

Generalised Linear Models (GLM): Part 1

Administration points

Lecture Outline Part 1

What are GLMs and why do we use them?

- EX1: Non-normal data

Components of a GLM

- EX2: Examples of non-normal data

Maximum likelihood and GLMs

Fitting in R

- EX3: Fit in R

Lecture Outline Part 2

More on the random part

- EX4: Choose a distribution

Practice with Poisson GLM

- EX5: Fit a Poisson GLM in R

Interpretation and GLMS

Checking model fit in GLMS

- EX6: Interpret and check

Reading

Chapter 8 – The New Statistics with R

Part 1

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

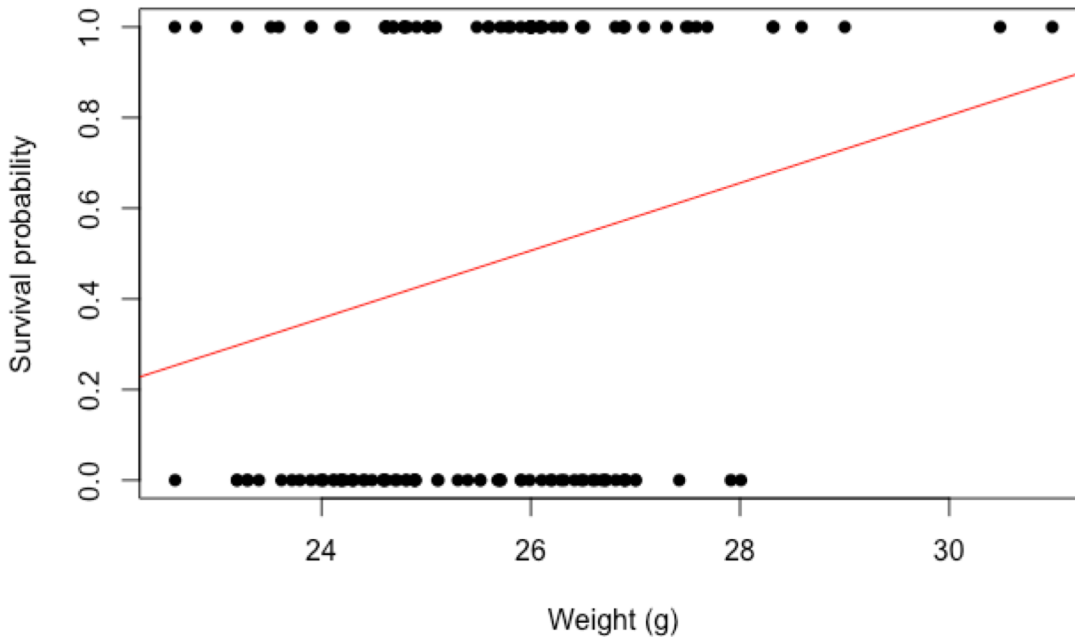
Assumptions:

- straight line (**linearity**)
- errors are independent
- errors have same variance (**homoscedasticity**)
- errors are normally distributed and zero mean
- no outliers

Exercise 1: Is a linear model appropriate?

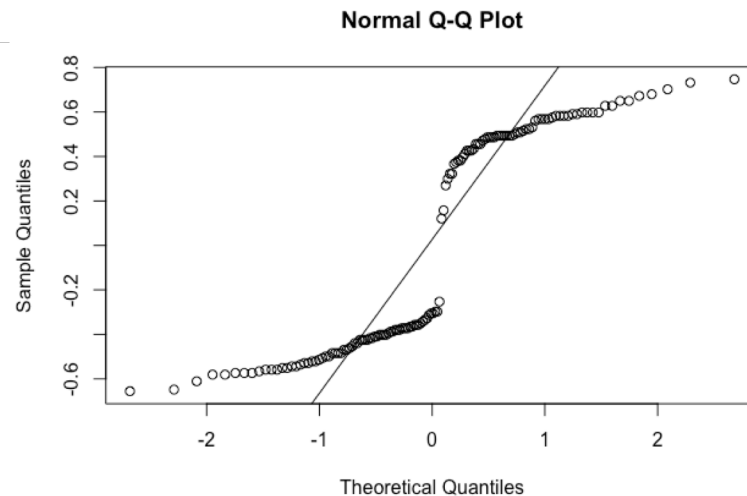
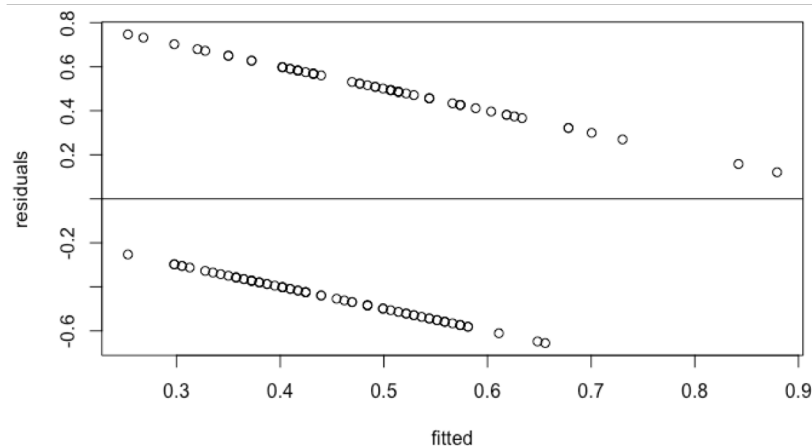
- Part A

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 1: Survival - ANSWER

Is a linear model a suitable model for this data?

No

If not, why not?

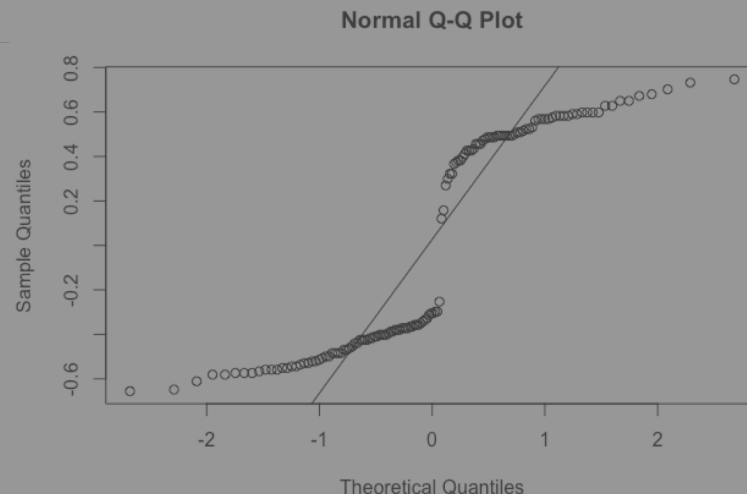
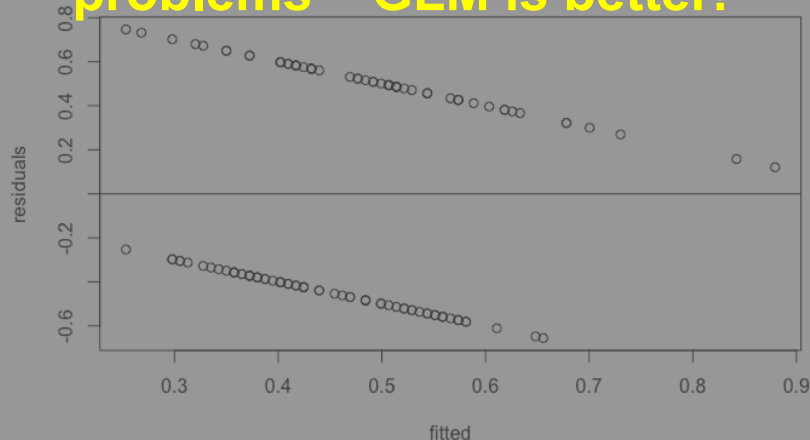
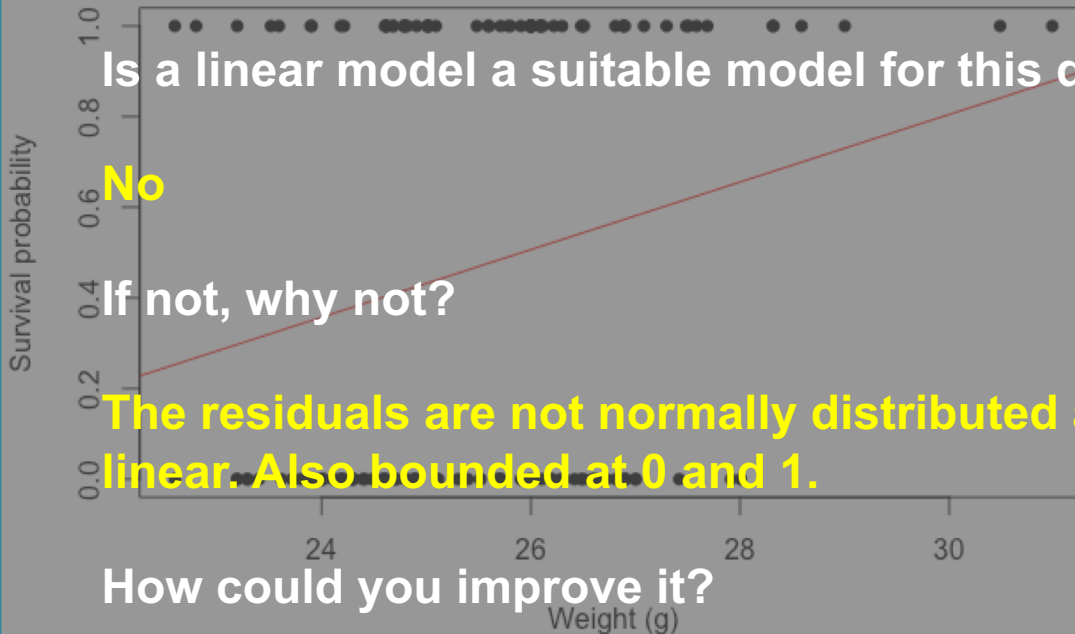
The residuals are not normally distributed and it is not linear. Also bounded at 0 and 1.

How could you improve it?

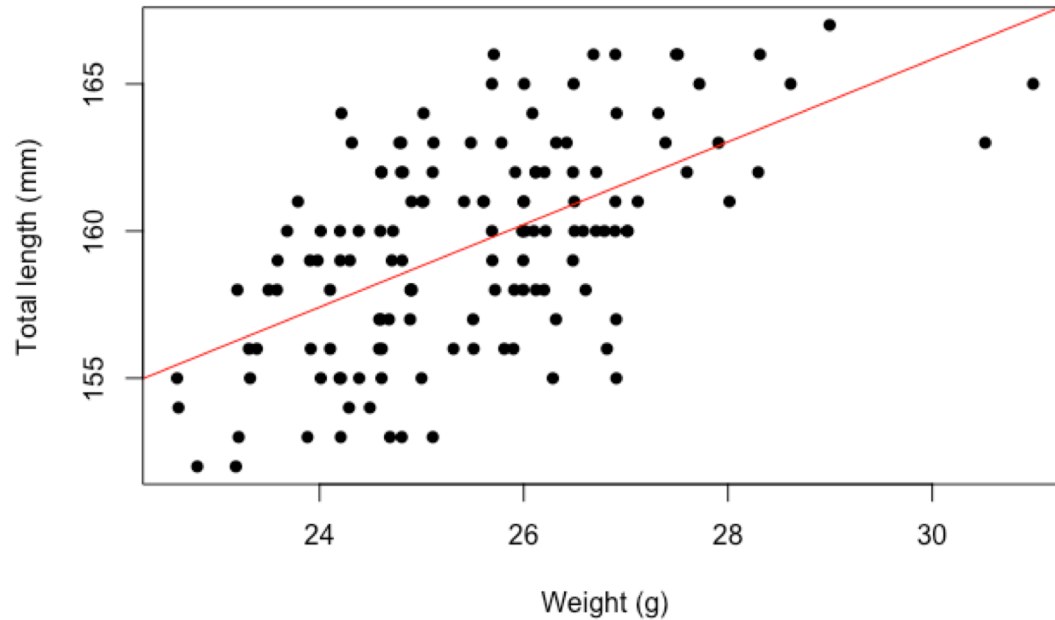
Could try transforming but won't deal with both problems – GLM is better!

