Generalised Linear Models (GLM): Part 1

Lecture Outline

Recap of the course so far

What are GLMs and why do we use them?

Components of a GLM

Maximum likelihood and GLMs

Fitting in R

Lecture Outline

Recap of the course so far

- EX1: Course so far

What are GLMs and why do we use them?

- EX2: Non-normal data

Components of a GLM

- EX3: Examples of non-normal data

Maximum likelihood and GLMs

Fitting in R

- EX4: Fit in R

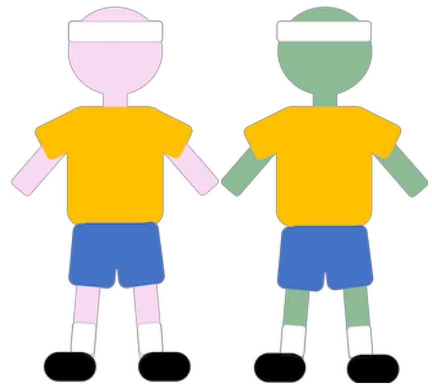
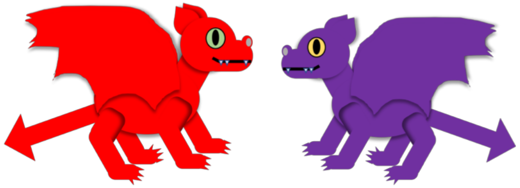
Reading

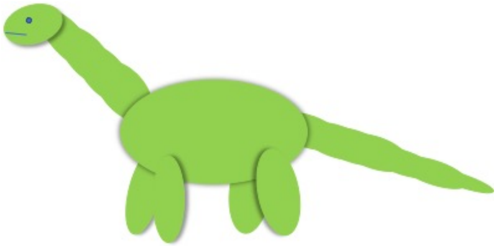
Chapter 8 – The New Statistics with R

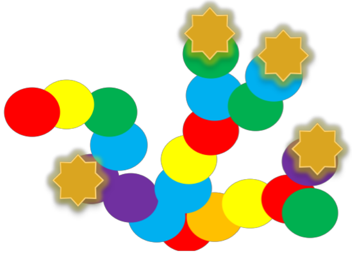
Recap of the course so far

Exercise 1: What have we covered so far?

- Think about the previous weeks, write on your boards some of the topics we have covered.




The modelling process

DATA

+

BIOLOGICAL QUESTION

The modelling process

Mathematical description of how the data were generated.

E.g.
- Distribution
- Linear equation (lines or groups)
- Defined by parameters

The modelling process

Inference

Choose a model

Get estimates of parameters

E.g. Maximum likelihood estimation

Find the parameters that give the highest likelihood given the data.

The modelling process

Inference

Choose a model

Get estimates of parameters

Quantify uncertainty in estimates

The modelling process

Choose a model

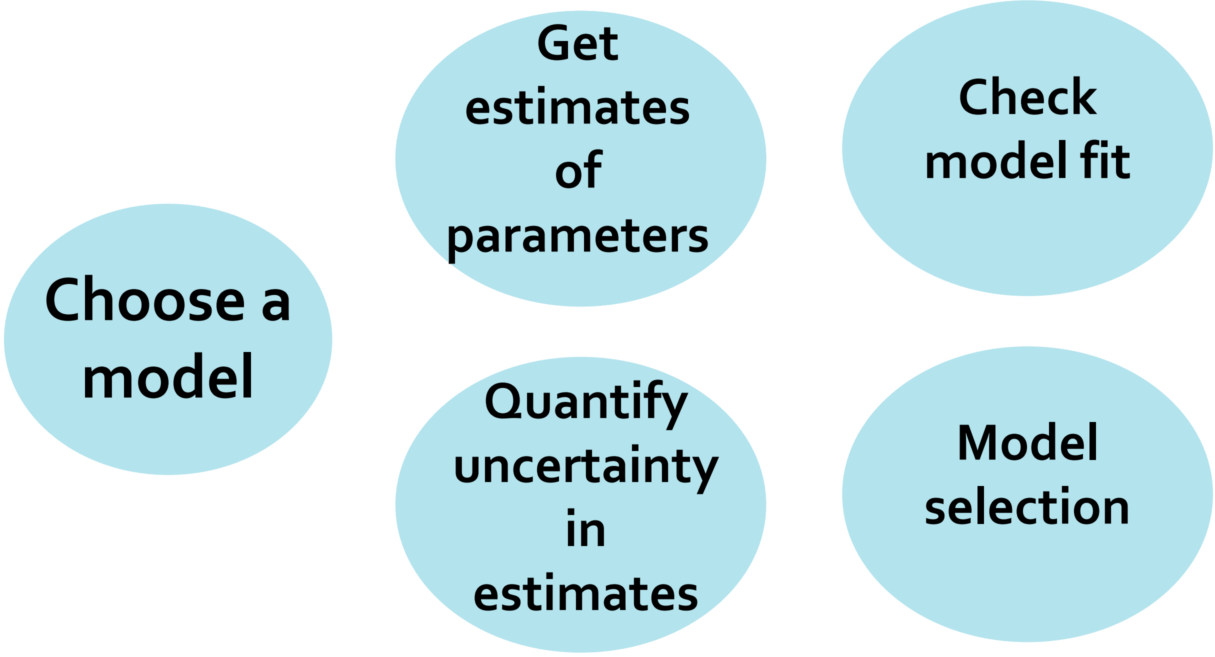
Get estimates of parameters

Quantify uncertainty in estimates

Check model fit

E.g. Check assumptions have been met

The modelling process



E.g. Exploratory or confirmatory

Using AIC, BIC, or anova and F-Tests

The modelling process

Choose a model

Get estimates of parameters

Quantify uncertainty in estimates

Check model fit

Model selection

Use results to make conclusions about the population

What are GLMs and why do we use them?

