Binomial/logistic GLM: Part 1

Lecture Outline

Recap of last week

What is the binomial GLM (logistic regression)?

When and why to use a binomial GLM?

More on the logit link function

Interpreting (and plotting) results from a binomial GLM

Lecture Outline

Recap of last week

What is the binomial GLM (logistic regression)?

When and why to use a binomial GLM?

- EX1: Why is a straight line bad?
- EX2: Running Binomial GLM

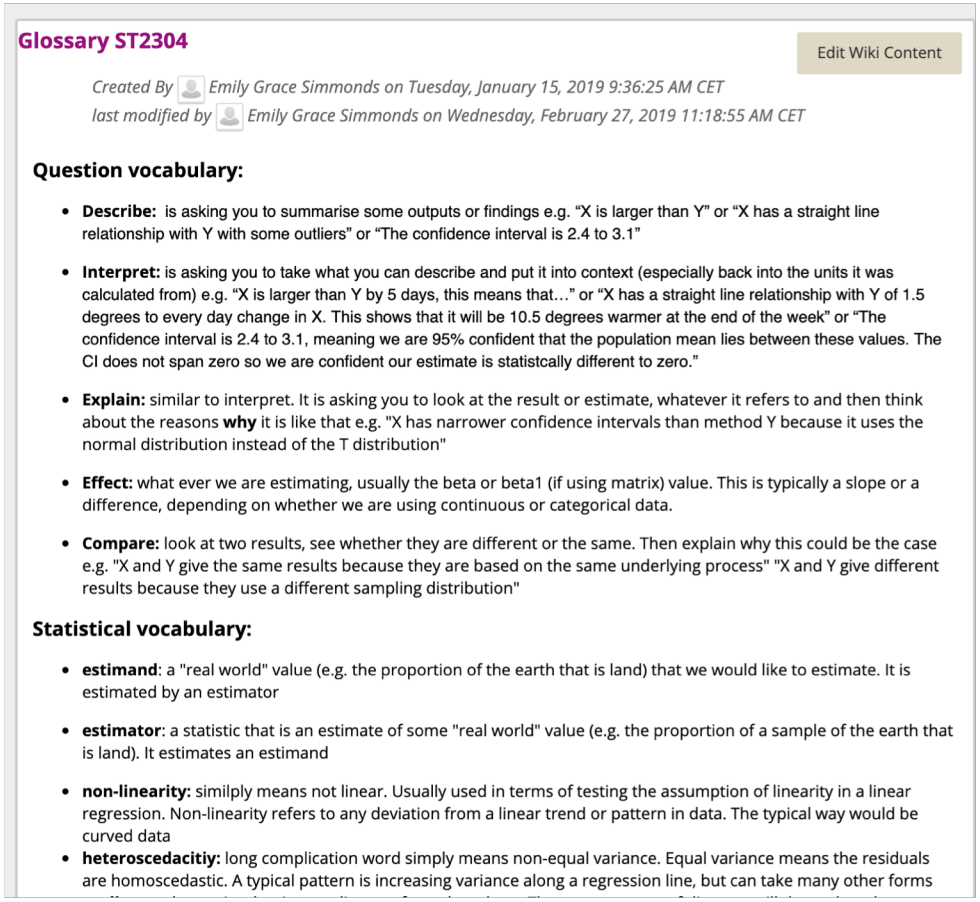
More on the logit link function

Interpreting (and plotting) results from a binomial GLM

- EX3: Interpretation
- EX4: Plotting

General

Glossary on Blackboard



Glossary ST2304

Edit Wiki Content

Created By Emily Grace Simmonds *on Tuesday, January 15, 2019 9:36:25 AM CET*
last modified by Emily Grace Simmonds *on Wednesday, February 27, 2019 11:18:55 AM CET*

Question vocabulary:

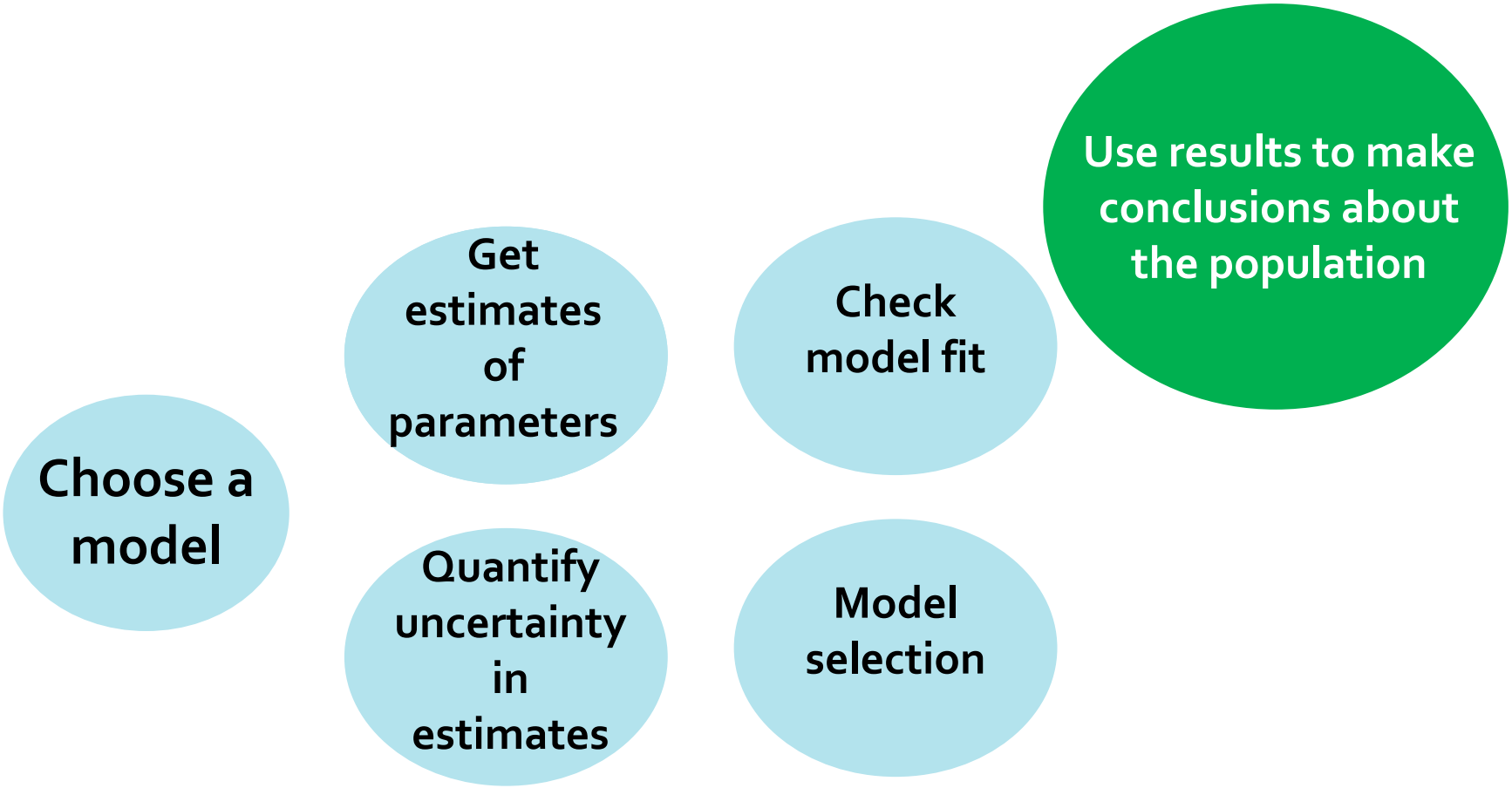
- **Describe:** is asking you to summarise some outputs or findings e.g. "X is larger than Y" or "X has a straight line relationship with Y with some outliers" or "The confidence interval is 2.4 to 3.1"
- **Interpret:** is asking you to take what you can describe and put it into context (especially back into the units it was calculated from) e.g. "X is larger than Y by 5 days, this means that…" or "X has a straight line relationship with Y of 1.5 degrees to every day change in X. This shows that it will be 10.5 degrees warmer at the end of the week" or "The confidence interval is 2.4 to 3.1, meaning we are 95% confident that the population mean lies between these values. The CI does not span zero so we are confident our estimate is statistcally different to zero."
- **Explain:** similar to interpret. It is asking you to look at the result or estimate, whatever it refers to and then think about the reasons **why** it is like that e.g. "X has narrower confidence intervals than method Y because it uses the normal distribution instead of the T distribution"
- **Effect:** what ever we are estimating, usually the beta or beta1 (if using matrix) value. This is typically a slope or a difference, depending on whether we are using continuous or categorical data.
- **Compare:** look at two results, see whether they are different or the same. Then explain why this could be the case e.g. "X and Y give the same results because they are based on the same underlying process" "X and Y give different results because they use a different sampling distribution"