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

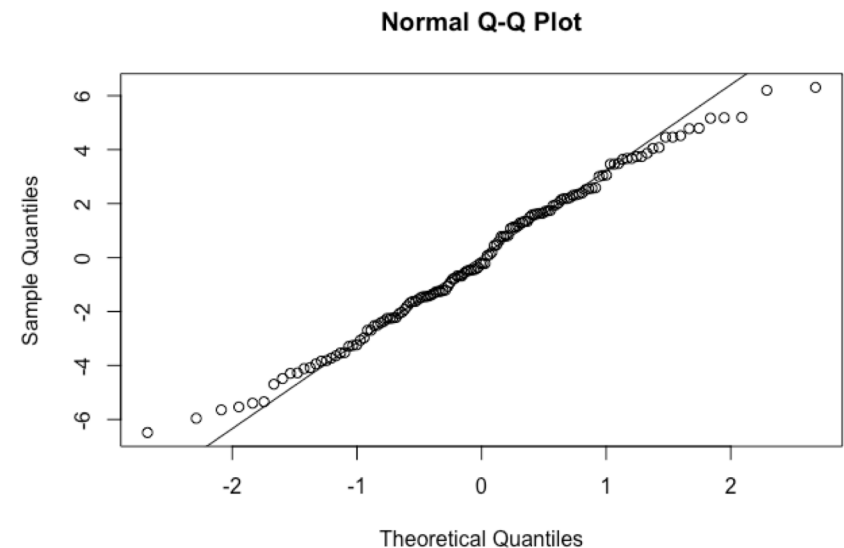
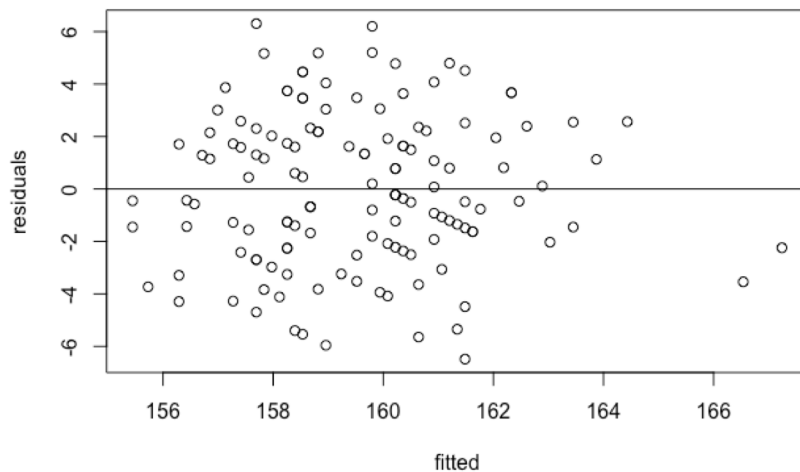


Example 2: Length and weight in sparrows



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

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



Example 2: Length and weight - ANSWER

Is a linear model a suitable model for this data?

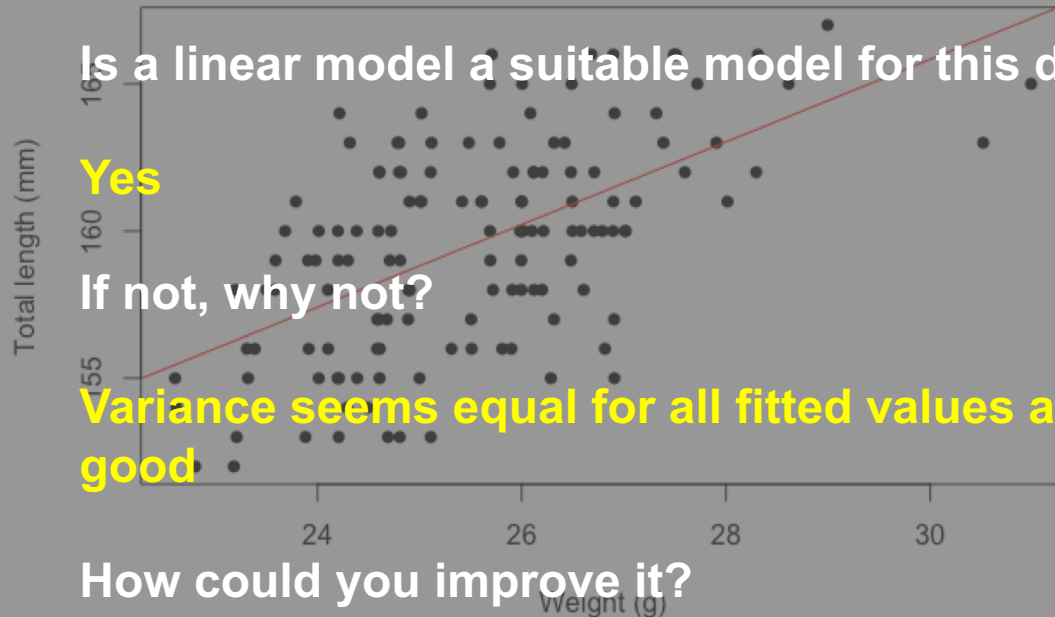
Yes

If not, why not?

Variance seems equal for all fitted values and linearity is good

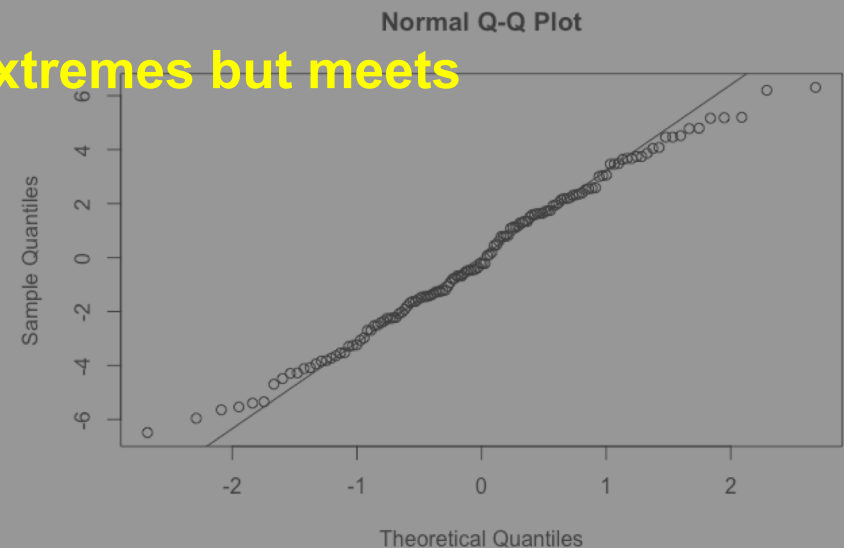
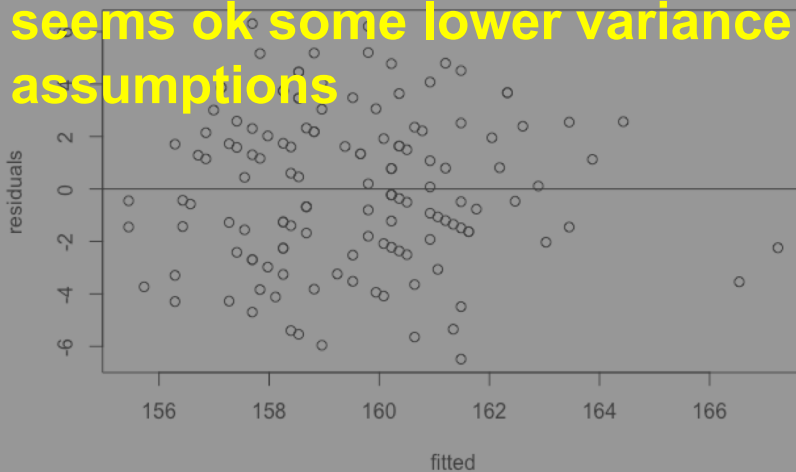
How could you improve it?

seems ok some lower variance at extremes but meets assumptions

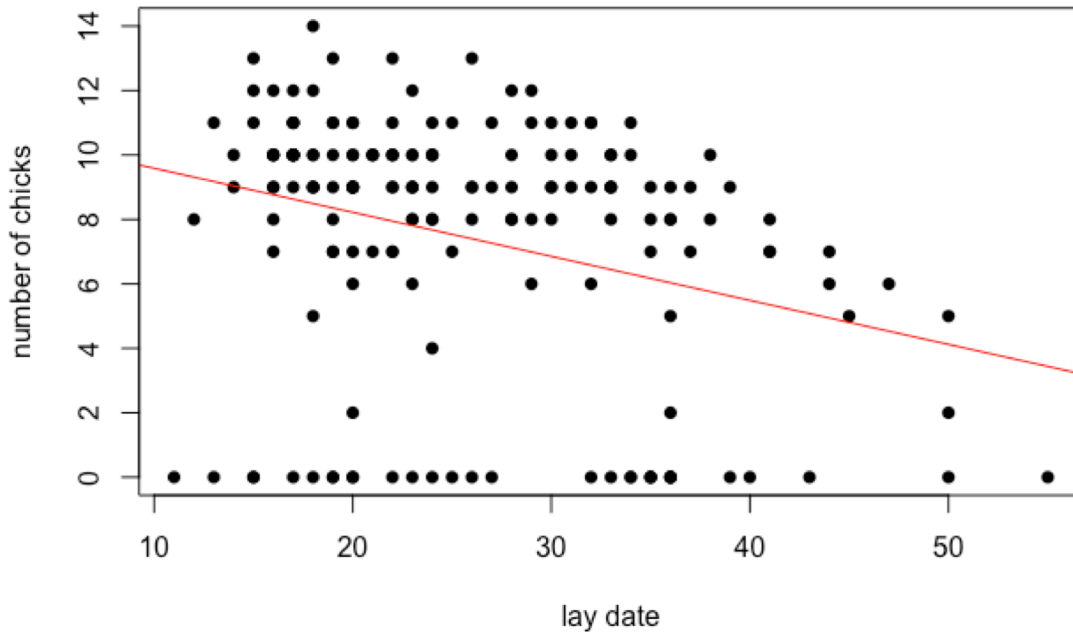


Question: How does body weight influence total length of the sparrows?

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

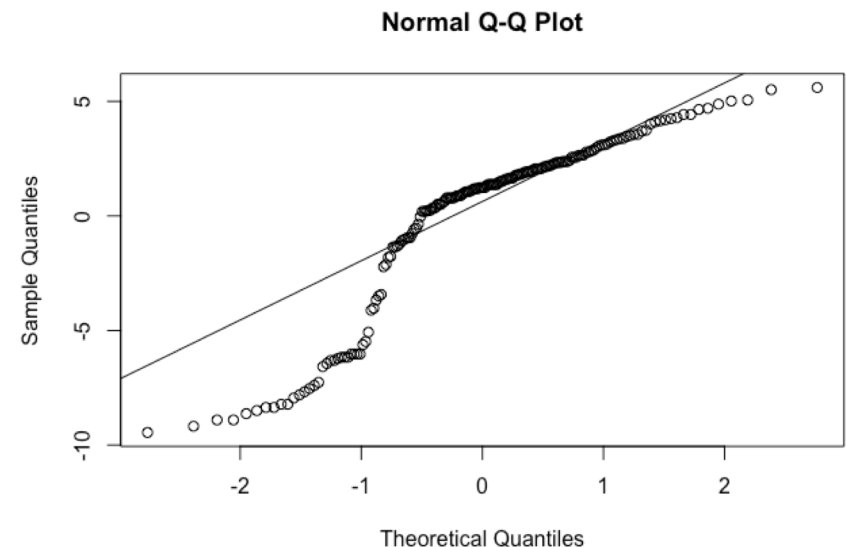
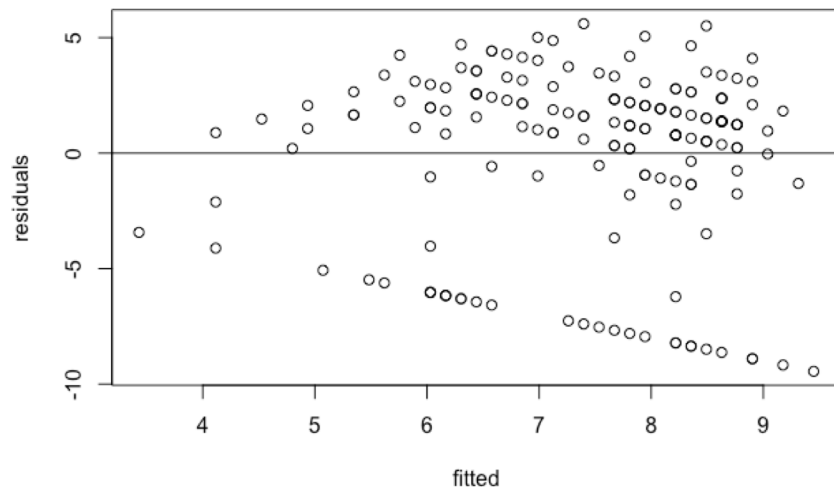


Example 3: Fledge success blue tits



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 3: Fledge success - ANSWER

Is a linear model a suitable model for this data?

