Generalised Linear Models (GLM): Part 2

Lecture Outline

Recap of yesterday

More on the Random part

Basics of the Poisson GLM

Model selection with GLMs

Checking model fit with GLMs

Lecture Outline

Recap of yesterday

More on the Random part

- EX1: Choose a distribution

Basics of the Poisson GLM

- EX2: Fit a Poisson GLM in R

Model selection with GLMs

- EX3: Conduct model selection

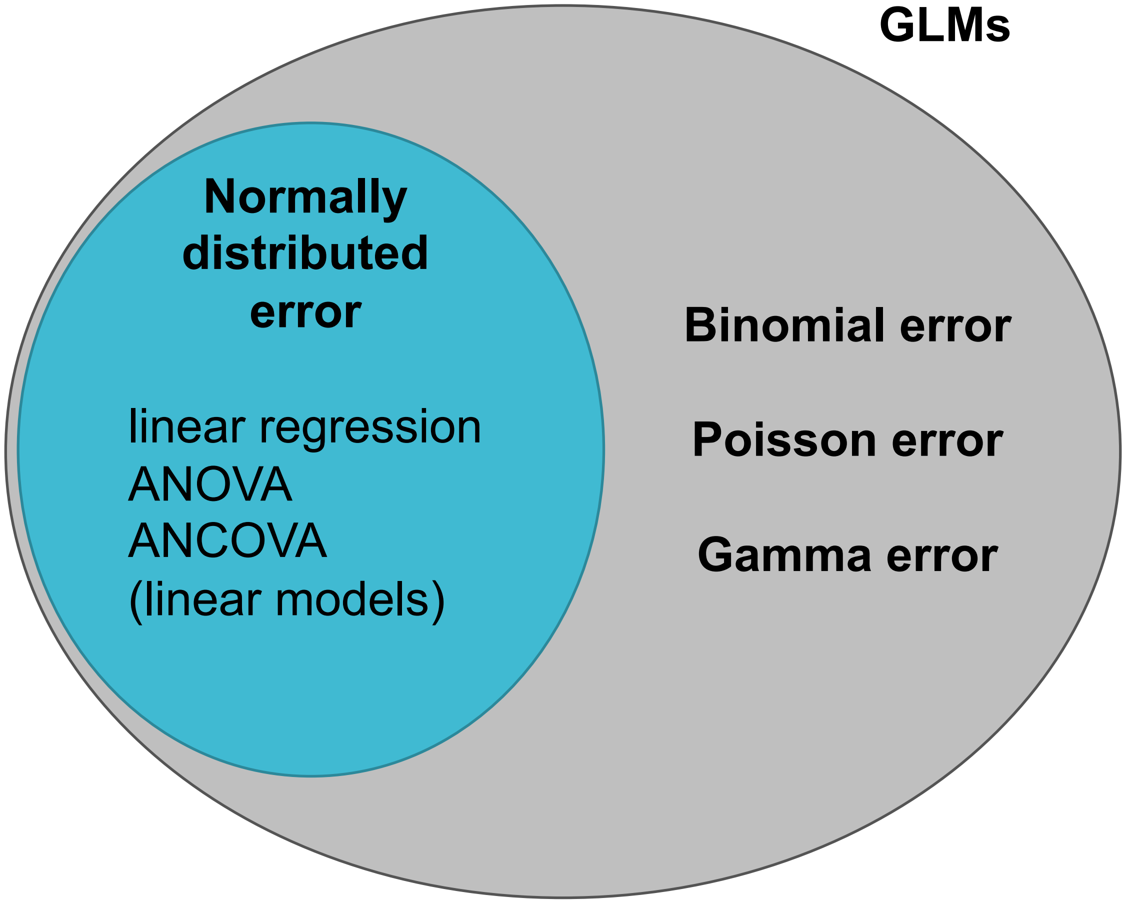
Checking model fit with GLMs

- EX4: Check your model fit/ interpret

Recap of yesterday

Generalised linear models

Useful for non-normal and non-linear data



GLMs

Normally distributed error

linear regression
ANOVA
ANCOVA
(linear models)

Binomial error

Poisson error

Gamma error

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?

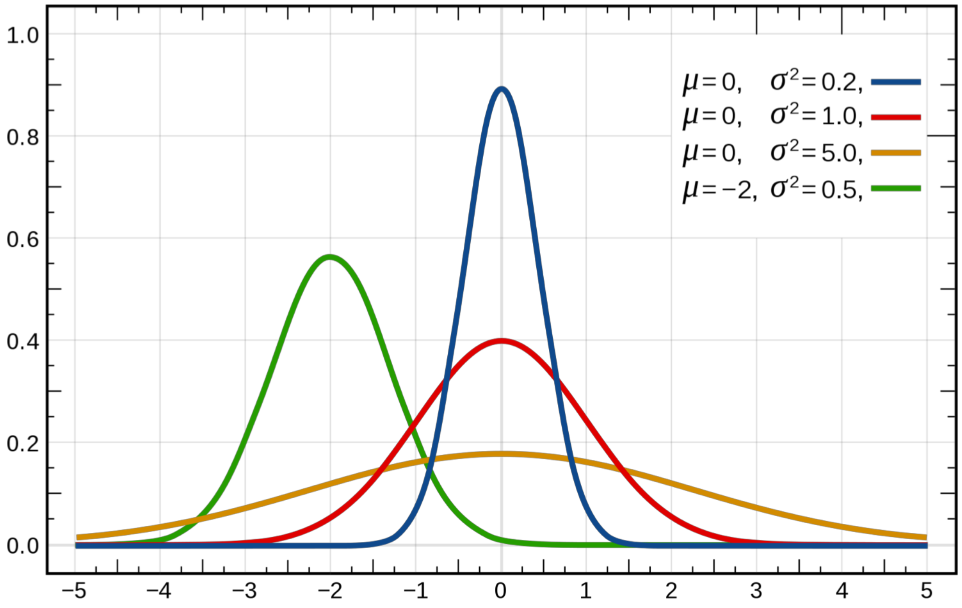
GLM can use **Normal, Binomial, Poisson**, Gamma, and some quasi- distributions

