Lecture 11: Review/Intro to GLMs

Bob O'Hara

bob.ohara@ntnu.no

Before we start. . .

- By my reckoning we'll have 4 lectures & exercises (plus revision)
- Number of Exercises you'll need to hand in 5 (out if 9 in total)
- This week's exercises don't have to be handed in until next Friday

Today

A re-cap, to try to put everything in context

Start on Generalised Linear Models

- like the linear models we have done, but with different types of data

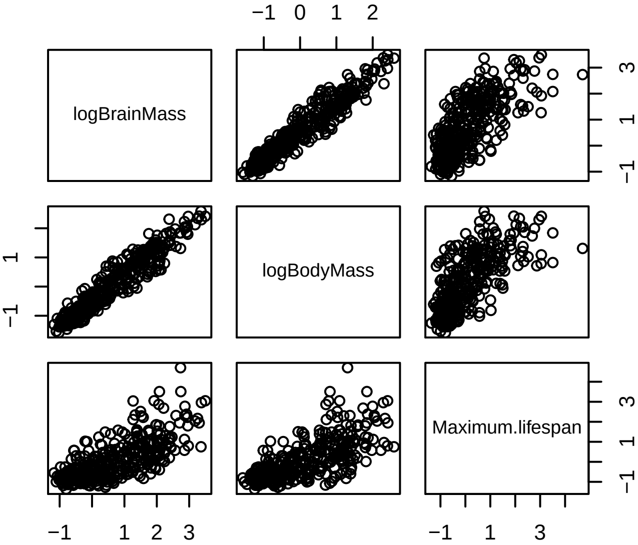
Example: Trying to Explain Bird Brain Size

The bird brain data we have been using was collected to look at the influence of life span on brain size in birds.

Easy, right? But. . .

Example: Trying to Explain Bird Brain Size

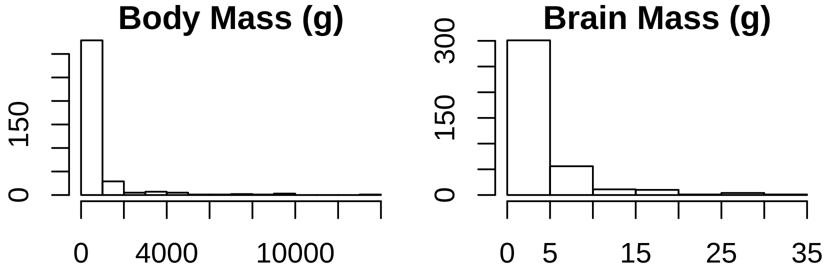
Body mass is correlated with everything



May be other correlations too

Before the modelling...

- Check the data
  - plot it
  - no obvious problems, except some horrible skew:



Before the modelling. . .

- Think about what what we want to do
- Ask if brain size is explained by lifespan (maximum lifespan)
- Correct for confounders, e.g. body size

Here: Age at first reproduction, Incubation length, Clutch size, Mean latitude, Mode of Development

Mode of Development

- how developed a chick is when it hatches
- precocial is most developed

The Strategy

Fit a model with lifespan & other confounders explaining brain size

We want to use the confounders that make a difference, but discard those that don't

Use model selection to chose the confounders

Check the model

Bird Brains: how I would approach it

For the bird brains, the data were collected to ask if life span was related to brain size. Other variables were included because they might have an effect (e.g. body size)

So, we have a specific hypothesis to test

But we want to find the other covariates that might have an effect

So, we use AIC to find these other covariates

Model

```r
# Don't use lifespan here, select best model with the rest
library(bestglm)
UseData <- cbind(BirdBrains[,Covars[-1]],
                 y=BirdBrains$logBrainMass)
AllSubsets <- bestglm(Xy=UseData, IC="AIC")
```

```
Morgan-Tatar search since factors present with more than 2
```

Now add max. lifespan

```r
TestLife <- update(AllSubsets$BestModel,
                   .~.+Maximum.lifespan, data=BirdBrains)
```