Linear models

Use linear equations to model a continuous response as a function of explanatory variables

Linear models

Use linear equations to model a continuous response as a function of explanatory variables

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Linear models

Use linear equations to model a continuous response as a function of explanatory variables

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

linear predictor

error

Linear models

Use linear equations to model a continuous response as a function of explanatory variables

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

linear predictor

error

Systematic part

Random part

Linear models

Assumptions:

- straight line **(linearity)**
- errors are independent
- errors have same variance **(homoscedasticity)**
- errors are normally distributed
- errors have zero mean

Exercise 2: Is a linear model appropriate?

- Take a look at the three datasets on the next slides (you have data plotted with the modelled line from a linear model, residual vs fitted plot, and a Normal Q-Q plot).

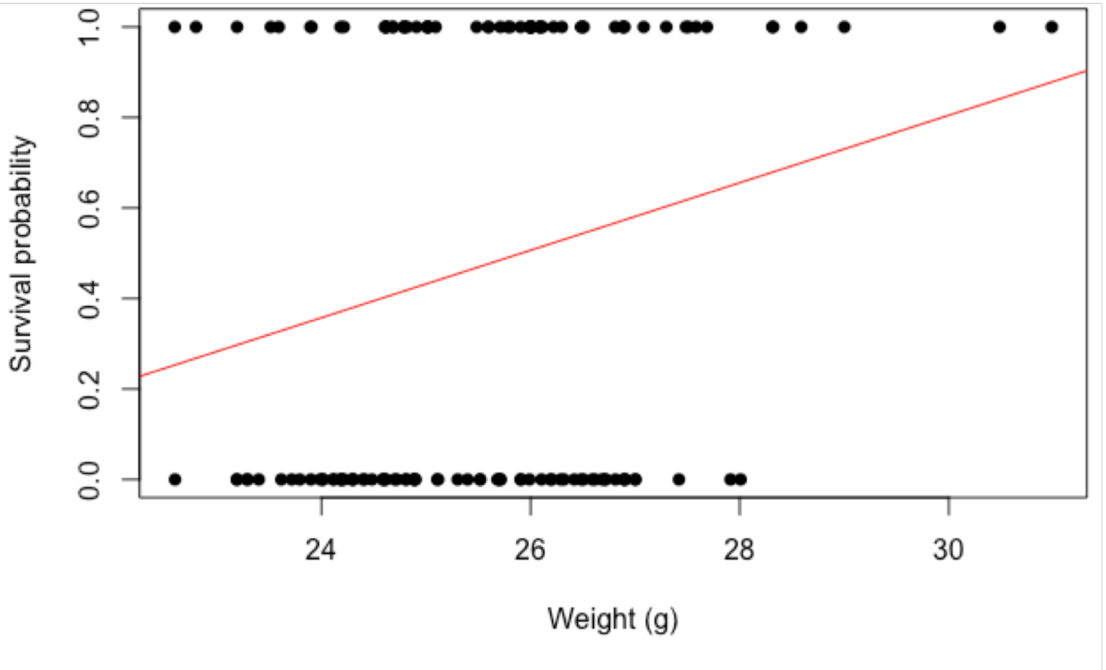
- For each, answer the questions:

Is a linear model a suitable model for this data?

If not, why not?

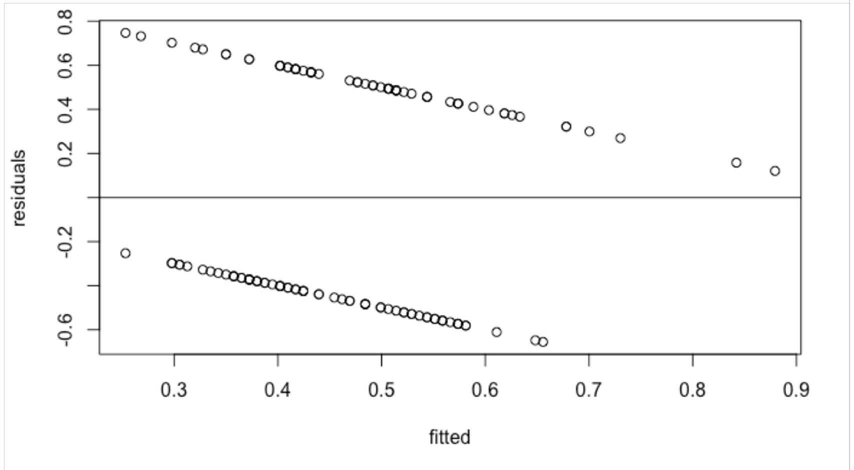
How could you improve it?