The Normal Distribution

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

Properties: Continuous, symmetrical around mean, single mode

Examples: height, biomass, running times



* Picture from Wikipedia

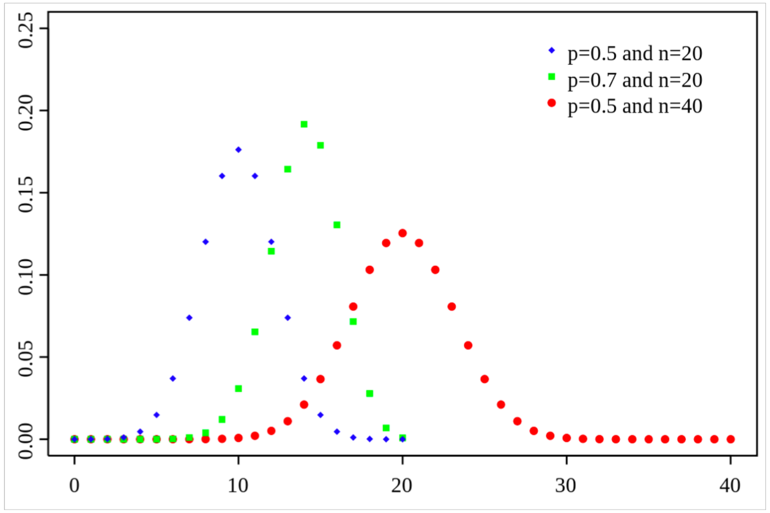
The Binomial Distribution

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



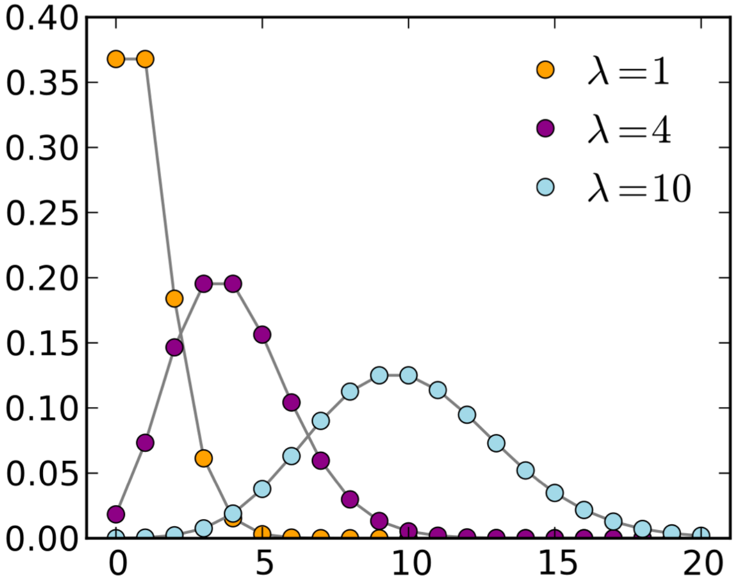
* Picture from Wikipedia

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