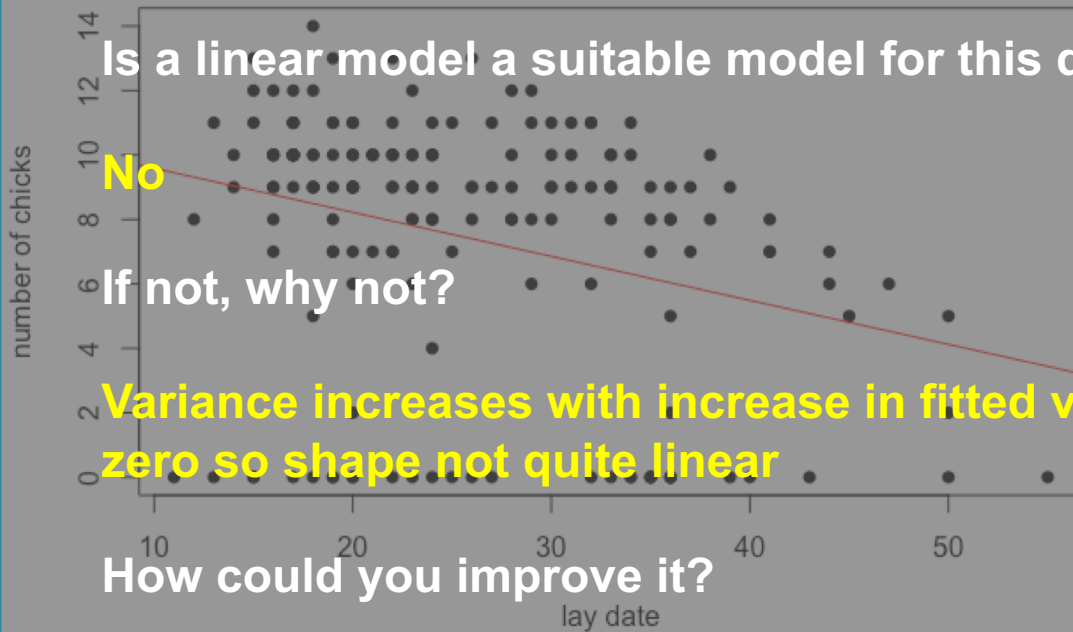
No

If not, why not?

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

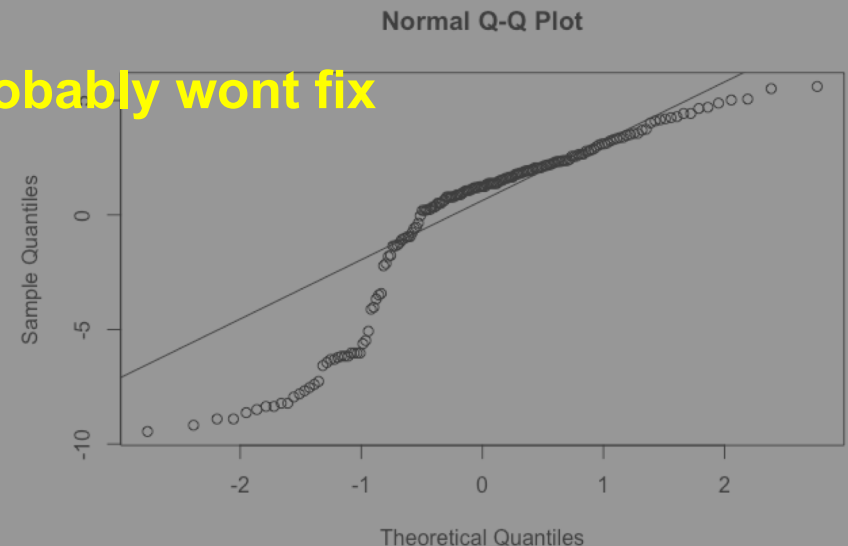
How could you improve it?

Could try log transformation but probably won't 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

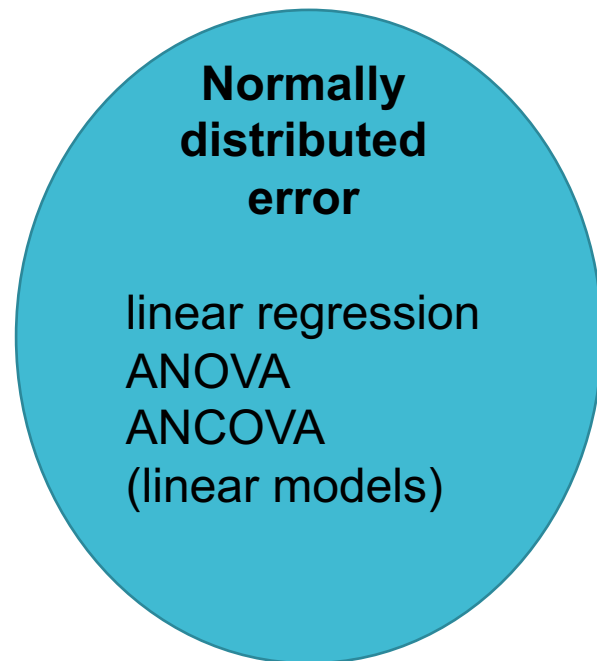
Luckily for us, very similar to `lm()` in R

Basis of many biological models

Key part of modern statistics!

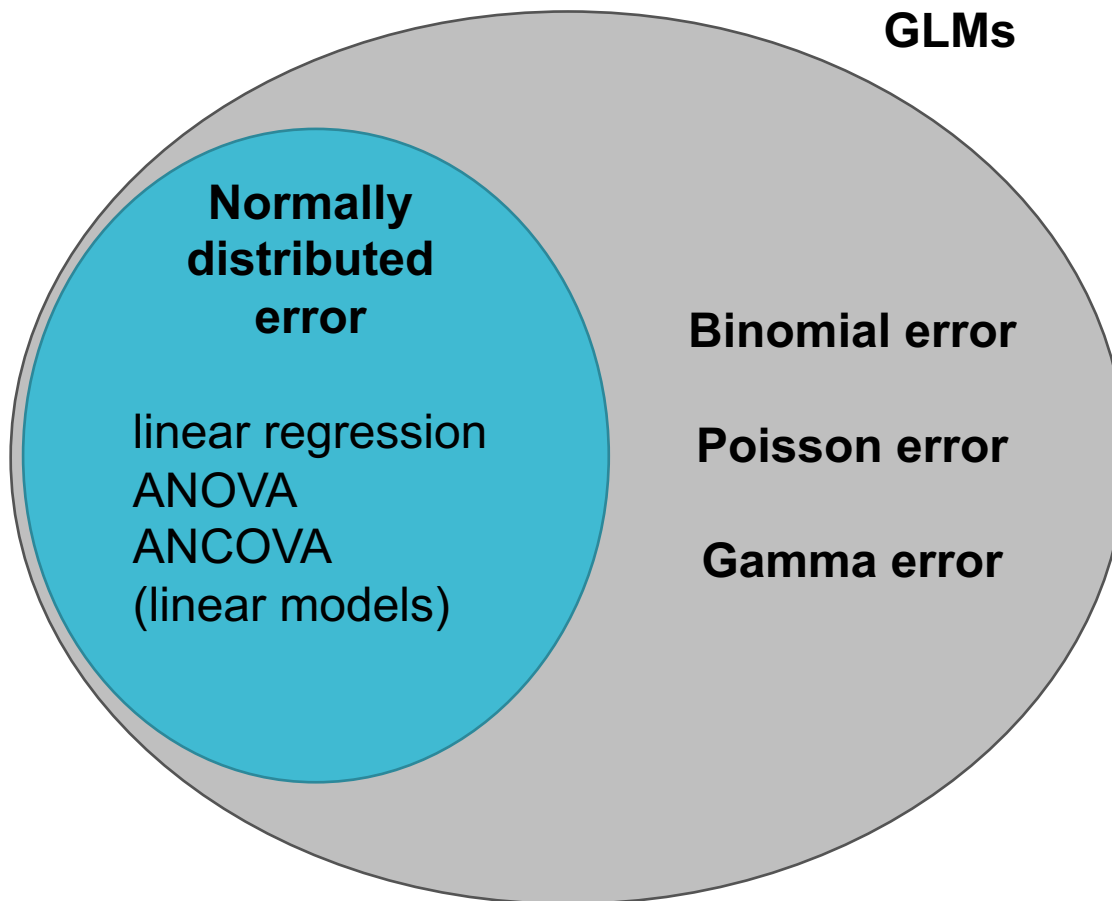
Generalised linear models

Similar to linear models but much more flexible



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 2: 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.
- Share one with 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

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.

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

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.


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

Expected value of Y_i
(from Poisson
distribution)

Key bits to remember:

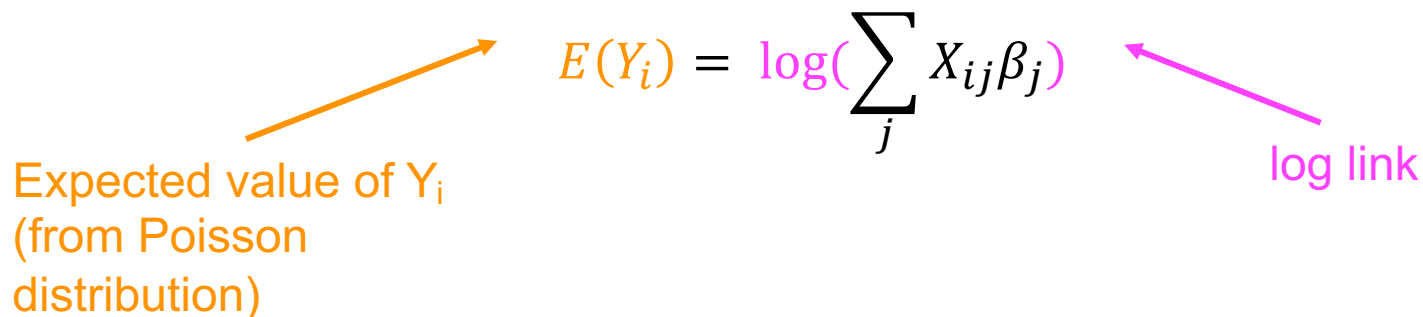
Different distributions use different link functions

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.



The diagram shows the equation $E(Y_i) = \log(\sum_j X_{ij}\beta_j)$. An orange arrow points from the text 'Expected value of Y_i (from Poisson distribution)' to the $E(Y_i)$ term. A purple arrow points from the text 'log link' to the \log function.

Expected value of Y_i
(from Poisson
distribution)

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

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

Formally it is the probability of the data given the parameter but also the likelihood of parameter given data (annoying – we know!)

$l(\theta|X)$ = 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 3: Fit the GLM and interpret

- Part B

Exercise 3: 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 – Part 1

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

Part 2

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. $\log(\sum_j X_{ij}\beta_j)$

More on the Random part

Which distribution do I use?