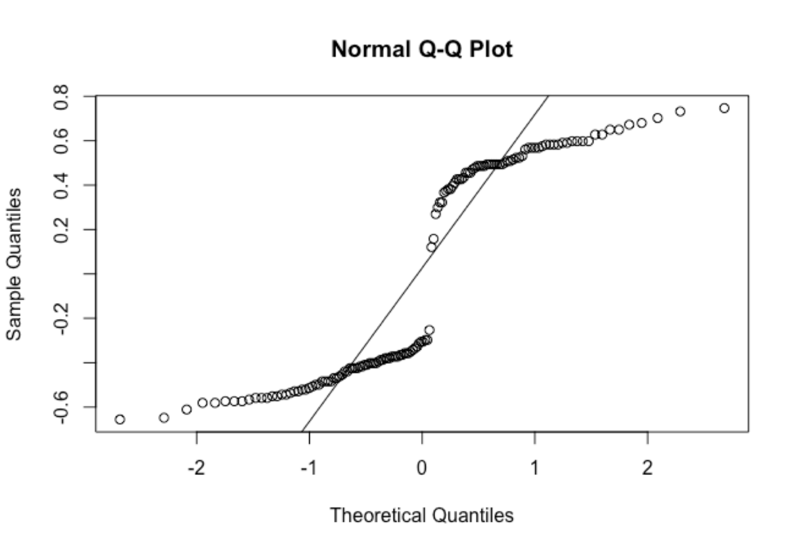
Example 1: Survival of sparrows



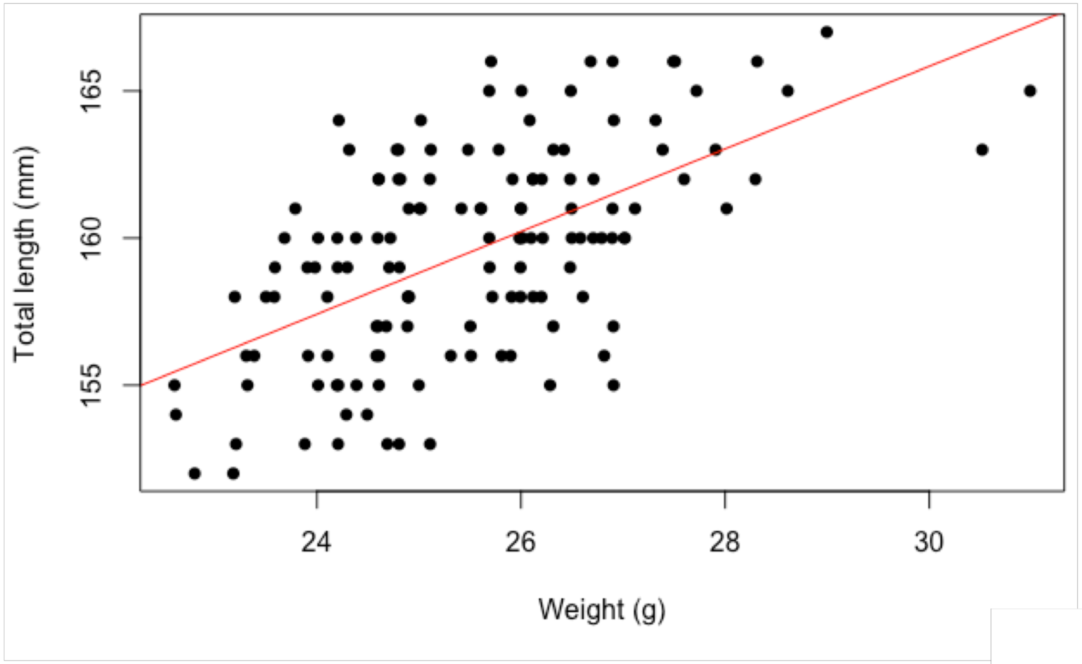
Question: How does body weight influence survival probability in sparrows?

Data: Response = whether the bird survived (1), or not (0). Explanatory = body weight in grams



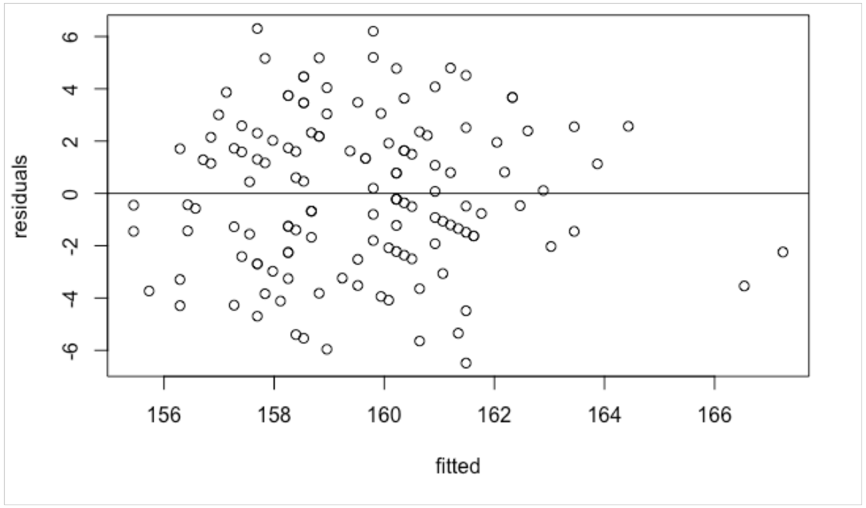


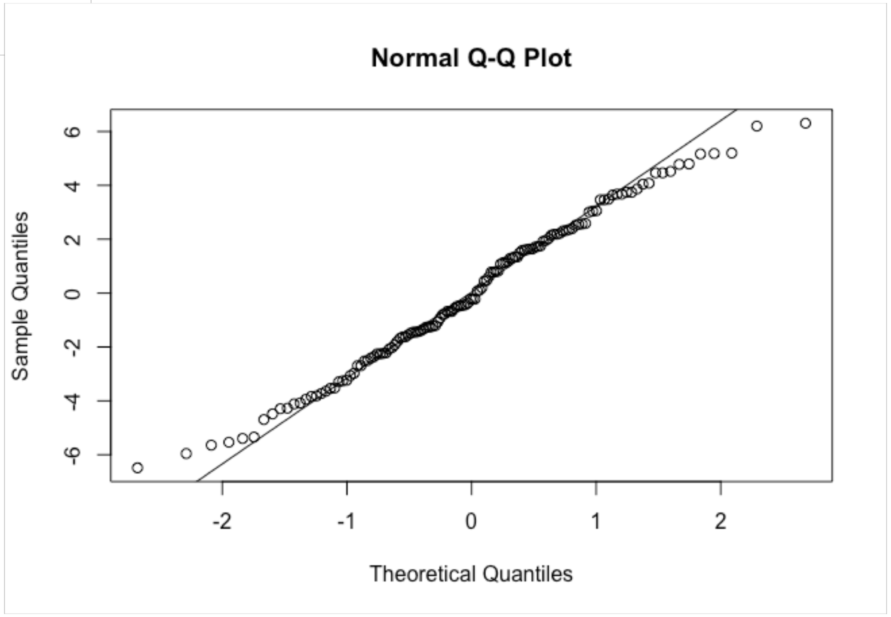
Example 2: Length and weight in sparrows