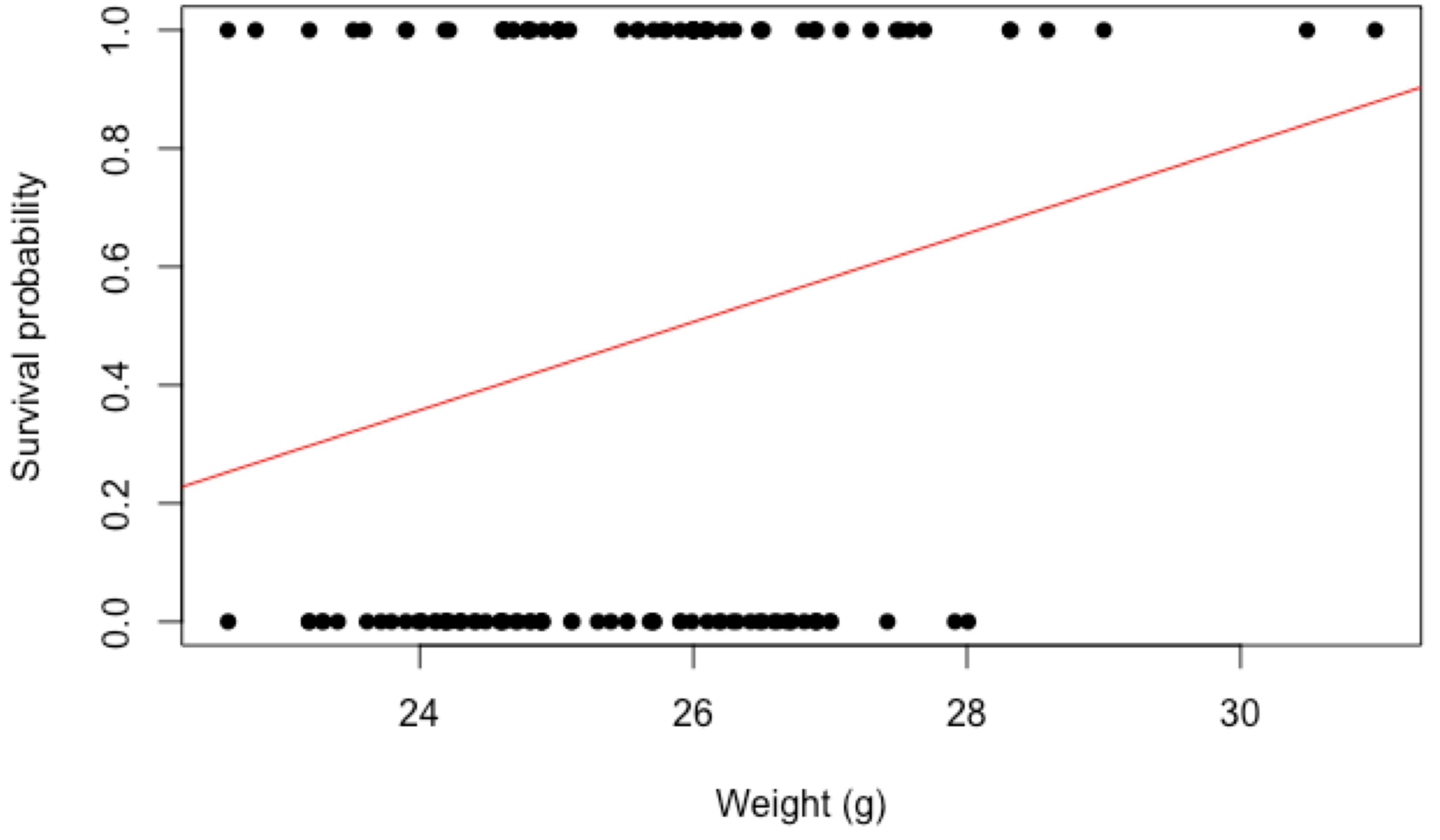
Exercise 1: Which distribution?

- Look at the examples of data on the next slides (the same as yesterday)
- This time you need to decide which distribution would be most appropriate to model these data

Example 1: Survival of sparrows

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

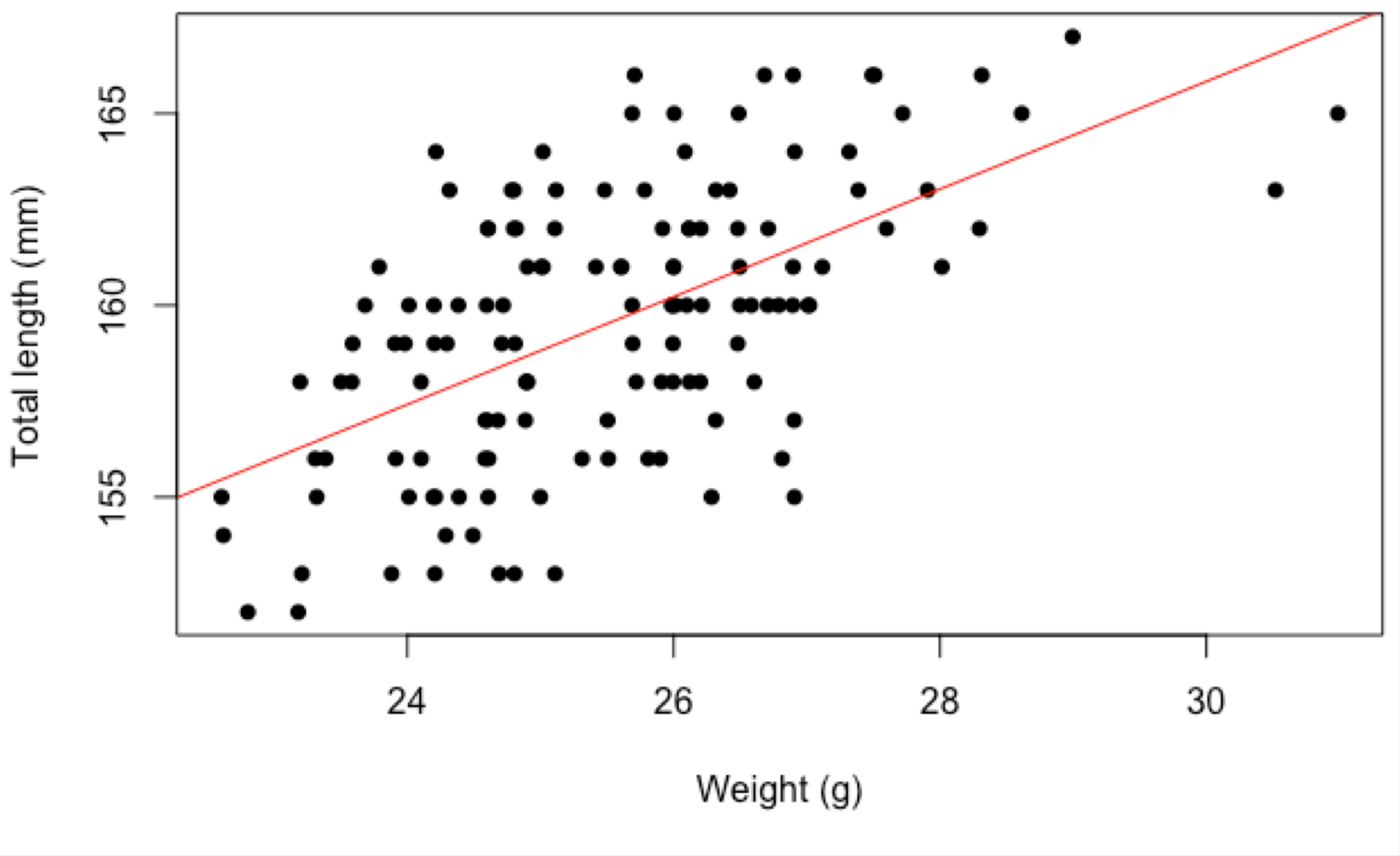
Data: Response = whether the bird survived (1), or not (0). Explanatory = body weight in grams