GLM can use Normal, Binomial, Poisson, Gamma, and some quasi- distributions

quasi = almost

Which distribution do I use?

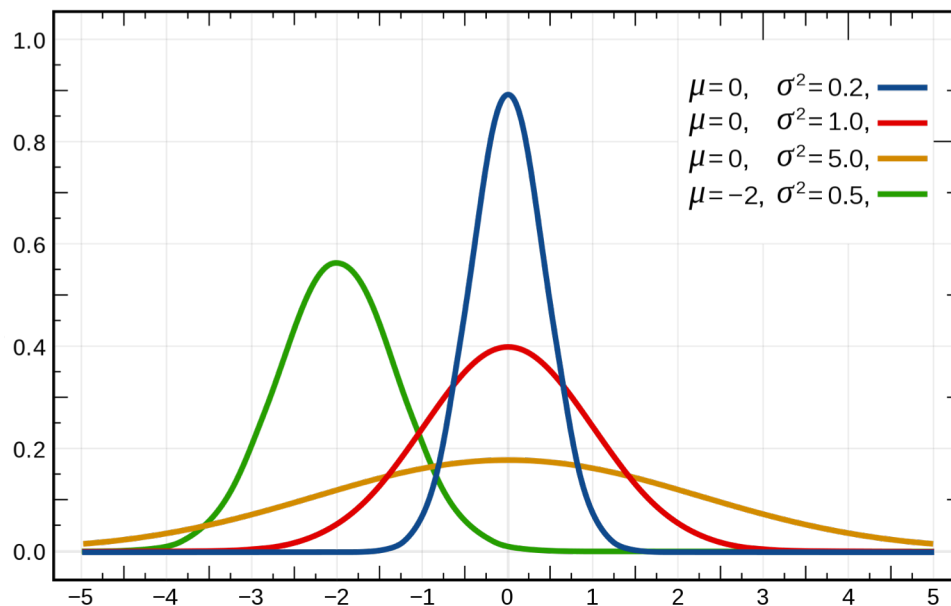
GLM can use **Normal**, **Binomial**, **Poisson**, Gamma, and some quasi- distributions

The Normal Distribution

Parameters: mean (μ) and variance (σ^2)

Properties: Continuous, symmetrical around mean, single mode

Examples: height, biomass, running times



The Binomial Distribution

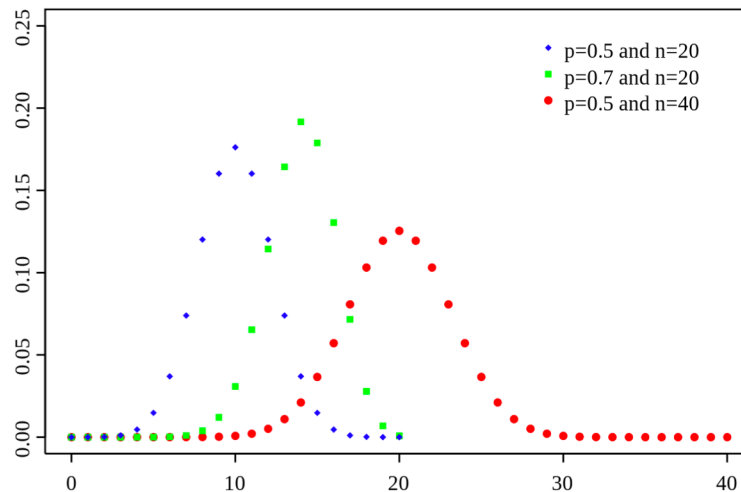
Parameters: probability (p)

mean = np (n = number of successes)

variance = $np(1 - p)$

Properties: Gives probability of success from two possible outcomes (bounded between 0 and 1)

Examples: survival, sex ratio, land or sea



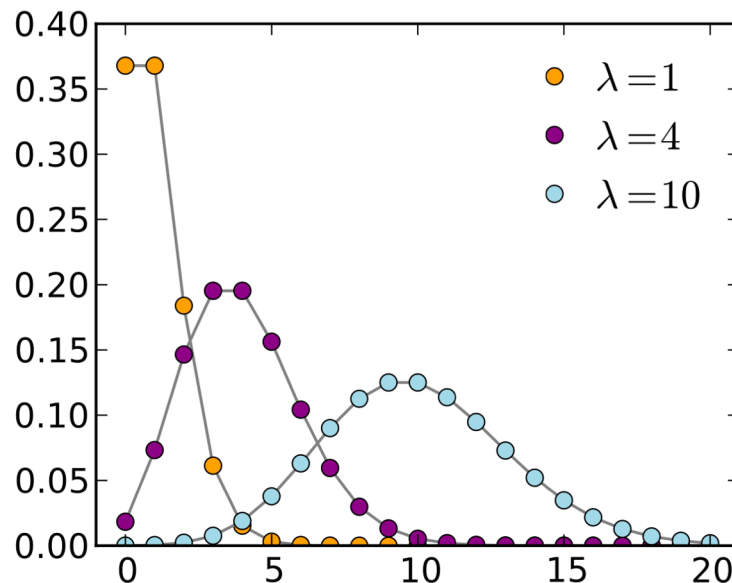
* Picture from Wikipedia

The Poisson Distribution

Parameters: mean (λ)
variance = mean

Properties: Successes in time or space (counts),
discrete, positive

Examples: number of plants, number of eggs,
population size



* Picture from Wikipedia

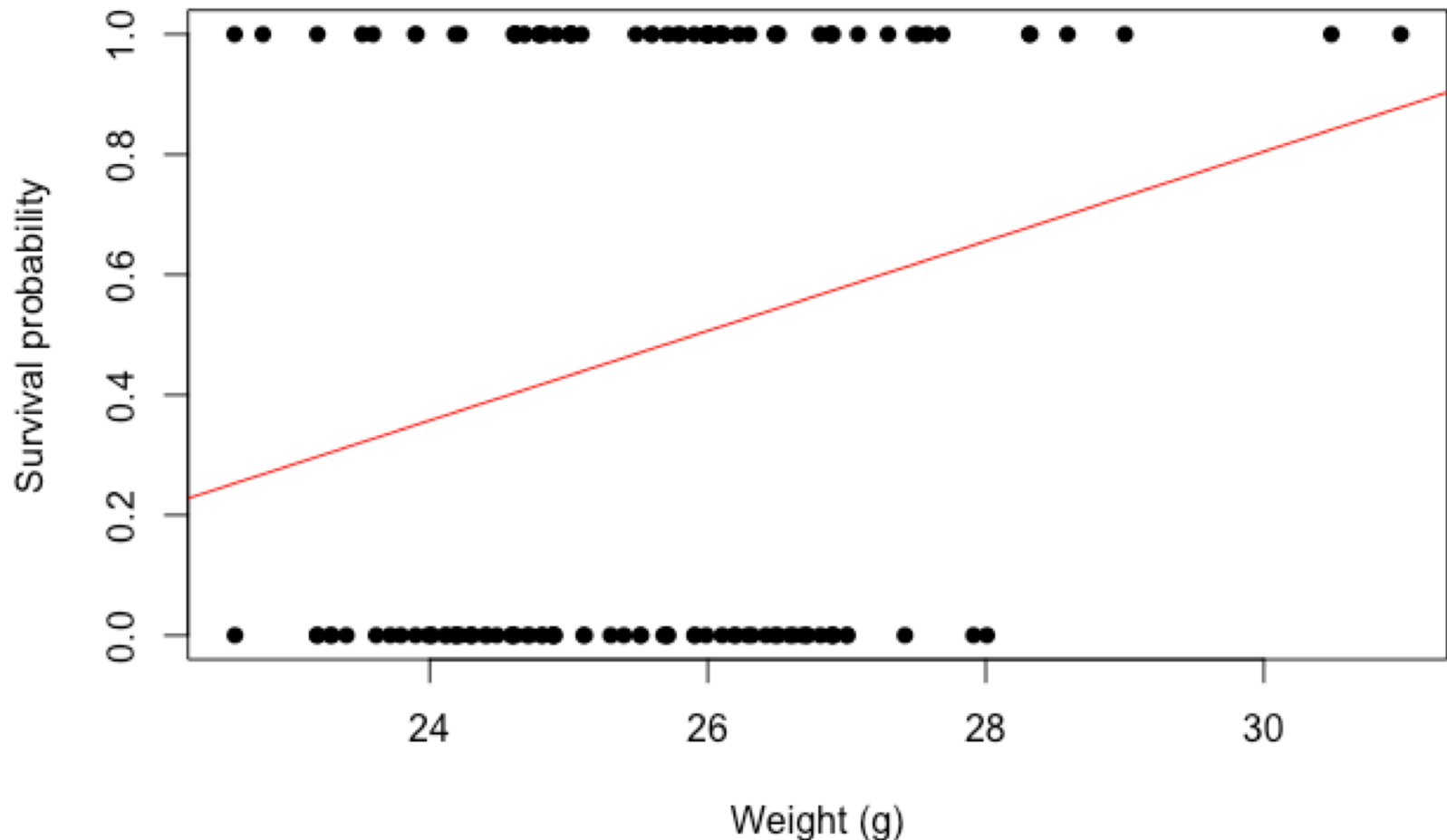
Exercise 4: Which distribution?

- Part C

Example 1: Survival of sparrows

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

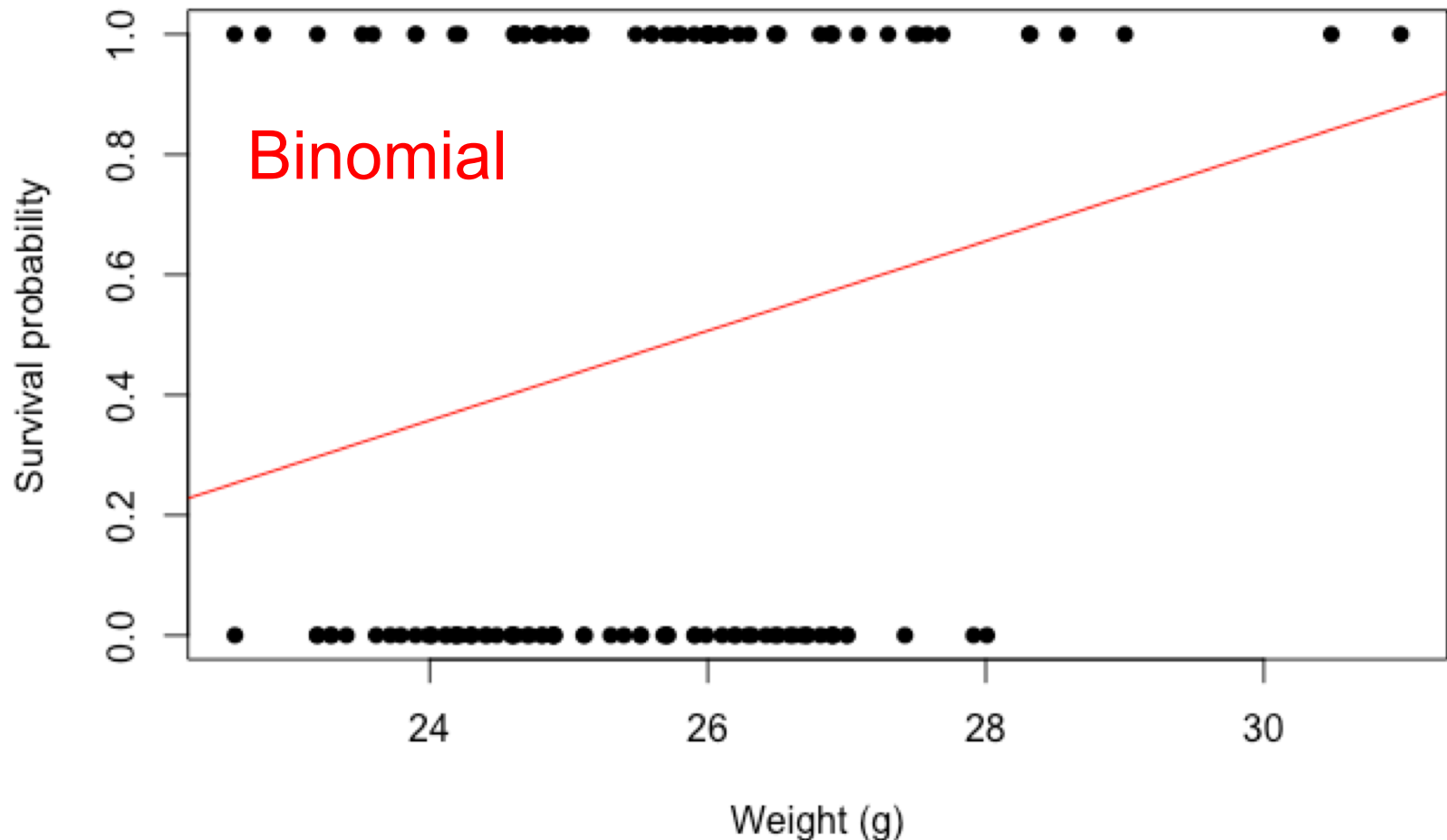
Data: Response = whether the bird survived (1), or not (0). Explanatory = body weight in grams