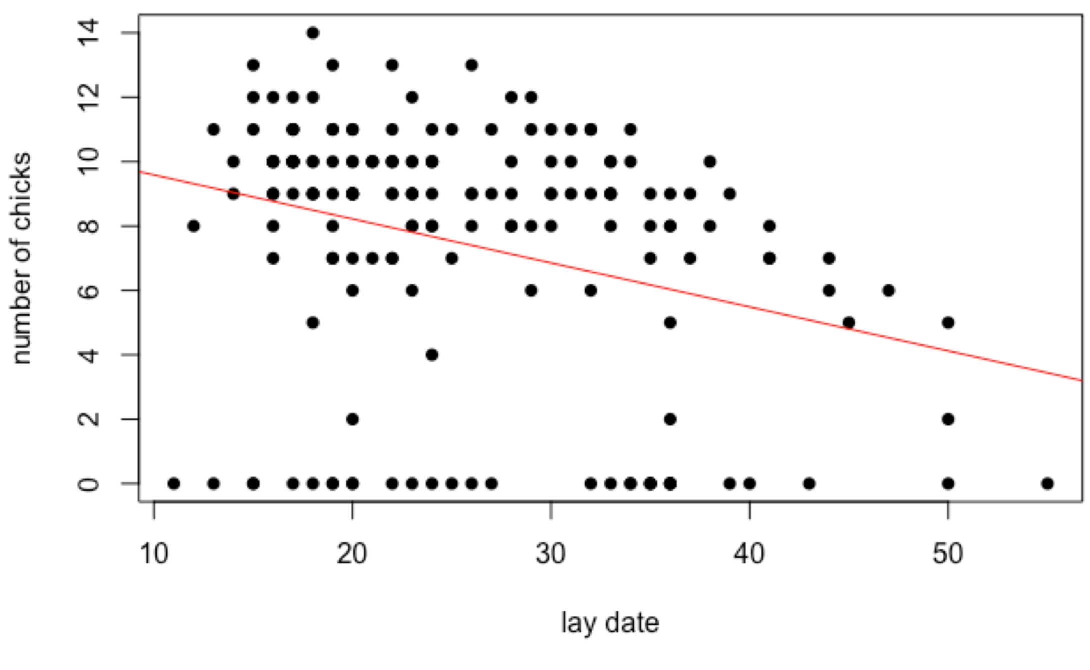
Question: How does body weight influence total length of the sparrows?

Data: Response = total length in mm. Explanatory = body weight in grams



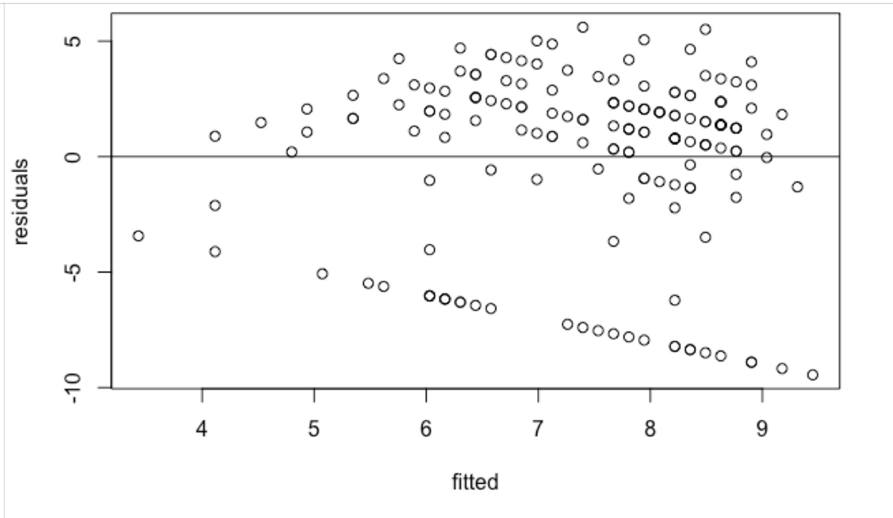


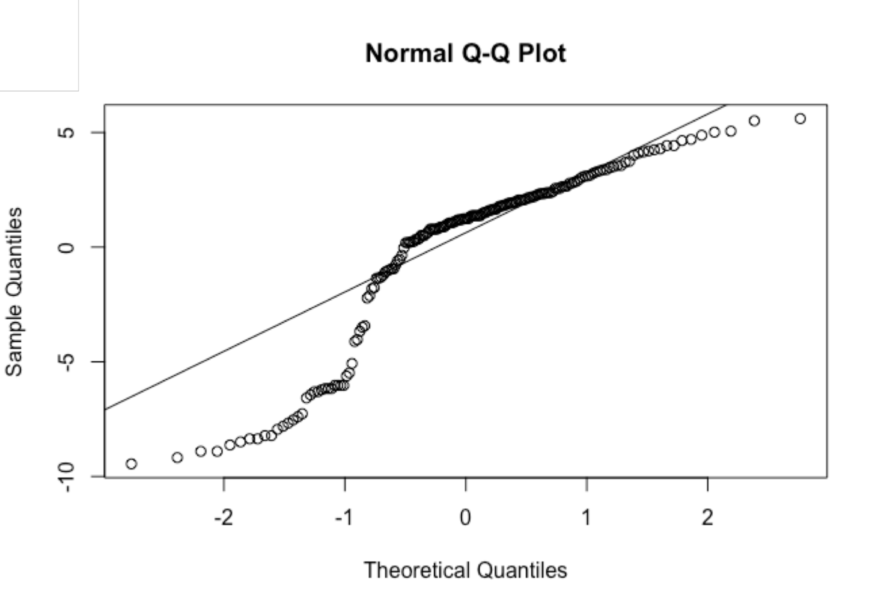
Example 3: Fledge success blue tits



Question: How does lay date influence the number of chicks that leave the nest?