The Result

```
anova(AllSubsets$BestModel, TestLife) # could use anova(Tes
```

```
Analysis of Variance Table

Model 1: y ~ Age.at.first.reprodction + logBodyMass + M.of
Model 2: y ~ Age.at.first.reprodction + logBodyMass + M.of
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    378 25.140
2    377 24.425  1   0.71441 11.027 0.0009857 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Yep, it still has an effect

The Model

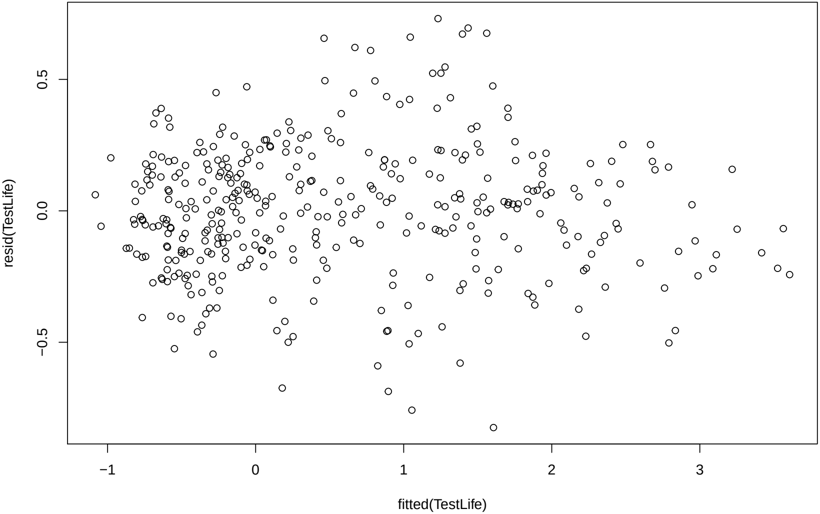
. . . but the effect is small compared to the other effects (note: these are standardised)

```r
round(summary(TestLife)$coefficients,2)
```

```
                          Estimate Std. Error t value Pr(>|t
(Intercept)                   0.70       0.02   37.21     0.
Age.at.first.reprodction      0.02       0.02    1.11     0.
logBodyMass                   1.09       0.02   45.40     0.
M.of.Dev.semi-altricial      -0.03       0.06   -0.47     0.
M.of.Dev.semi-precocial      -0.35       0.06   -5.66     0.
M.of.Dev.precocial           -0.48       0.04  -12.04     0.
Maximum.lifespan              0.07       0.02    3.32     0.
```

Model Fit 1

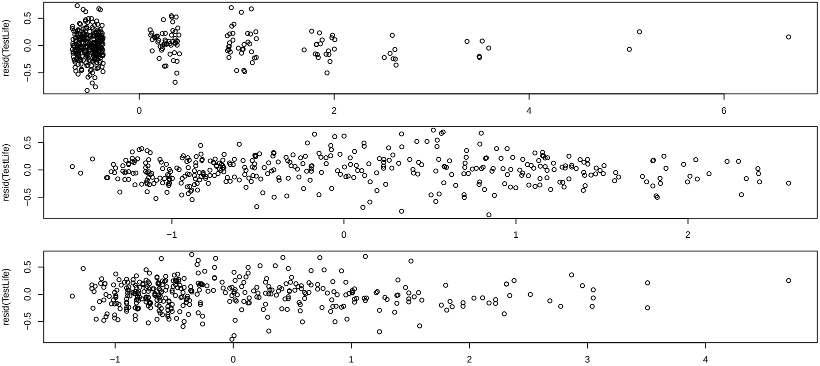
```
plot(fitted(TestLife), resid(TestLife))
```



Looks OK

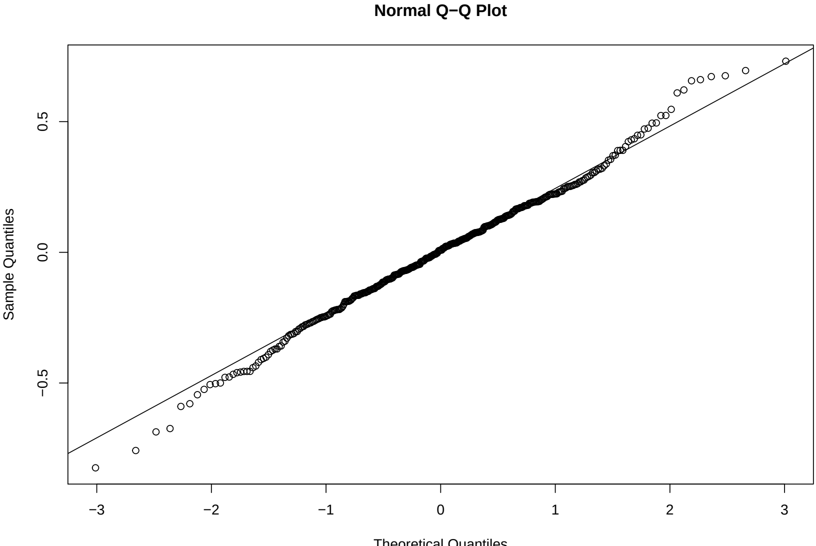
Model Fit 2