Example 2: Length and weight in sparrows

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

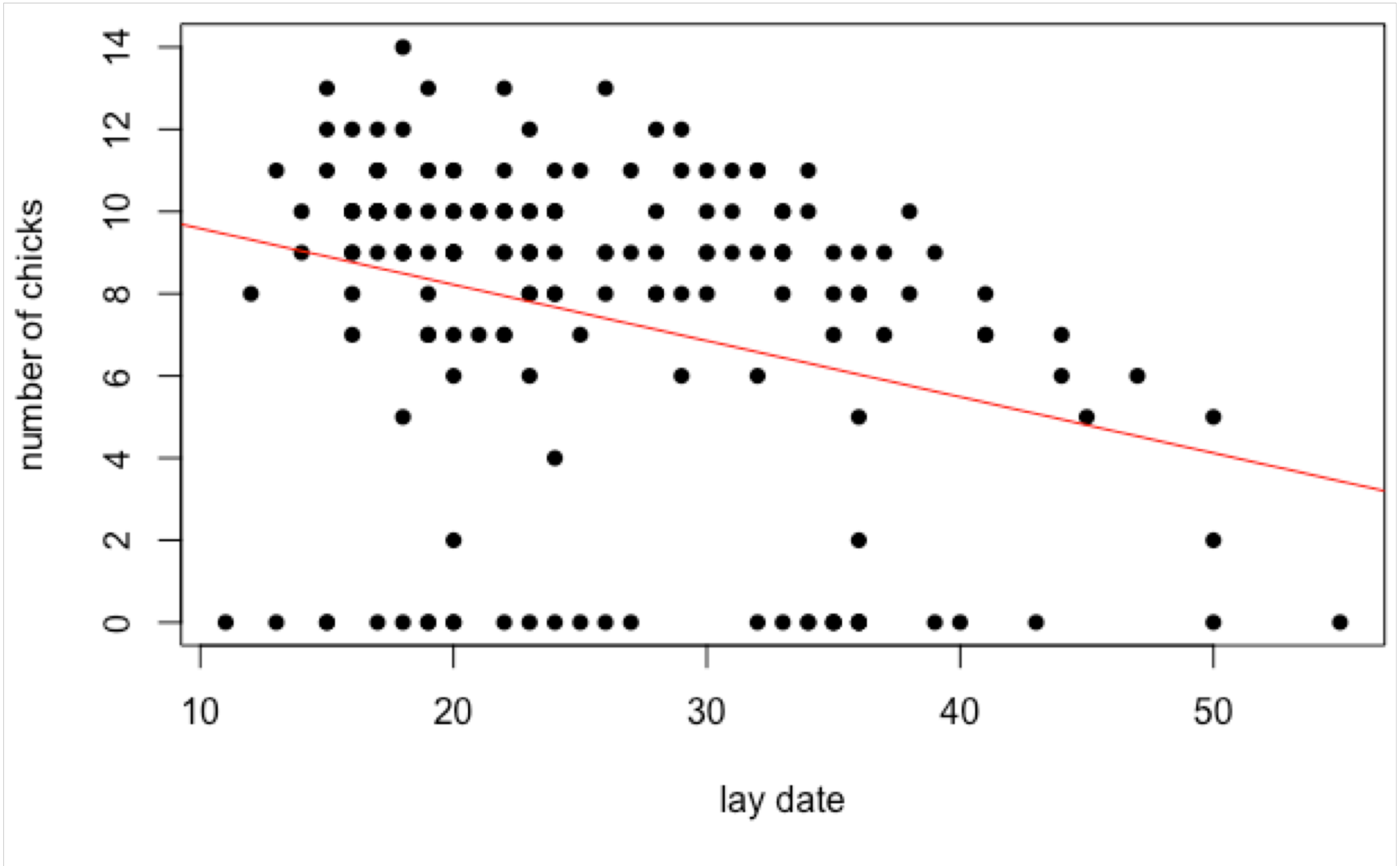
Data: Response = total length in mm. Explanatory = body weight in grams