Data: Response = number of chicks that fledge (leave nest alive). Explanatory = lay date (day since 1st April)





Example 1: Survival - ANSWER

Question: How does body weight influence survival probability in sparrows?

Data: Response = whether the bird survived (1), or not (0). Explanatory = body weight in grams

Is a linear model a suitable model for this data?

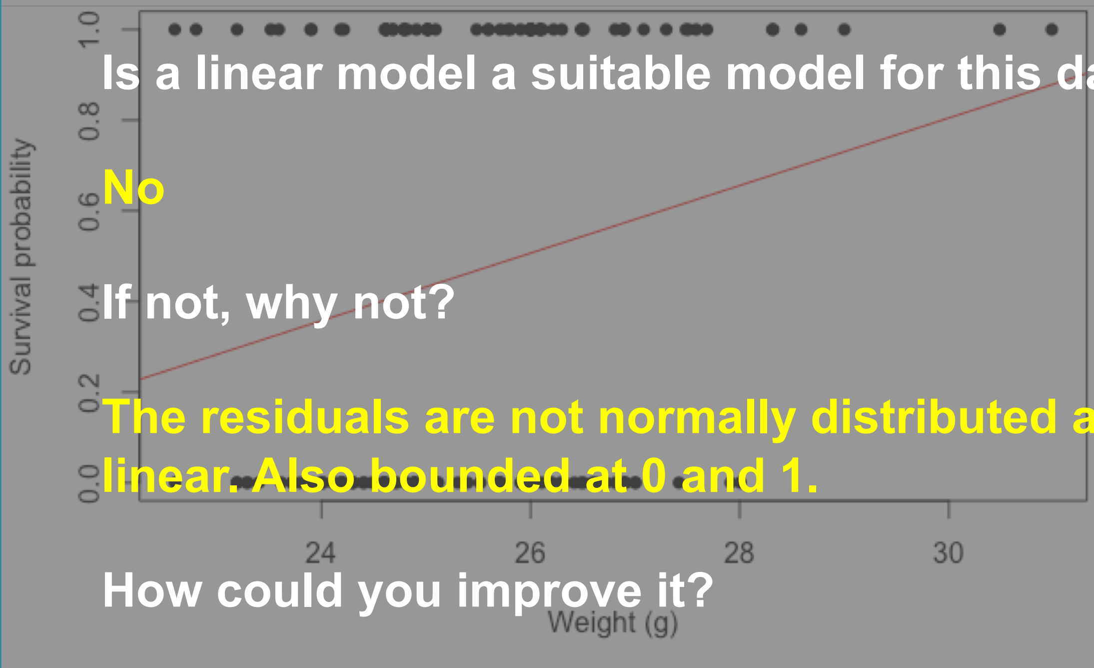
No

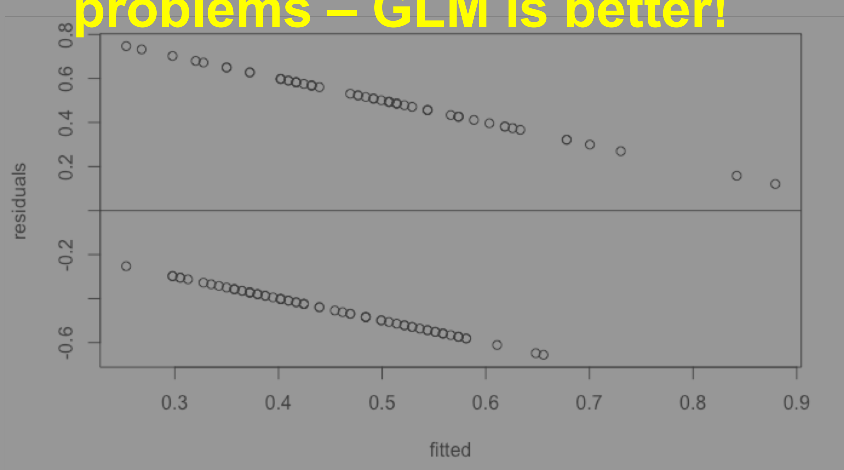
If not, why not?

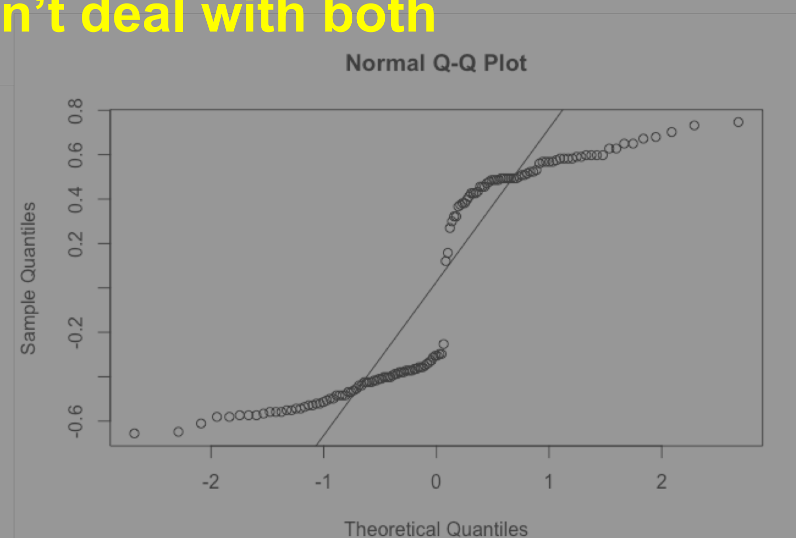
The residuals are not normally distributed and it is not linear. Also bounded at 0 and 1.

How could you improve it?

Could try transforming but won't deal with both problems – GLM is better!







Example 2: Length and weight - ANSWER

Is a linear model a suitable model for this data?

Yes

If not, why not?