```r
par(mfrow=c(3,1), mar=c(2,4,1,1))
plot(jitter(BirdBrains$Age.at.first.reprodction), resid(Tes
plot(BirdBrains$logBodyMass, resid(TestLife))
plot(BirdBrains$Maximum.lifespan, resid(TestLife))
```



Model Fit 3

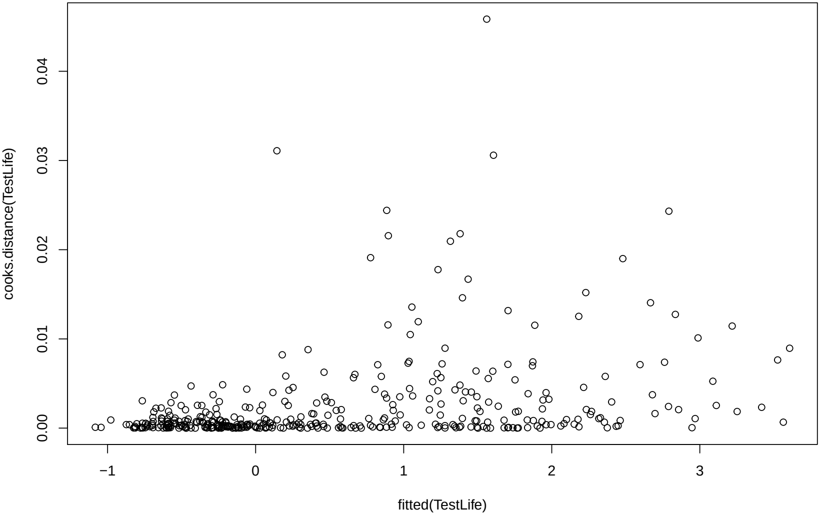
```
qqnorm(resid(TestLife))
qqline(resid(TestLife))
```



Looks

Model Fit 3