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)



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:

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

Poisson likelihood:

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

$a(\phi) = 1$
$\theta = \log(\mu)$

The likelihood

General likelihood for GLM: $l(\theta|y) = \frac{y\theta - \boxed{b(\theta)}}{a(\phi)} + c(y, \phi)$

Poisson likelihood: $l(\mu|r) = \boxed{-\mu} + r\log(\mu) - \log(r!)$

$a(\phi) = 1$
$\theta = \log(\mu)$
$b(\theta) = -e^{\theta} = -e^{\log(\mu)} = -\mu$

The likelihood

General likelihood for GLM: $$l(\theta|y) = \frac{y\theta - b(\theta)}{a(\phi)} + \boxed{c(y,\phi)}$$

Poisson likelihood: $$l(\mu|r) = -\mu + r\log(\mu) - \boxed{\log(r!)}$$

$a(\phi) = 1$

$\theta = \log(\mu)$

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

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

The likelihood

General likelihood for GLM: $l(\theta|y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$

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

$a(\phi) = 1$
$\theta = \log(\mu)$
$b(\theta) = -e^{\theta} = -e^{\log(\mu)} = -\mu$
$c(y,\phi) = -\log(r!)$

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 2: Fit the Poisson GLM in R

- Take the data for the phoenix clutch size
- https://www.math.ntnu.no/emner/ST2304/2019v/Week11/Phoenix.csv
- Fit a GLM with a Poisson family and log link to look at whether location of nest influences number of eggs
- Basic formula is below, you will need to edit
- Look at results using coef() THEN STOP

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

Model selection with GLMs

Model selection

A bit different for GLMs

Terminology changes

Model selection

A bit different for GLMs

Terminology changes

Exploratory model selection with AIC/BIC:

AIC/BIC = Deviance + penalty

Deviance = $-2 * l(\theta|y)$

Model selection

A bit different for GLMs

Terminology changes

Exploratory model selection with AIC/BIC:

AIC/BIC = Deviance + penalty

SAME

Deviance = $-2 * l(\theta | y)$

Model selection

A bit different for GLMs

Terminology changes

Exploratory model selection with AIC/BIC:

AIC/BIC = Deviance + penalty

SAME

Deviance = $-2 * l(\theta|y)$

Confirmatory model selection using the anova() function:

Becomes analysis of deviance

Model selection

A bit different for GLMs

Terminology changes

Exploratory model selection with AIC/BIC:

AIC/BIC = Deviance + penalty

SAME

Deviance = $-2 * l(\theta|y)$

Confirmatory model selection using the anova() function:

Becomes analysis of deviance

Not quite the same

Analysis of deviance

Compares deviance instead of sum of squares

Residual deviance = twice the difference in loglikelihood of **saturated model** (parameter for each data point) and the **proposed model**

Deviance = difference in residual deviances

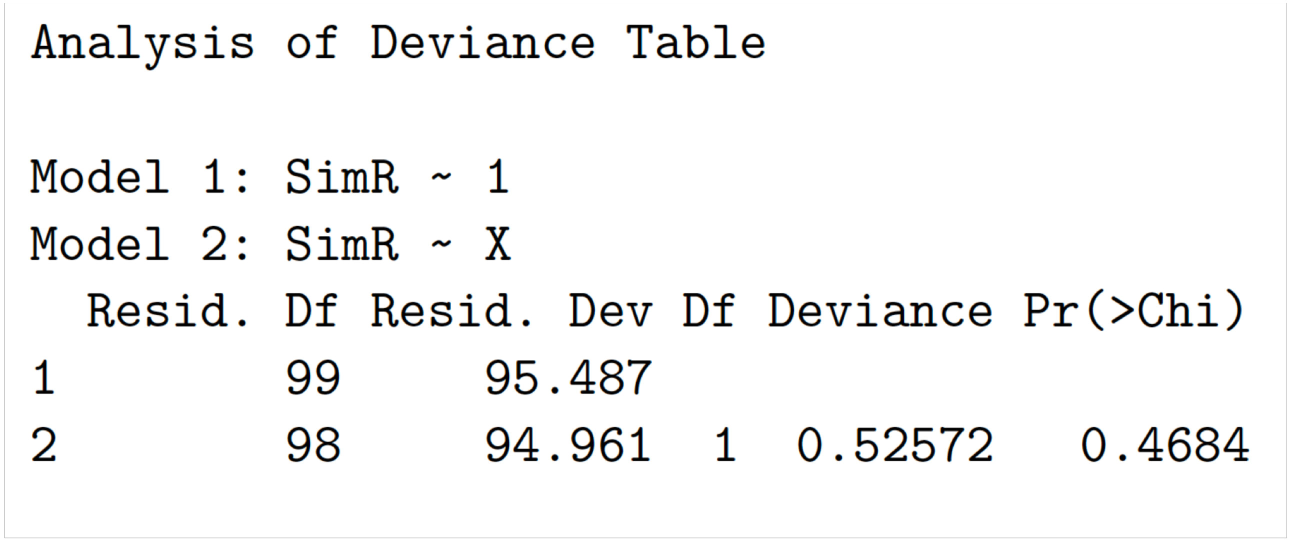
Analysis of deviance

Compares deviance instead of sum of squares

Residual deviance = twice the difference in loglikelihood of **saturated model** (parameter for each data point) and the **proposed model**