Example 1: ANSWER

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

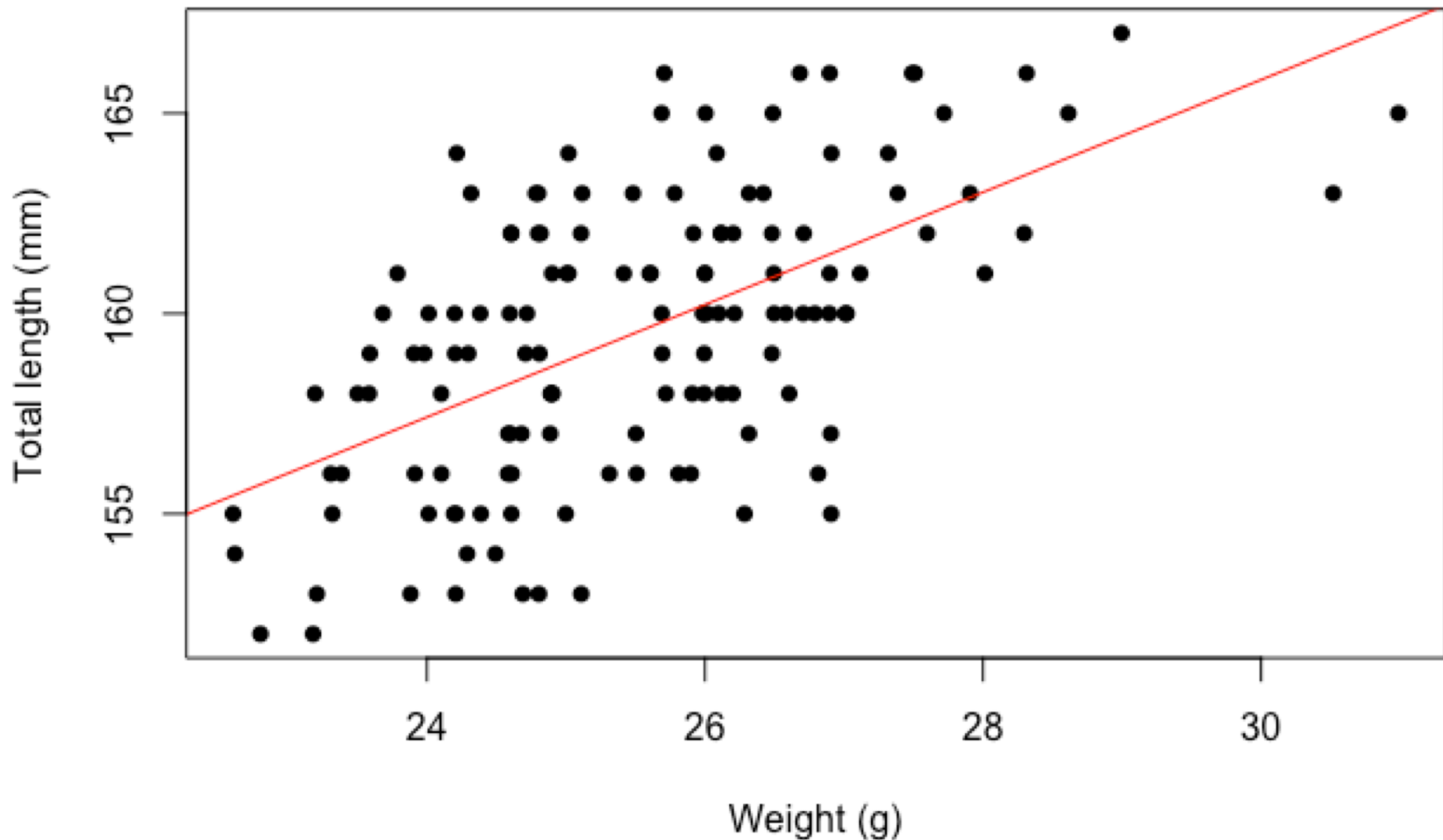
Data: Response = whether the bird survived (1), or not (0). Explanatory = body weight in grams



Example 2: Length and weight in sparrows

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

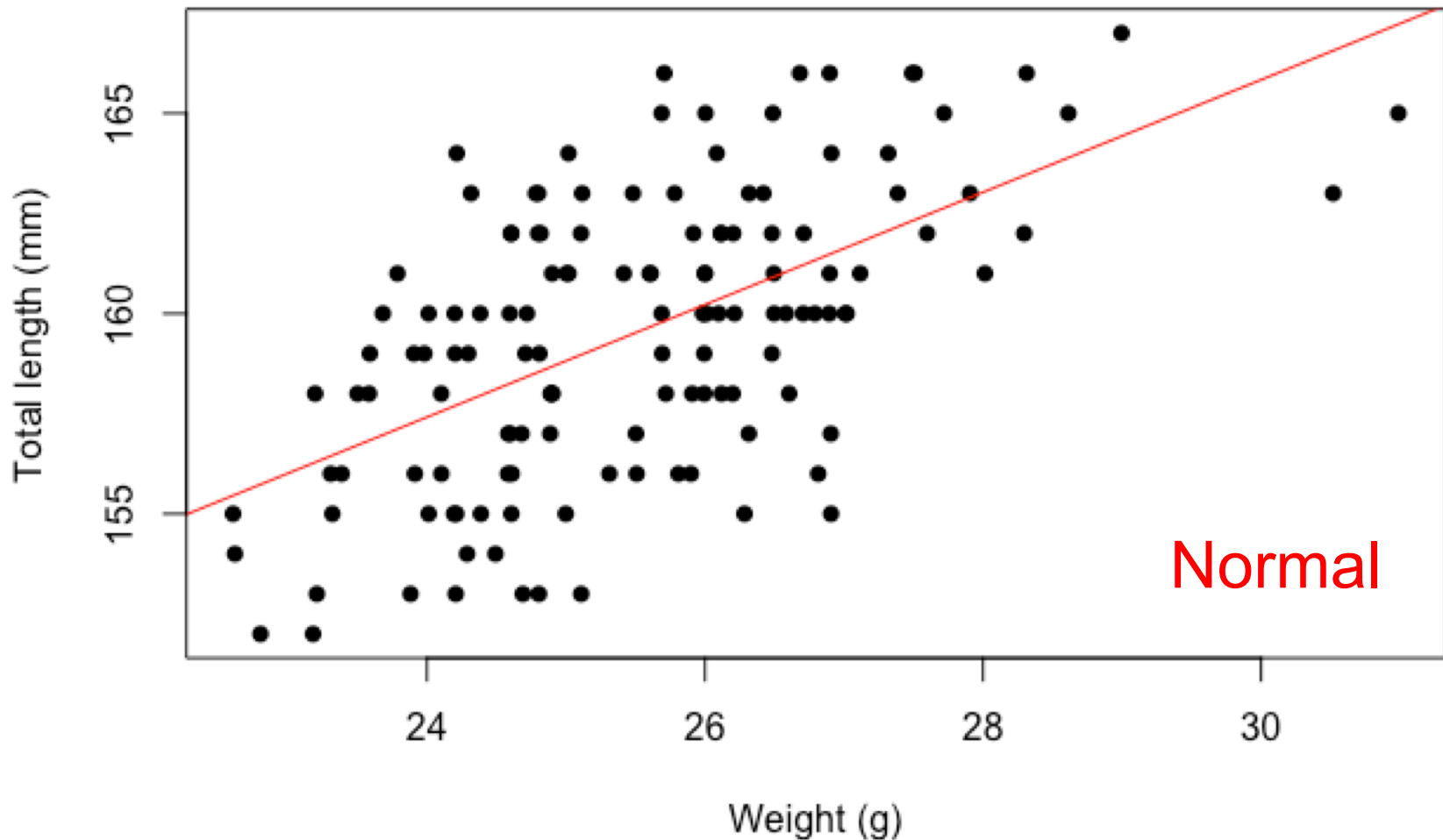
Data: Response = total length in mm. Explanatory = body weight in grams



Example 2: ANSWER

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

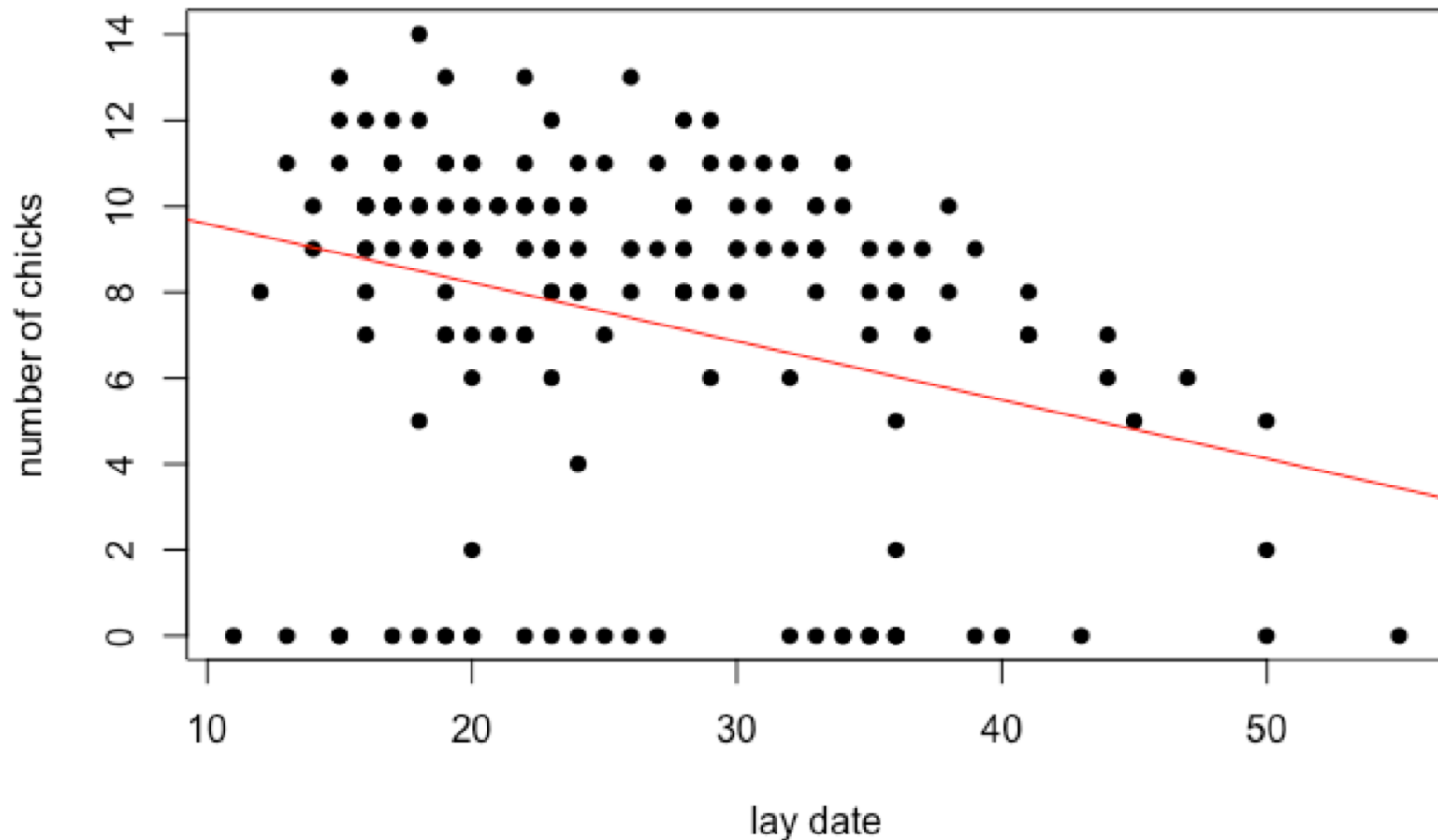
Data: Response = total length in mm. Explanatory = body weight in grams