```
plot(fitted(TestLife), cooks.distance(TestLife))
```



Looks OK

Conclusion

Yes, life span does help to explain brain size

But effect is small: the change from smallest to largest lifespan is 0.42 which is a bit less than hte change from one extreme to the other of mode of development, and is less than half the effect of changing body mass by 1 standard deviation.

- in part this is because body mass and lifespan are correlated

Break

Figure 1: Eric

Generalised Linear Models

Not everything is normal

Question: what other data types are there? (e.g. counts of things)

A slight diversion

We have 2 colour morphs, with an initial frequency of the white morph of p.

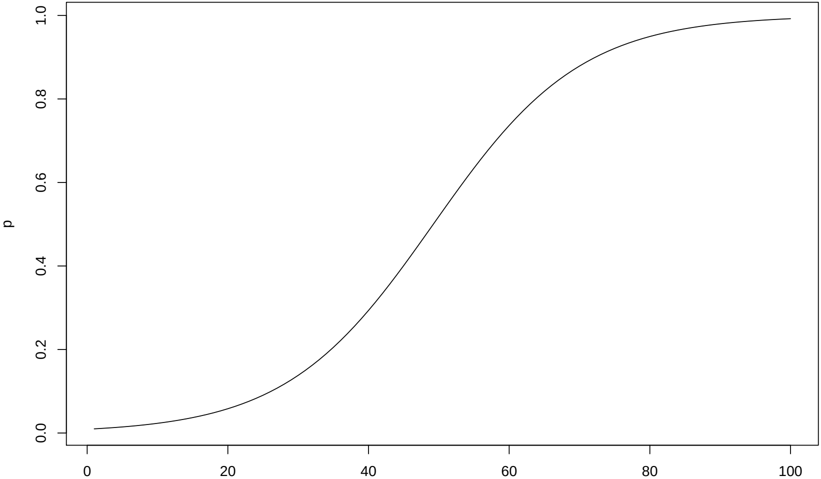
For every individual the red morph produces, the white morph produces $1 + s$

How does the frequency of the white morph change?

Asexual population genetics

The change in frequency:

$$p_1 = \frac{(1+s)p_0}{(1+s)p_0 + 1(1-p_0)} = \frac{(1+s)p_0}{1+sp_0}$$



Asexual population genetics

Another way of making the same calculation…

$$\frac{p_1}{1-p_1} = \frac{(1+s)p_0}{1-p_0}$$

So…

$$\frac{p_2}{1-p_2} = (1+s)\frac{p_1}{1-p_1} = (1+s)(1+s)\frac{p_0}{1-p_0}$$

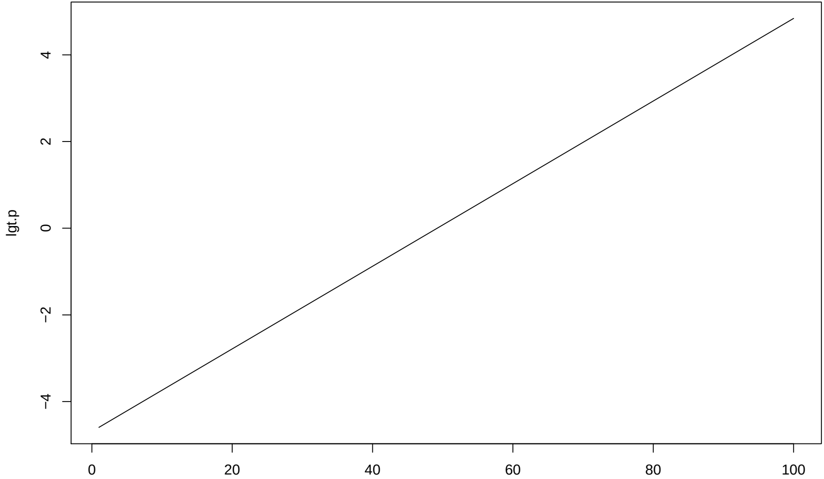
or

$$\log\frac{p_t}{1-p_{t+r}} = \mathbf{t}\log(1+s) + \log\frac{p_0}{1-p_0}$$

We can get a straight Line!

This is linear in t

$$\log \frac{p_t}{1 - p_{t+r}} = \mathbf{t} \log(1 + s) + \log \frac{p_0}{1 - p_0}$$



The (modelling) point

We can transform the model so that it is linear

$$log \frac{p_i}{1 - p_i} = \alpha + \beta x_i$$

Inference

Suppose we know p_0, and then sample N individuals to estimate p_1: of these r are white.

We want to maximise the binomial likelihood

$$l(p|r,N) = \log \left(\begin{array}{c} N \\ r \end{array} \right) + r \log p + (N-r) log(1-p)$$

Tidying up

Re-arrange to collect terms in r:

$$\begin{aligned} l(p|r,N) &= \log \begin{pmatrix} N \\ r \end{pmatrix} + r \left(\log p - \log\left(1-p\right)\right) + N \log\left(1-p\right) \\ &= \log \begin{pmatrix} N \\ r \end{pmatrix} + r \log\left(\frac{p}{1-p}\right) + N \log\left(1-p\right) \end{aligned}$$

Back to $\log p/(1-p)$!

Generalising Linear Models

It turn out that it's really convenient to model $\log p/(1-p)$:

$$\log \frac{p_i}{1-p_i} = \sum_j X_{ij}\beta_j$$

and we can work with the right hand side like we do with multiple regression.

Components of a Generalized Linear Model

Random Part

- the data (we assume a distribution, e.g. binomial)

Systematic Part

- the model for the mean for each data point: $\sum_j X_{ij}\beta_j$

The link Function

- transforms the model (which is linear) onto the scale of the data
- e.g. the *logit* link function, $\log p/(1-p)$

Binomial

We look at selection over time. We take samples of size N_t at time t, and look at the colour, and record the number that are white, r_t. Then we have

Random part:

$$r_t \; Binom(N_t, p_t)$$

Link function:

$$\log \frac{p_t}{1 - p_t} = \theta_t$$

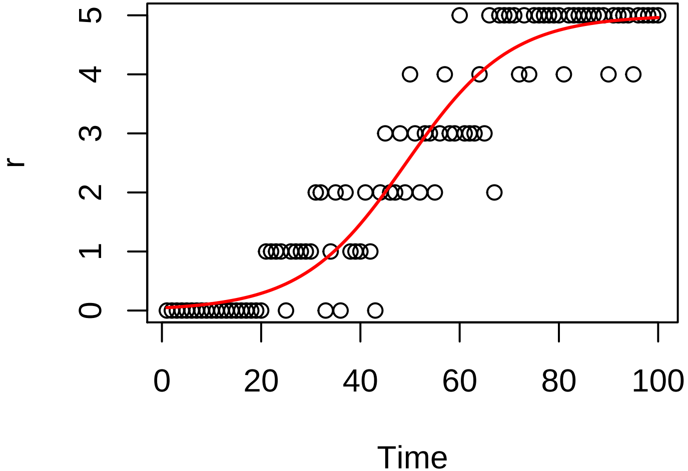
Systematic Part:

$$\theta_t = \alpha + \beta t$$

(*) this is not the best model, but is OK for now

Simulated data