Deviance = difference in residual deviances

```
anova(mod, mod1, test="LRT")
```



```
Analysis of Deviance Table

Model 1: SimR ~ 1
Model 2: SimR ~ X
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         99     95.487
2         98     94.961  1  0.52572   0.4684
```

Analysis of deviance

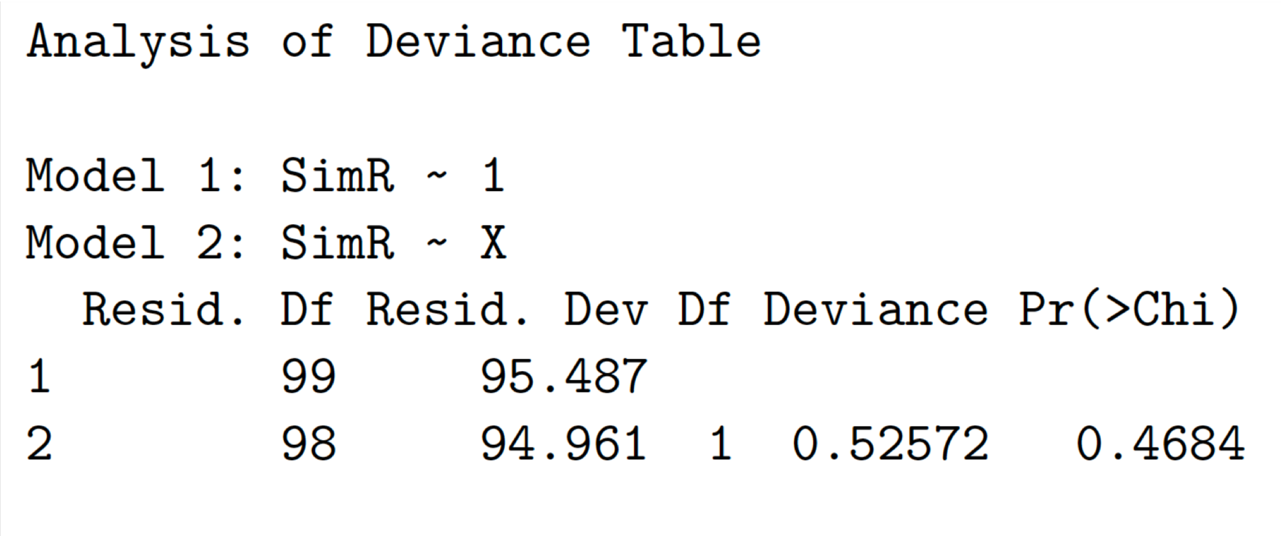
Compares deviance instead of sum of squares

Residual deviance = twice the difference in loglikelihood of **saturated model** (parameter for each data point) and the **proposed model**

Deviance = difference in residual deviances

`anova(mod, mod1, test="LRT")`

LRT = likelihood ratio test



```
Analysis of Deviance Table

Model 1: SimR ~ 1
Model 2: SimR ~ X
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        99     95.487
2        98     94.961  1  0.52572   0.4684
```

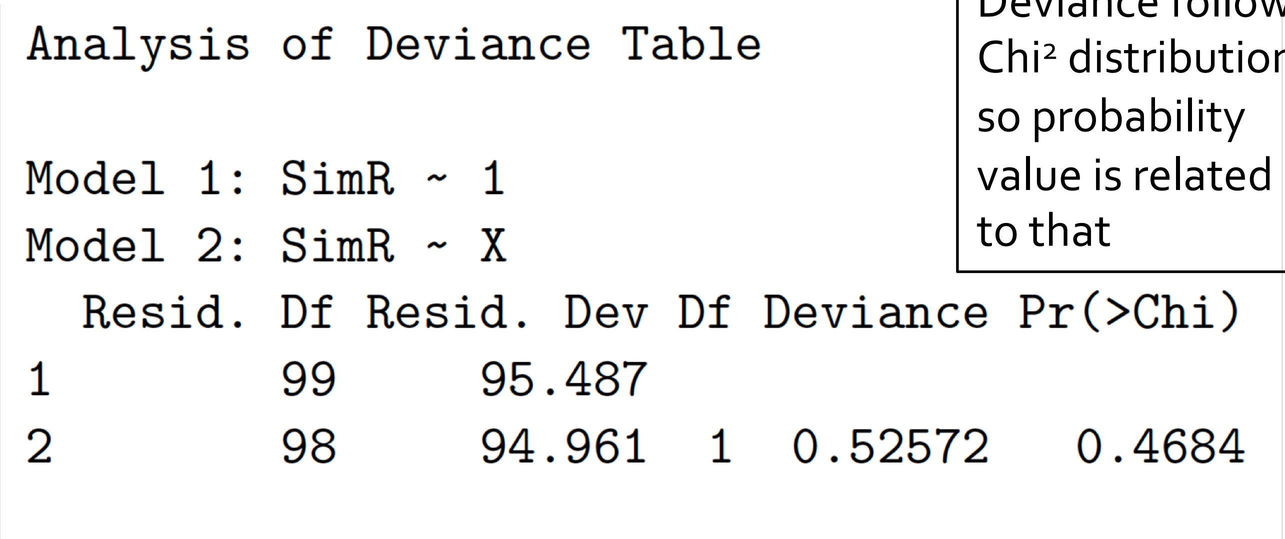
Analysis of deviance

Compares deviance instead of sum of squares

Residual deviance = twice the difference in loglikelihood of **saturated model** (parameter for each data point) and the **proposed model**