Example 3: Fledge success blue tits

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

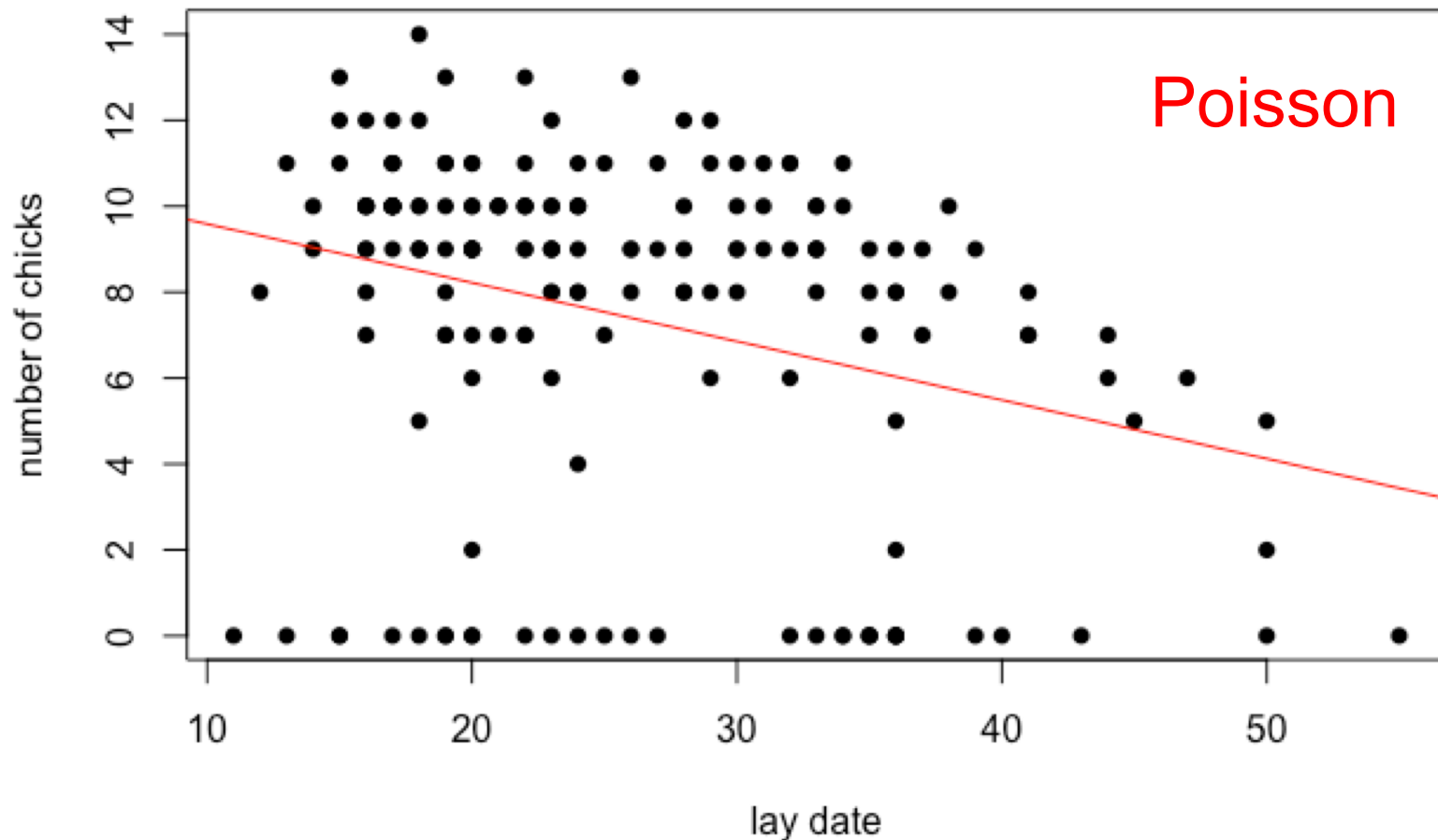
Data: Response = number of chicks that fledge (leave nest alive). Explanatory = lay date (day since 1st April)



Example 3: ANSWER

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)



Link functions and distributions

| Family (distribution) | Default link function (canonical) | Other common link functions |
|--------------------------|---|--------------------------------|
| Gaussian | Identity (μ) | |
| Binomial | Logit ($\log(\frac{\mu}{1-\mu})$) | Probit, cloglog |
| Poisson | Log ($\log(\mu)$) | Identity |

Basics of a Poisson GLM in R (log-linear model)

Does location of nest influence clutch size?

Phoenix clutch size

Mythical bird. Counted eggs in nests.

Counted eggs in two places Scotland and Norway.

Want to see if the location of the nest influences the number of eggs laid.



The likelihood

General likelihood for GLM:

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

Poisson likelihood:

$$l(\mu|r) = -\mu + r \log(\mu) - \log(r!)$$

The likelihood

General likelihood for GLM:

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

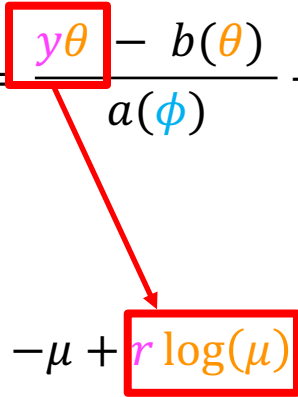
Poisson likelihood:

$$l(\mu|r) = -\mu + r \log(\mu) - \log(r!)$$

$$a(\phi) = 1$$

The likelihood

General likelihood for GLM:

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


Poisson likelihood:

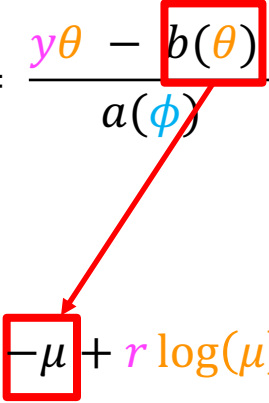
$$l(\mu|r) = -\mu + r \log(\mu) - \log(r!)$$

$$a(\phi) = 1$$

$\theta = \log(\mu)$ and (μ) comes from the linear equation $\alpha + \beta X_i$

The likelihood

General likelihood for GLM:

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


Poisson likelihood:

$$l(\mu|r) = -\mu + r \log(\mu) - \log(r!)$$


$$a(\phi) = 1$$

$$\theta = \log(\mu)$$

$$b(\theta) = -e^{\theta} = -e^{\log(\mu)} = -\mu$$

The likelihood

General likelihood for GLM:

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


Poisson likelihood:

$$l(\mu|r) = -\mu + r \log(\mu) - \log(r!)$$

$$a(\phi) = 1$$

$$\theta = \log(\mu)$$

$$b(\theta) = -e^{\theta} = -e^{\log(\mu)} = -\mu$$

$$c(y, \phi) = -\log(r!), \text{ where } r = \text{the data we observed, the count}$$

The likelihood

General likelihood for GLM:

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

Poisson likelihood:

$$l(\mu|r) = -\mu + r \log(\mu) - \log(r!)$$

$$a(\phi) = 1$$

$$\theta = \log(\mu)$$

$$b(\theta) = -e^\theta = -e^{\log(\mu)} = -\mu$$

$$c(y, \phi) = -\log(r!)$$

Yay, it fits the same format!

The likelihood

General likelihood for GLM:

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

Poisson likelihood:

$$l(\mu|r) = -\mu + r \log(\mu) - \log(r!)$$

$$a(\phi) = 1$$

$$\theta = \log(\mu)$$

$$b(\theta) = -e^\theta = -e^{\log(\mu)} = -\mu$$

$$c(y, \phi) = -\log(r!)$$

Also – we can see our link function

Exercise 5: Fit the Poisson GLM in R

- Part D

Exercise 5: ANSWER

```
> model1 <- glm(ClutchSize~Location, data = phoenix, family=poisson(link=log))  
> coef(model1)  
      (Intercept) LocationScotland  
          1.098612          -1.272966
```

Interpreting

Exercise 5: Interpreting

- Continue Part D.

Exercise 5: ANSWER

```
> coef(model1)
      (Intercept) LocationScotland
           1.098612           -1.272966
> confint(model1)
Waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept)    0.9341921  1.2544806
LocationScotland -1.6270659 -0.9408995
```

But what do they mean?

Exercise 5: ANSWER

```
> coef(model1)
```

| | |
|-------------|------------------|
| (Intercept) | LocationScotland |
| 1.098612 | -1.272966 |

Mean for Norge

```
> confint(model1)
```

Waiting for profiling to be done...

| | | |
|------------------|------------|------------|
| | 2.5 % | 97.5 % |
| (Intercept) | 0.9341921 | 1.2544806 |
| LocationScotland | -1.6270659 | -0.9408995 |

Exercise 5: ANSWER

```
> coef(model1)
```

| | |
|-------------|------------------|
| (Intercept) | LocationScotland |
| 1.098612 | -1.272966 |

Difference between mean
Norge and mean Scotland

```
> confint(model1)
```

Waiting for profiling to be done...

| | | |
|------------------|------------|------------|
| | 2.5 % | 97.5 % |
| (Intercept) | 0.9341921 | 1.2544806 |
| LocationScotland | -1.6270659 | -0.9408995 |

Exercise 5: ANSWER

```
> coef(model1)
```

| | |
|-------------|------------------|
| (Intercept) | LocationScotland |
| 1.098612 | -1.272966 |

But – we need to
remember the link

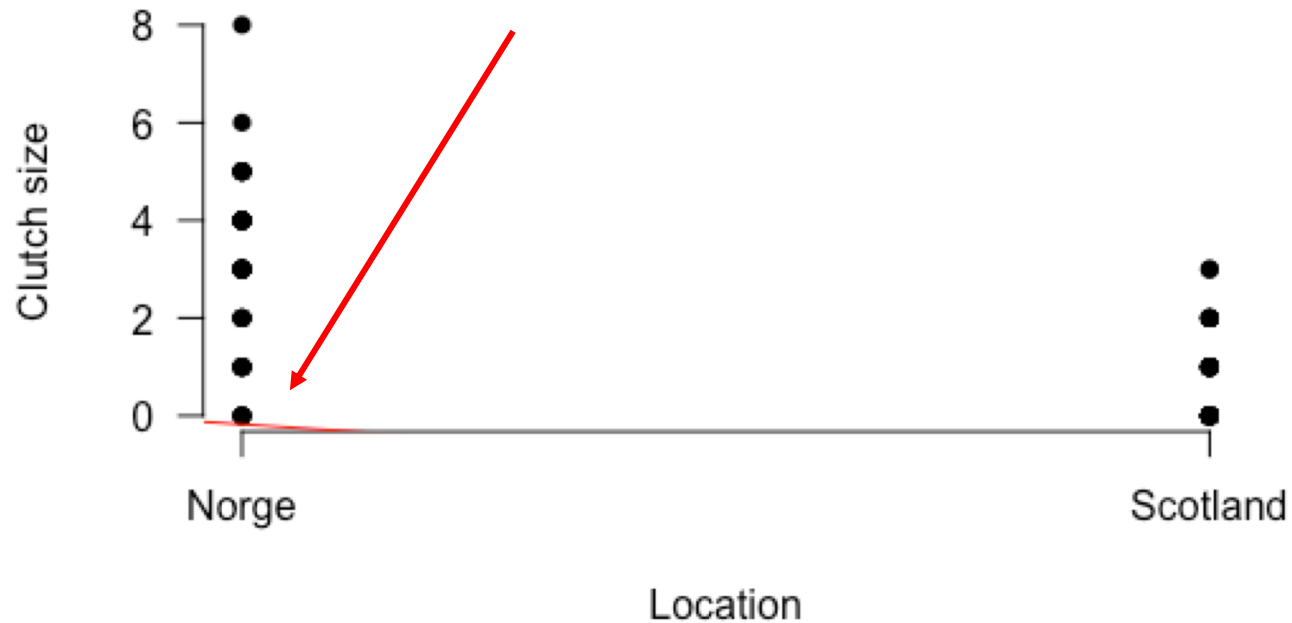
```
> confint(model1)
```

Waiting for profiling to be done... These are on log scale!