```
N <- 5
r <- rbinom(length(p), N, p)
plot(Time, r)
lines(Time, p*N, col=2, lwd=1.6)
```



Fitting the Model

```r
glmod <- glm(cbind(r, N-r) ~ Time,
             family=binomial("logit"))
round(summary(glmod)$coefficients, 3)
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.572      0.407 -11.241        0
Time           0.093      0.008  11.971        0
```

The technical bit (i.e. not on the exam)

Nelder & Wedderburn showed that a lot of common models have likelihoods (for one data point) that look similar:

$$l(\theta|y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

- y is the data
- θ is the expected value (the mean)
- ϕ is the dispersion (the variance)
- $a()$, $b()$ & $c()$ are functions

Often ϕ is known, so this is also written as

$$l(\theta|y) = a(y)t(\theta) + c(\theta) + d(y)$$

For example

$$l(\theta|y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

For the binomial we'll focus on the proportion, so $y = r/N$

$$l(p|r, N) = \log \left(\begin{array}{c} N \\ r \end{array} \right) + N \left(\frac{r}{N} \log \left(\frac{p}{1-p} \right) + \log (1-p) \right)$$

$$a(\phi) = 1/N$$

$$\theta = \log \left(\frac{p}{1-p} \right)$$

and $b(\theta) = \log (1-p)$ & $c = \log \left(\begin{array}{c} N \\ r \end{array} \right)$

Likelihood for everything

$$l(\theta_i|y_i) = \sum \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)$$

We can make θ_i a linear model (like we do for regression):

$$\theta_i = \sum_j X_{ij}\beta_j$$

The Workhorse of Modern Statistics

It turns out that this has some nice properties

- the likelihood is easy to maximise
- the standard errors don't need any special tricks to calculate
- the model often behaves really nicely

Lots of models come under this framework, and many extensions come from it

What's coming

There are several distributions we can use in GLMs. We will look at these:

- Poisson (for counts)
- Binomial (for counts of proportions)

But there are others (e.g. Gamma)

Distribution, link functions are new: the systematic part is the same

- but interpretation of the parameters needs care