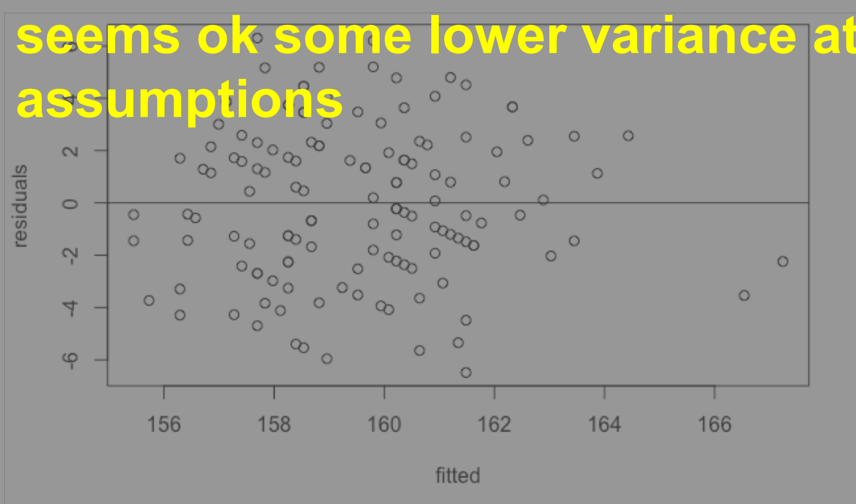
Variance seems equal for all fitted values and linearity is good

How could you improve it?

seems ok some lower variance at extremes but meets assumptions

Question: How does body weight influence total length of the sparrows?

Data: Response = total length in mm. Explanatory = body weight in grams



Example 3: Fledge success - ANSWER

Is a linear model a suitable model for this data?

No

If not, why not?

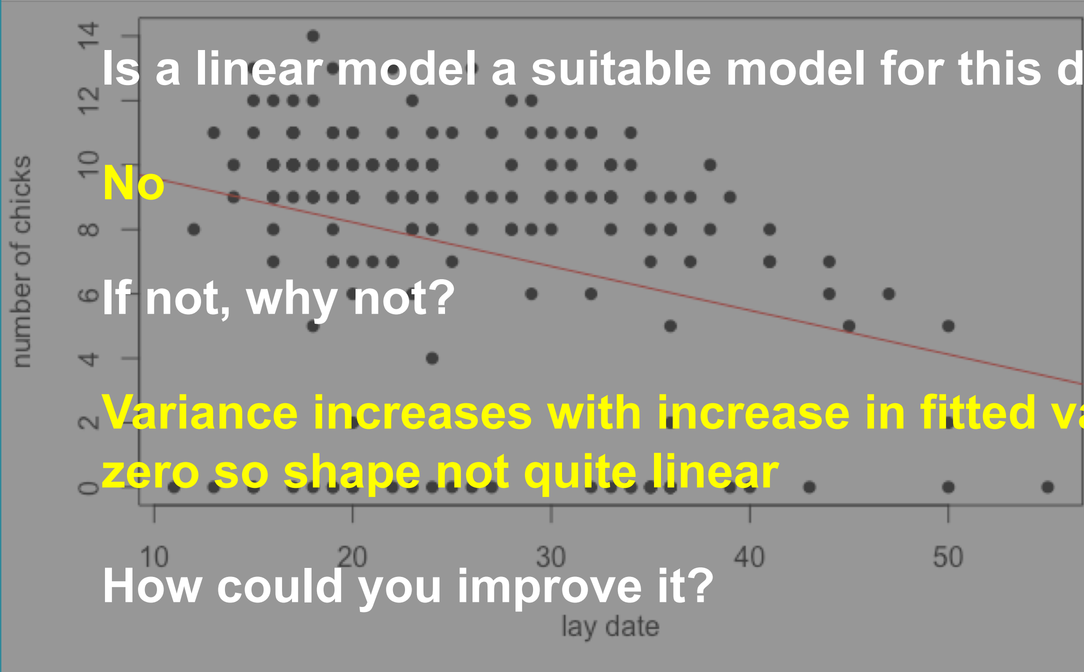
Variance increases with increase in fitted value. Bounded at zero so shape not quite linear

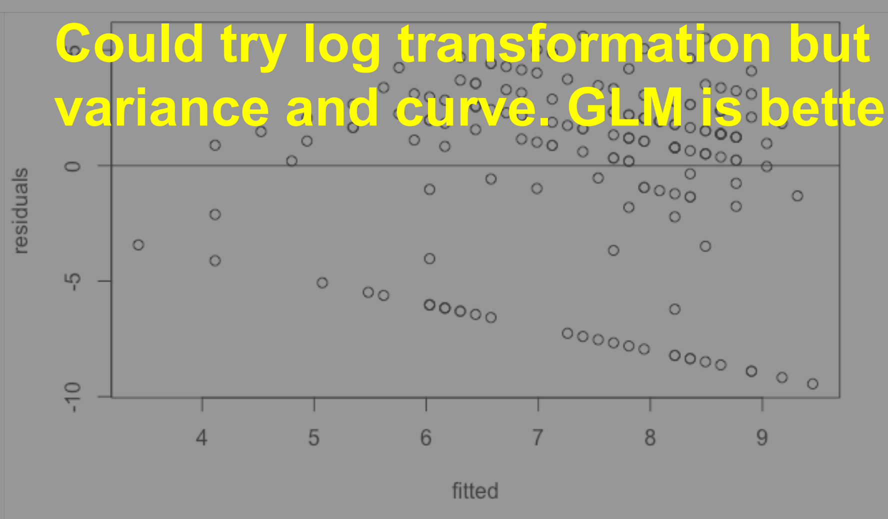
How could you improve it?