| | | |
|------------------|------------|------------|
| | 2.5 % | 97.5 % |
| (Intercept) | 0.9341921 | 1.2544806 |
| LocationScotland | -1.6270659 | -0.9408995 |

Exercise 5: ANSWER

```
> coef(model1)
(Intercept) LocationScotland
1.098612      -1.272966
```



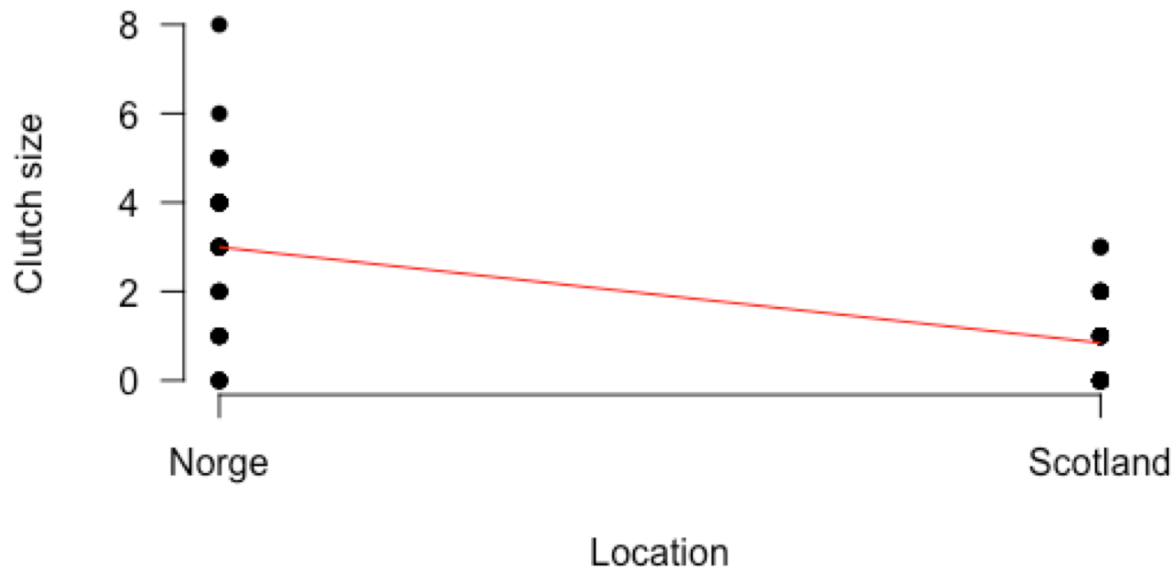
Exercise 5: ANSWER

```
> coef(model1)
      (Intercept) LocationScotland
           1.098612           -1.272966
```

Use `exp()` to take the inverse of the link function and get predictions on scale of Y

For $\beta_0 > 0$ need to take `exp()` of whole equation (predicting)

```
> lines(x=c(1,2), y=c(exp(1.098), exp(1.098-1.2729)), col=2)
```



Checking model fit with GLMs

Assumptions of a GLM

Assumptions of a GLM:

- Lack of outliers
- Correct distribution used
- Correct link function is used
- Correct variance function is used
- Dispersion parameter is constant
- Independence of y

Checking the model fit

For linear models we used:

Residuals vs fitted plots

Normal Q-Q plots

Cook's distance

These are easy to interpret – we know what we are looking for

This is not the case for GLMs – non-normal variance!

Checking the model fit

For linear models we used:

Residuals vs fitted plots – **equal variance and linearity**

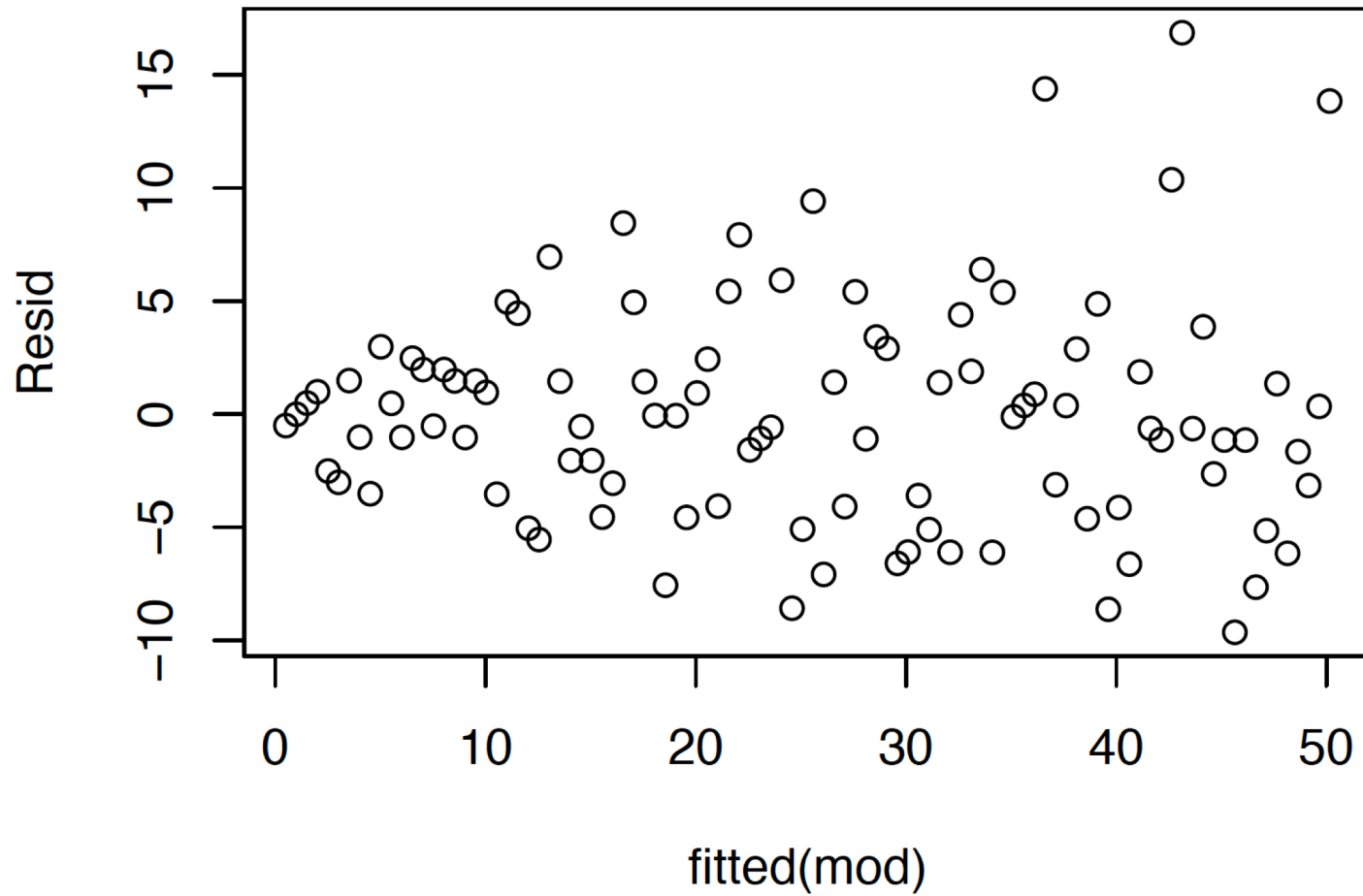
Normal Q-Q plots – **normality of residuals**

Cook's distance - **outliers**

These are easy to interpret – we know what we are looking for

This is not the case for GLMs – non-normal variance!

Checking the model fit



Checking the model fit

Need a way to handle non-constant variance

Want to produce plots that are roughly normal

Two ways: **Pearson** and **Deviance** residuals (neither is perfect)

Both scale residual by variance (in some way)

Pearson residuals: $(x - \mu_x)/\sigma_x$

Deviance residuals: $\text{sgn}(y_i - E(y_i)) \sqrt{D_i}$

$\text{sgn}(x) = 1$ when $x > 0$ and -1 when $x < 0$

Checking the model fit


Need a way to handle non-constant variance

Want to produce plots that are roughly normal

Two ways: **Pearson** and **Deviance** residuals (neither is perfect)

Both scale residual by variance (in some way)

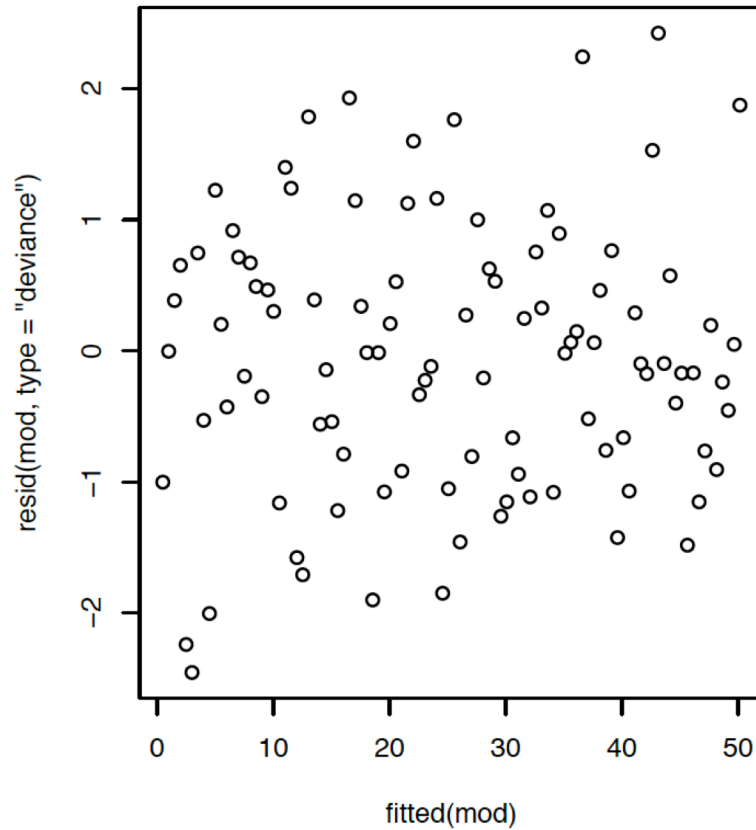
Pearson residuals: $(x - \mu_x)/\sigma_x$

Deviance residuals: $\text{sgn}(y_i - E(y_i)) \sqrt{D_i}$  Default for
glm

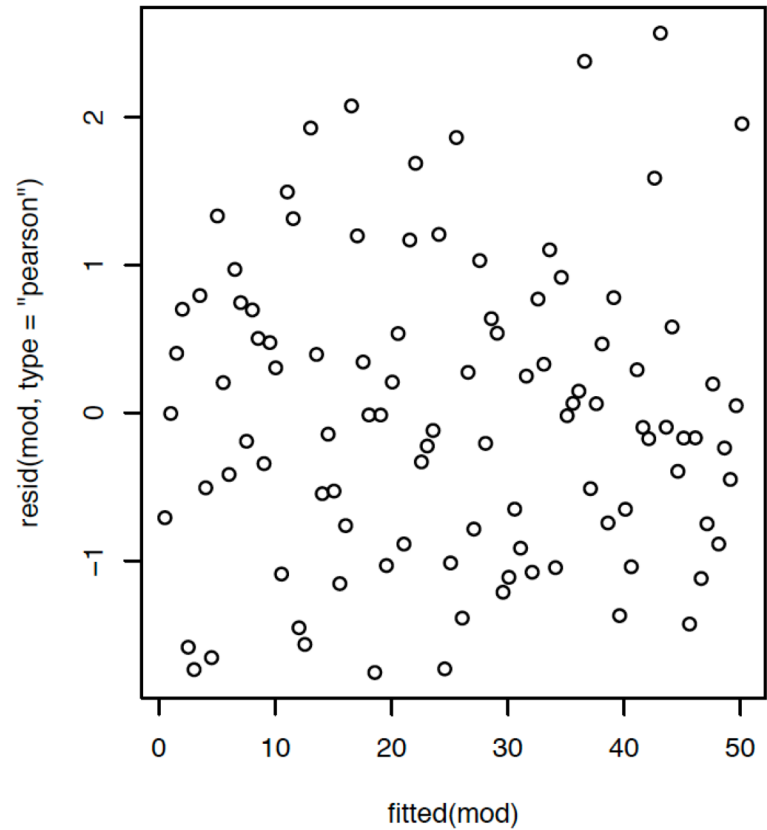
$\text{sgn}(x) = 1$ when $x > 0$ and -1 when $x < 0$