Deviance = difference in residual deviances

```
anova(mod, mod1, test="LRT")
```



```
Analysis of Deviance Table

Model 1: SimR ~ 1
Model 2: SimR ~ X
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        99     95.487
2        98     94.961  1  0.52572   0.4684
```

Deviance follows Chi^2 distribution so probability value is related to that

Exercise 3: Model selection

- Look back at the previous slides and the data
- What is our question here?
- Is this a confirmatory or exploratory question?
- Conduct model selection for this question using code given on previous slide
- What do you conclude about the question?

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

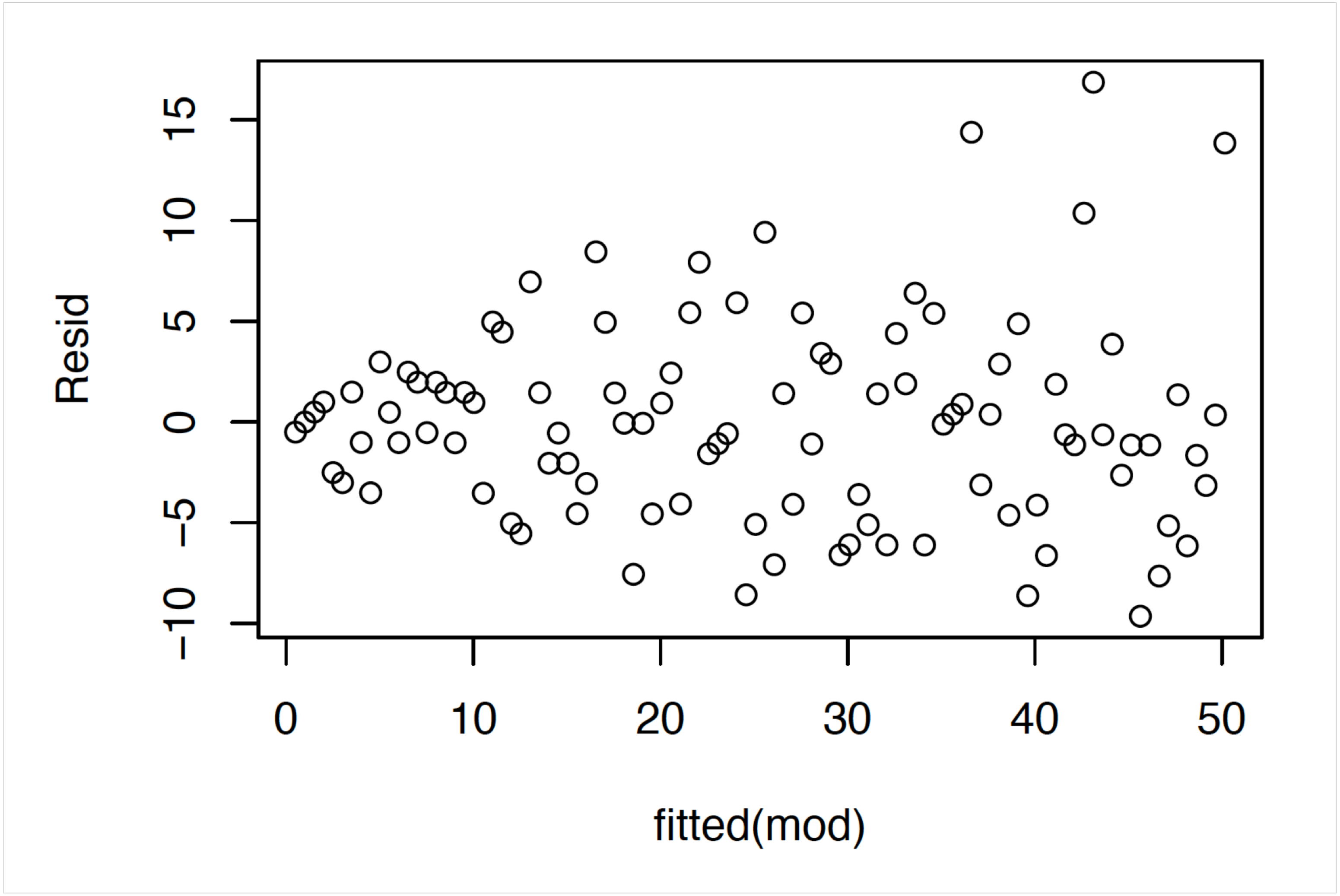
Normal Q-Q plots – normality of residuals

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: $\mathrm{sgn}\left(y_i - E(y_i)\right)\sqrt{D_i}$

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)