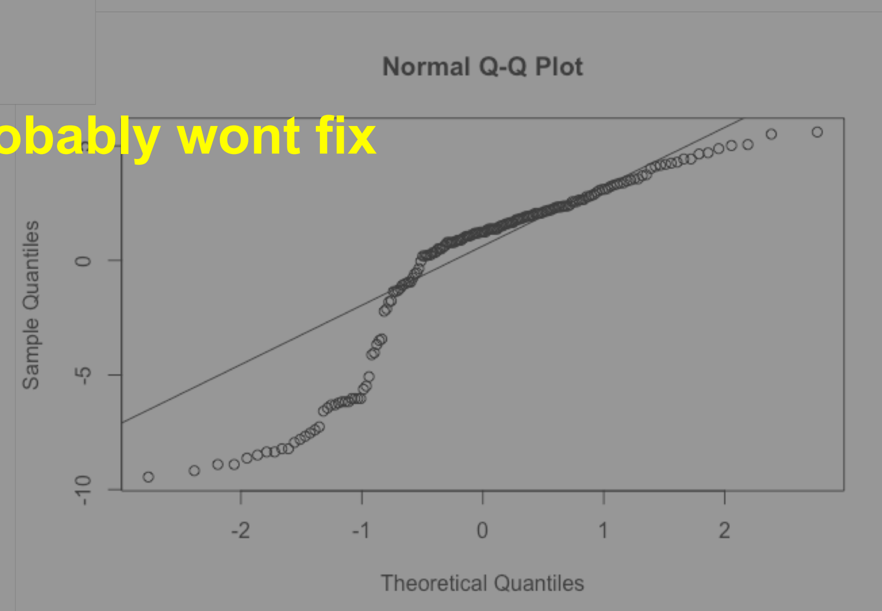
Could try log transformation but probably wont fix variance and curve. GLM is better!

Question: How does lay date influence the number of chicks that leave the nest?

Data: Response = number of chicks that fledge (leave nest alive). Explanatory = lay date (day since 1st April)







What to do with non-normality or non-linearity

Transformation of response?

Different, specialized models?

What to do with non-normality or non-linearity

Transformation of response?

Different, specialized models?

Or

Generalised linear models

A brief intro to Generalised Linear Models

Introduced in 1972 by Nelder and Wedderburn
https://docs.ufpr.br/~taconeli/CE225/Artigo.pdf

Can address variance and linearity in single model

Response unchanged

Luckily for us, very similar to lm() in R

Basis of many biological models

Key part of modern statistics!

Generalised linear models

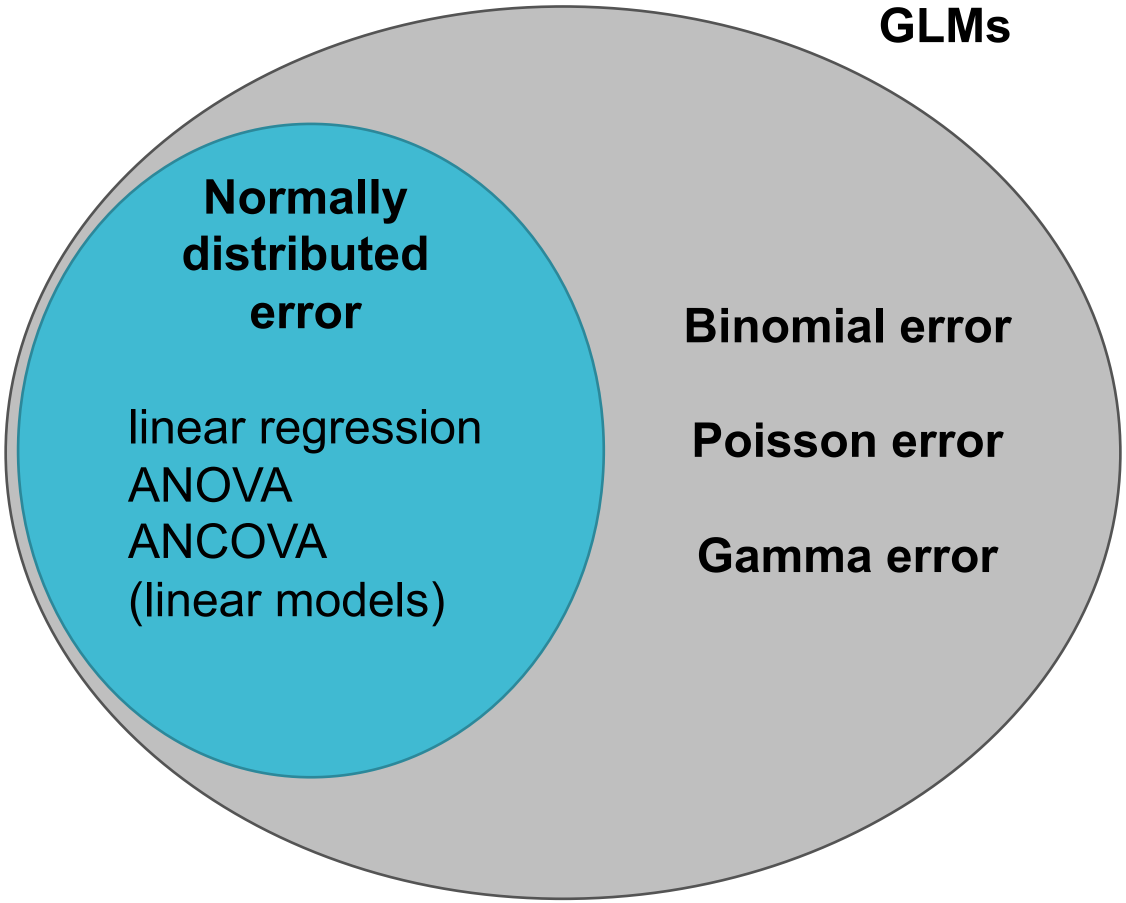
Similar to linear models but much more flexible

Normally distributed error

linear regression
ANOVA
ANCOVA
(linear models)

Generalised linear models

Similar to linear models but much more flexible



Biological examples

Clutch size

Sex ratio

Population size

Number of plants in a quadrat

Two colour morphs

Biological examples

Clutch size

Sex ratio

Population size

Number of plants in a quadrat

Two colour morphs

Counts and binary data

Exercise 3: Think of examples of non-normal data

- In your groups see if you can think of any other biological examples of non-normal data.
- This can be from your practical classes, just things you are interested in or anything else.
- Try and think of 3 examples in each group and write on white boards.
- Present one to the class.

