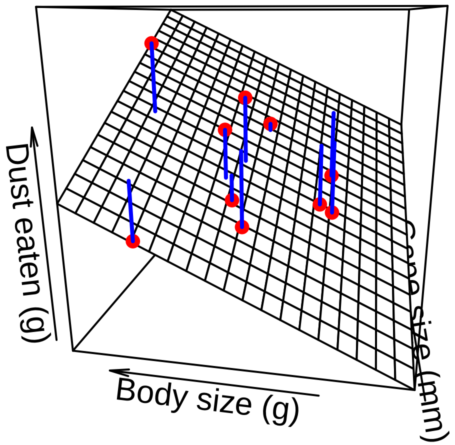


The obvious model

The model for the data is thus

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

The points deviate from the plane



Fitting in R

In R we can just use the same function as we did before.

```
FullMod <- lm(Dust ~ GapeSize + BodySize, data=Schey)
```

The only change is in the formula. It was

```
Y ~ X
```

now it is

```
Y ~ X1 + X2
```

Your Turn

```
FullMod <- lm(Dust ~ GapeSize + BodySize, data=Schey)
```

Fit the model, and get the coefficients (with confidence intervals), and the amount of variation

Compare these with the results you got from the single regression models

Regression More Generally

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

- we have p covariates, labelled from $j = 1$ to p
- we have p covariate effects
- the jth covariate values for the ith individual is x_{ij}

Design Matrices

We can write this more compactly. First, we turn the intercept into a covariate by using a covariate with a value of 1 for every data point. Then we write all of the covariates in a matrix, X:

$$X = \begin{pmatrix} 1 & 2.3 & 3.0 \\ 1 & 4.9 & -5.3 \\ 1 & 1.6 & -0.7 \\ \vdots & \vdots & \vdots \\ 1 & 8.4 & 1.2 \end{pmatrix}$$

So, the first column is the intercept, the second is the first covariate, and the third is the second covariate.

This is called the *Design Matrix*: it is helpful for writing down the model

Writing the Model

Using matrix algebra, the regression model becomes

$$\mathbf{Y} = X\beta + \varepsilon$$

where $\mathbf{Y}$, β and ε are now all vectors of length n, where there are n data points. X is am $n \times p$ matrix.

We will not look at the mathematics in any detail: the point here is that the model for the effect of covariates can be written in the design matrix.

Writing the Model

$$\mathbf{Y} = X\beta + \varepsilon$$

is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 2.3 & 3.0 \\ 1 & 4.9 & -5.3 \\ 1 & 1.6 & -0.7 \\ \vdots & \vdots & \vdots \\ 1 & 8.4 & 1.2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- β_0 is the intercept

The Solution (just so you can see it)

After a bit of matrix algebra, one can find the ML solution:

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$$

where **b** is the MLE for β.

In practice:

- you won't have to calculate this: the computer does it, and
- the computer actually doesn't use this

Multiple Regression Today

We can now write a multiple regression model

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

We can fit it in R

```
lm(Dust ~ GapeSize + BodySize, data=Schey)
```

We know what a design matrix looks like

$$X = \begin{pmatrix} 1 & 2.3 & 3.0 \\ 1 & 4.9 & -5.3 \\ 1 & 1.6 & -0.7 \\ \vdots & \vdots & \vdots \\ 1 & 8.4 & 1.2 \end{pmatrix}$$

Tomorrow

Polynomials