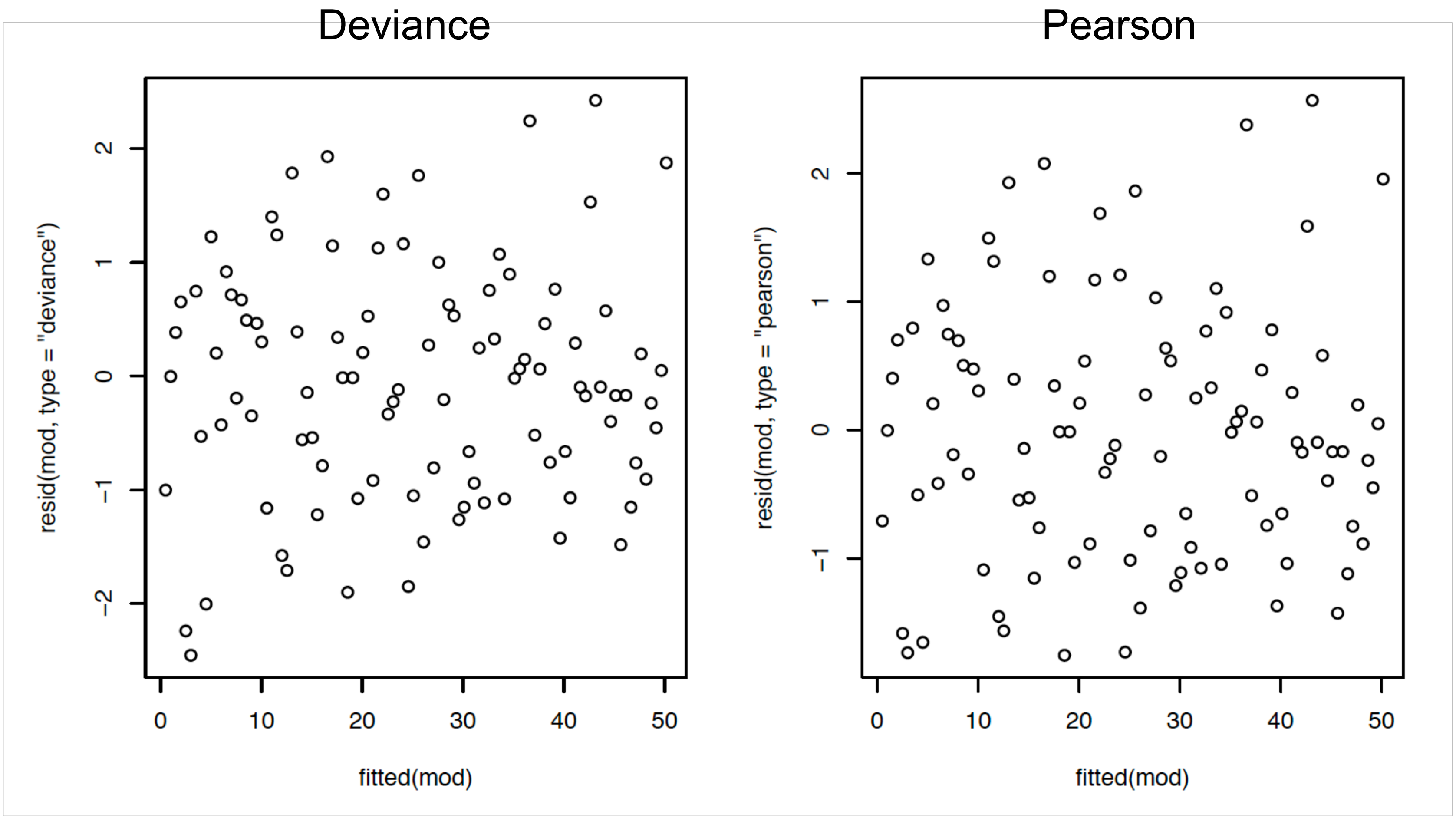
Pearson residuals: $(x - \mu_x)/\sigma_x$

Deviance residuals: $\mathrm{sgn}\left(y_i - E(y_i)\right)\sqrt{D_i}$

Default for glm

sgn(x) = 1 when x > 0 and -1 when x < 0

Checking the model fit



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 4: Check model fit

- After exercise 3 you should have a final model
- Check the fit of the model using Pearson and Deviance residuals. Check linearity, normality, and outliers
- What do you think?

code:

```
resid(model, type=”pearson”)
resid(model, type=”deviance”)
fitted(model)

plot(fitted, residuals)
qqnorm(residuals)
qqline(residuals)
plot(model, which=4) # cook’s distance
```

Exercise 4: Interpret

- Now you have checked your model fit interpret the output

- Remember the link function! The parameters (coefficients) are for the linear predictor, which sits inside the link function

- Our link here was log(), the inverse is exp()

Lecture Summary

Recap of yesterday

More on the Random part

Basics of the Poisson GLM

Model selection with GLMs

Checking model fit with GLMs

Lecture Summary

Recap of yesterday

More on the Random part
Choose a distribution based on your data

Basics of the Poisson GLM

Model selection with GLMs

Checking model fit with GLMs

Lecture Summary

Recap of yesterday

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

Model selection with GLMs

Checking model fit with GLMs

Lecture Summary

Recap of yesterday

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

Model selection with GLMs
Bit different for confirmatory selection – uses analysis of deviance

Checking model fit with GLMs

Lecture Summary

Recap of yesterday

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

Model selection with GLMs
Bit different for confirmatory selection – uses analysis of deviance

Checking model fit with GLMs
Bit more difficult for GLMs but can still use similar tools