Components of a GLM

Components of a GLM

Three main components of a GLM:

Random part

- the data (with an assumed distribution e.g. Binomial)

Systematic part

- the model for each data point (linear predictor) e.g. $\sum_j X_{ij}\beta_j$

The link function

- transforms the model (linear) onto scale of data e.g. $\log(\sum_j X_{ij}\beta_j)$

Random

Key bits to remember:

Think about the correct distribution for the data

GLM can use Normal, Binomial, Poisson, and Gamma

Different distributions use different link functions

Systematic

Key bits to remember:

This part is the same as a linear model

Link

Key bits to remember:

Different distributions use different link functions

Which you use will alter the interpretation

Connects the Systematic part to the Random data

Describes how the mean depends on the linear predictor

e.g.

$$E(Y_i) = \log(\sum_j X_{ij}\beta_j)$$

Link

Key bits to remember:

Different distributions use different link functions

Which you use will alter the interpretation

Connects the Systematic part to the Random data

Describes how the mean depends on the linear predictor

e.g.

$$E(Y_i) = \log(\sum_j X_{ij}\beta_j)$$

Expected value of Y_i (from Poisson distribution)

Link

Key bits to remember:

Different distributions use different link functions

Which you use will alter the interpretation

Connects the Systematic part to the Random data

Describes how the mean depends on the linear predictor

e.g.

$$E(Y_i) = \log(\sum_j X_{ij}\beta_j)$$

Expected value of Y_i (from Poisson distribution)

log link

Maximum likelihood and GLMs

Definitions/synonyms

Explanatory variable = covariate = predictor

Normal distribution = Gaussian distribution

Dispersion = how wide or narrow a distribution is, measured by variance or standard deviation

Parameter estimation reminder

Use maximum likelihood to estimate parameters

Likelihood is an equation that represents how the data were generated

Likelihood of parameters(θ) given the data (X):

$l(\theta|X)$ = likelihood equation for appropriate distribution

General formulation of likelihoods – not in exam

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

θ is the expected value (e.g. the mean)

y is the data

$l(\theta|y)$ is likelihood of expected value given the data

ϕ is the variance (dispersion)

a, b, and c are functions – will depend on the distribution used

Fitting GLMs in R

100m times data

Previously fit using lm() now try with glm()

Data are here:

https://www.math.ntnu.no/emner/ST2304/2019v/Week5/Times.csv

100m times data

Fit in R using glm(   )

↓

```
glm(Y ~ X, data,  family = gaussian(link=identity))
```

Fit in R using glm( )

glm(Y ~ X, data, family = gaussian(link=identity))

Exactly like lm()

Systematic part

Fit in R using glm( )

glm(Y ~ X, data, family = gaussian(link=identity))

defines the distribution you are using for the **random** part of the glm

today we use gaussian, aka Normal

Fit in R using glm( )

```
glm(Y ~ X, data,  family = gaussian(link=identity))
```

defines the **link function** to relate the **systematic** part to the **random** part

Exercise 4: Fit the GLM and interpret

- Take 100m data you used in earlier weeks
- Use code on slides before to fit a glm() and an lm() for WomenTimes
- Stick with gaussian family and identity link
- Compare the results
- Use **coef() and <u>confint.lm() or summary()</u>**

Exercise 4: ANSWER

- Results should be the same
- Can see that lm() is a special case of glm()
- But we can do much more with glm() – will start tomorrow!
- confint() on a glm uses profile likelihood

```
> coef(mod1)
(Intercept)        Year
42.18938095 -0.01573214
> coef(mod2)
(Intercept)        Year
42.18938095 -0.01573214
```

```
> round(confint.lm(mod1),2)
            2.5 % 97.5 %
(Intercept) 29.19  55.19
Year        -0.02  -0.01
> round(confint.lm(mod2),2)
            2.5 % 97.5 %
(Intercept) 29.19  55.19
Year        -0.02  -0.01
```

Lecture Outline

Recap of the course so far

What are GLMs and why do we use them?

Components of a GLM

Maximum likelihood and GLMs

Fitting in R

Lecture Outline

Recap of the course so far

We have covered many parts of the modelling process, now bringing them all together

What are GLMs and why do we use them?

Components of a GLM

Maximum likelihood and GLMs

Fitting in R

Lecture Outline

Recap of the course so far

We have covered many parts of the modelling process, now bringing them all together

What are GLMs and why do we use them?

Very flexible models that we can use for non-normal

Components of a GLM

Maximum likelihood and GLMs

Fitting in R

Lecture Outline

Recap of the course so far

We have covered many parts of the modelling process, now bringing them all together

What are GLMs and why do we use them?

Very flexible models that we can use for non-normal

Components of a GLM

Random part (data), systematic part (linear predictor), link function

Maximum likelihood and GLMs

Fitting in R

Lecture Outline

Recap of the course so far

We have covered many parts of the modelling process, now bringing them all together

What are GLMs and why do we use them?

Very flexible models that we can use for non-normal

Components of a GLM

Random part (data), systematic part (linear predictor), link function

Maximum likelihood and GLMs

General formula for the likelihood that works for all GLMs but exact functions depend on distribution of data

Fitting in R

Lecture Outline

Recap of the course so far

We have covered many parts of the modelling process, now bringing them all together

What are GLMs and why do we use them?

Very flexible models that we can use for non-normal data

Components of a GLM

Random part (data), systematic part (linear predictor), link function

Maximum likelihood and GLMs

General formula for the likelihood that works for all GLMs but exact functions depend on distribution of data

Fitting in R

Use glm(), very similar to lm() but with extra arguments for link random part and link function