Checking the model fit

Deviance



Pearson



Checking the model fit - summary

These plots are still important (with tweaks):

Residuals vs fitted plots

Normal Q-Q plots

Cook's distance

Once we have scaled the residuals to account for non-equal variance, they should be approximately normal

Outliers still important

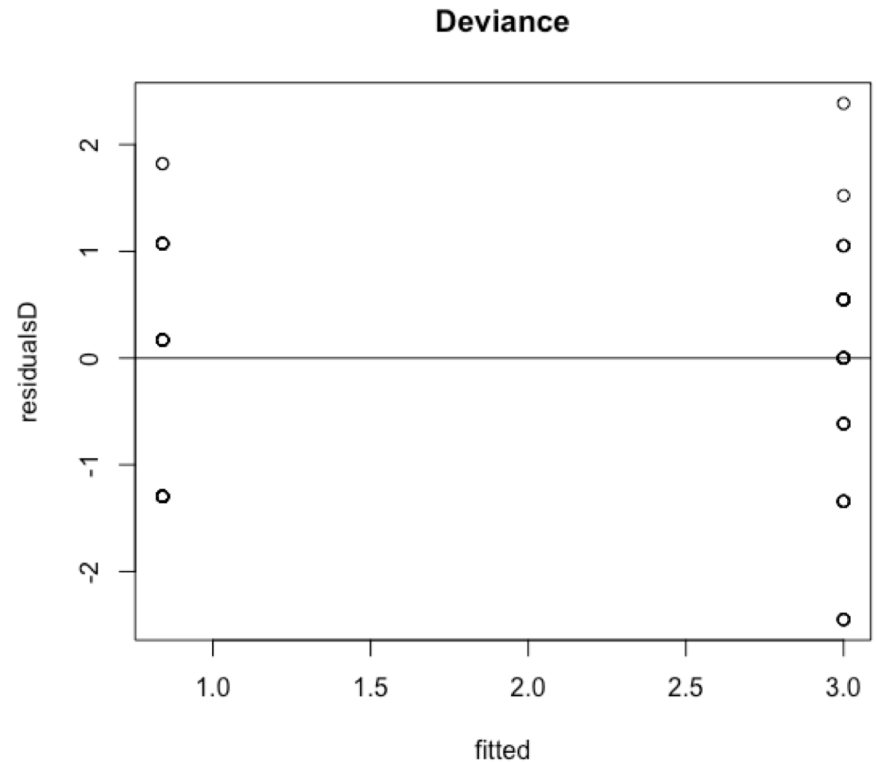
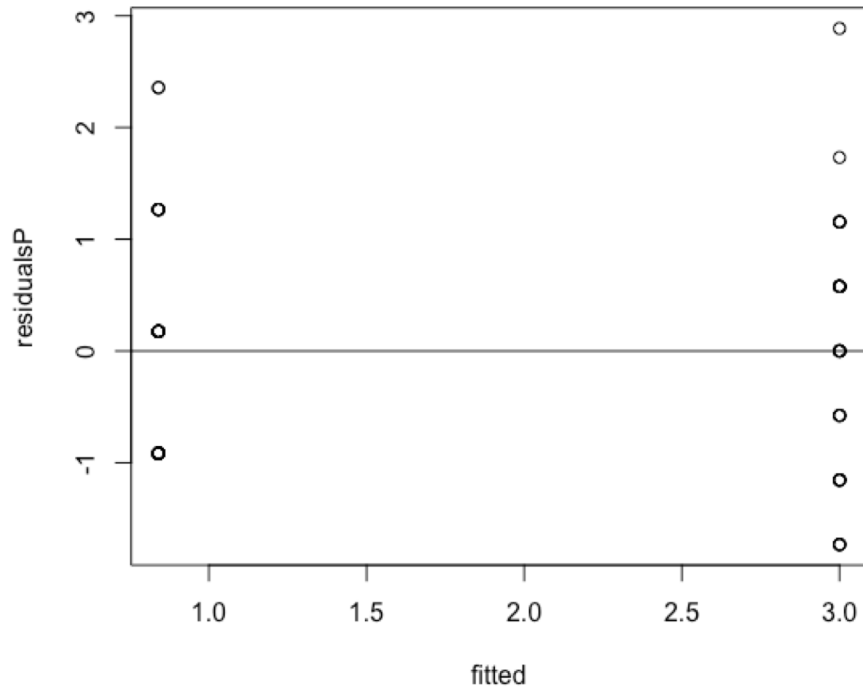
Plots still useful even if they look weird

Exercise 6: Check model fit

- Part E

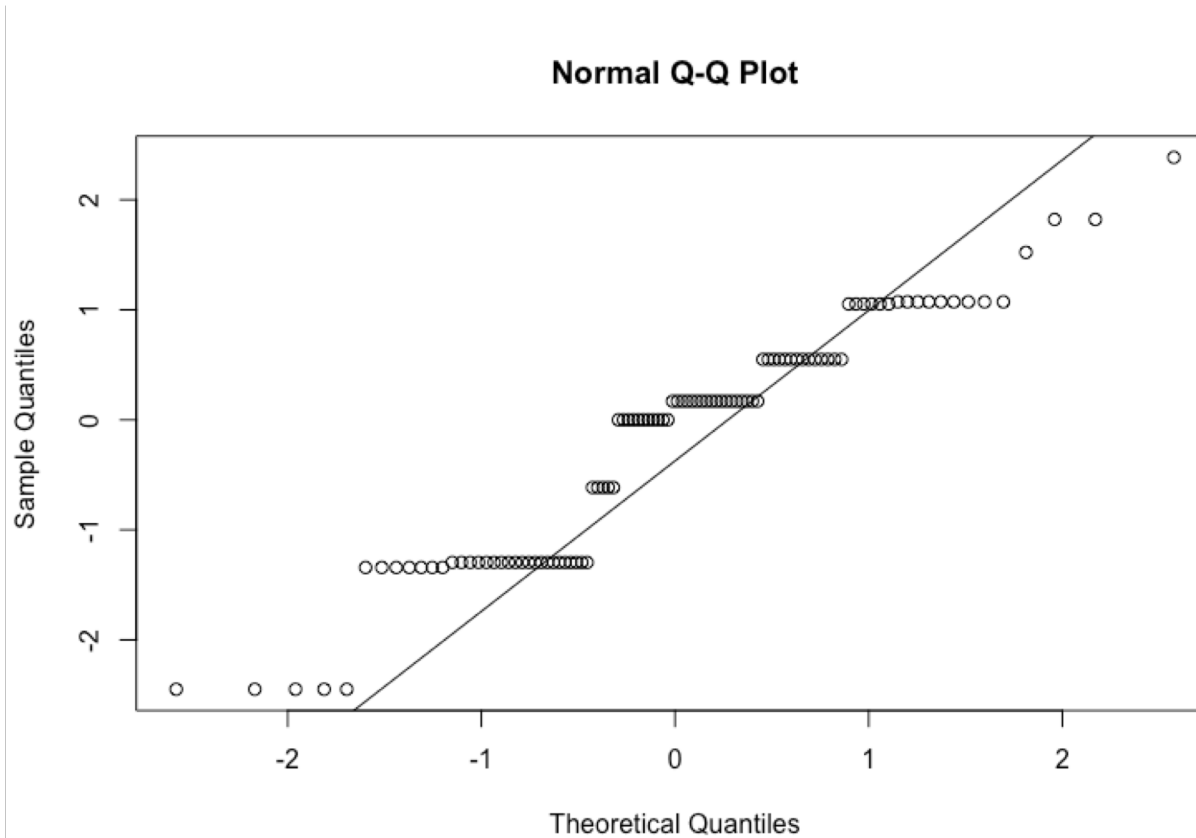
Exercise 6: ANSWER

```
> residualsP <- resid(model1, type="pearson")  
> residualsD <- resid(model1, type="deviance")  
>  
> fitted <- fitted(model1)  extract fitted values  
>  
> par(mfrow=c(1,2))  makes two plots next to each other  
>  
> plot(fitted, residualsP, main="Pearson")  
> abline(h=0)          plot  
> plot(fitted, residualsD, main="Deviance")  
> abline(h=0)
```



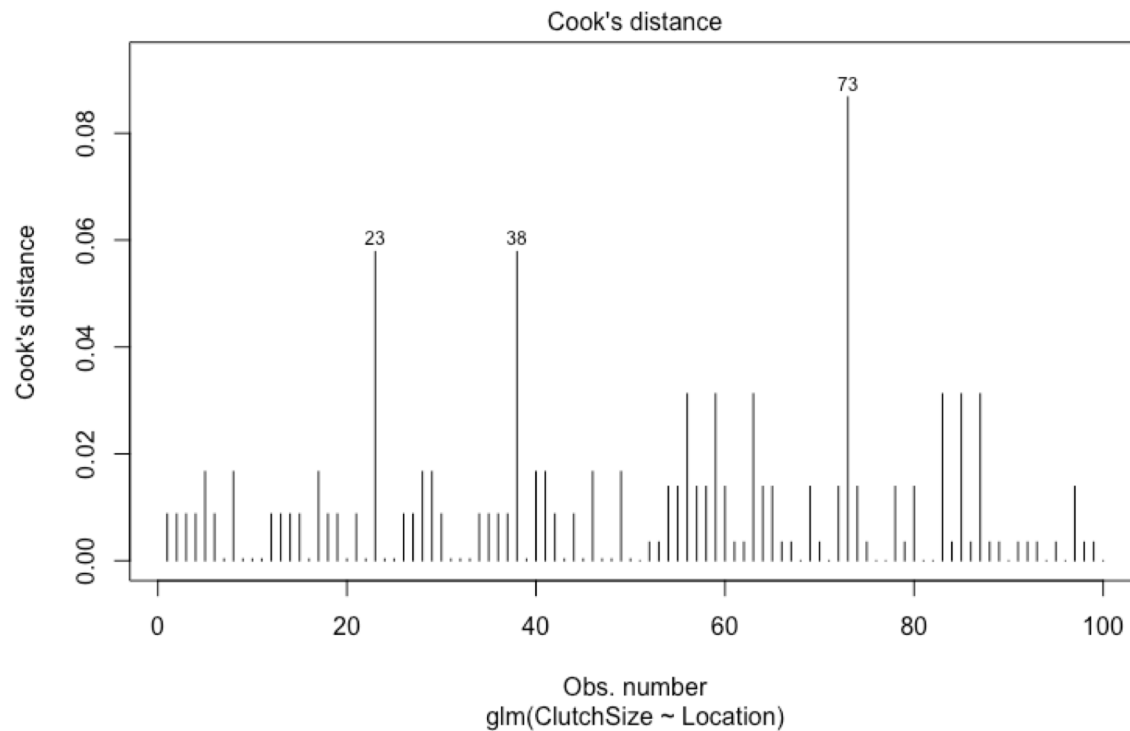
Exercise 6: ANSWER

```
> qqnorm(residualsD)  
> qqline(residualsD)
```



Exercise 6: ANSWER

```
> plot(model1, which=4)
```



Lecture Outline – Part 2

More on the Random part

Choose a distribution based on your data

Basics of the Poisson GLM

Uses log link as default and used for count data

Checking model fit with GLMs

Bit more difficult for GLMs but can still use similar tools