Statistical vocabulary:

- **estimand**: a "real world" value (e.g. the proportion of the earth that is land) that we would like to estimate. It is estimated by an estimator
- **estimator**: a statistic that is an estimate of some "real world" value (e.g. the proportion of a sample of the earth that is land). It estimates an estimand
- **non-linearity:** similply means not linear. Usually used in terms of testing the assumption of linearity in a linear regression. Non-linearity refers to any deviation from a linear trend or pattern in data. The typical way would be curved data
- **heteroscedacitiy:** long complication word simply means non-equal variance. Equal variance means the residuals are homoscedastic. A typical pattern is increasing variance along a regression line, but can take many other forms

Recap of last week

The modelling process



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

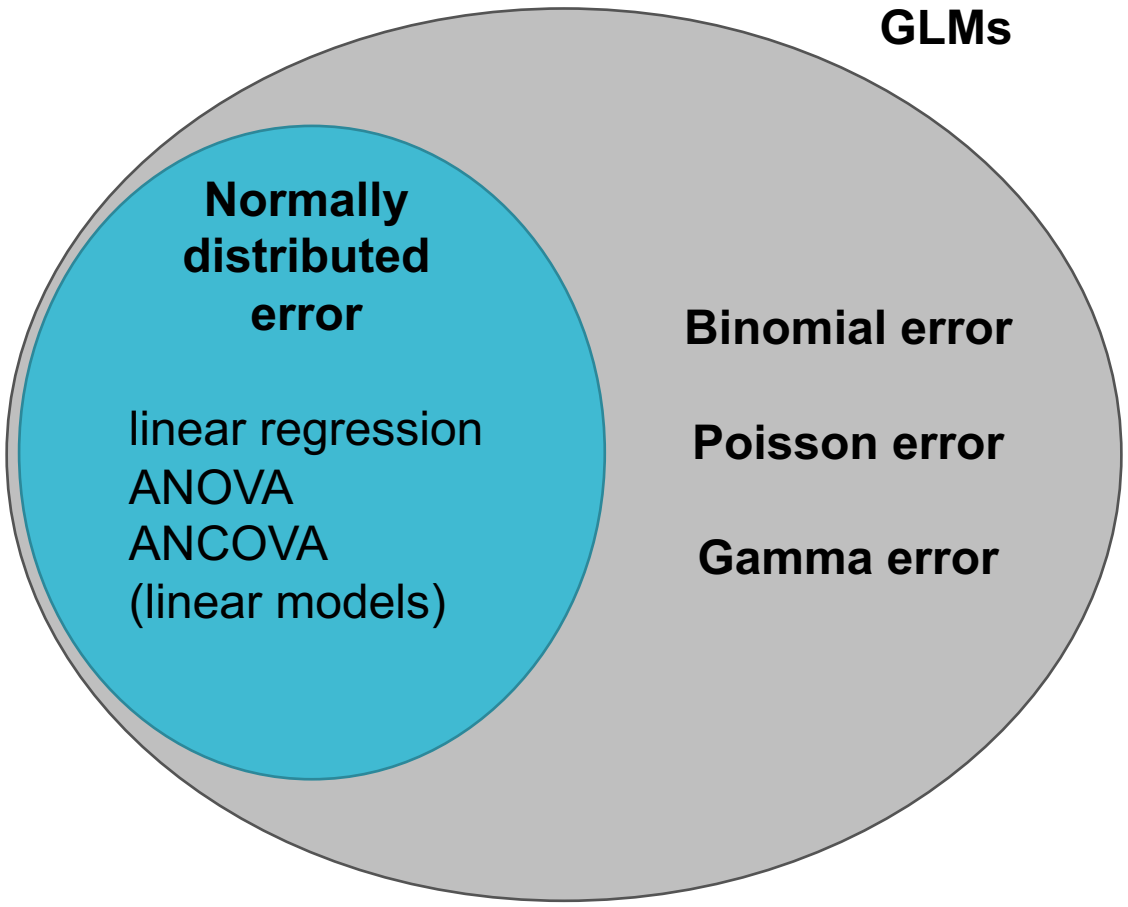
$l(\theta|y)$ is likelihood of expected value given the data

ϕ is the variance (dispersion)

$a, \mathrm{b}, \mathrm{and}\ \mathrm{c}$ are functions – will depend on the distribution used

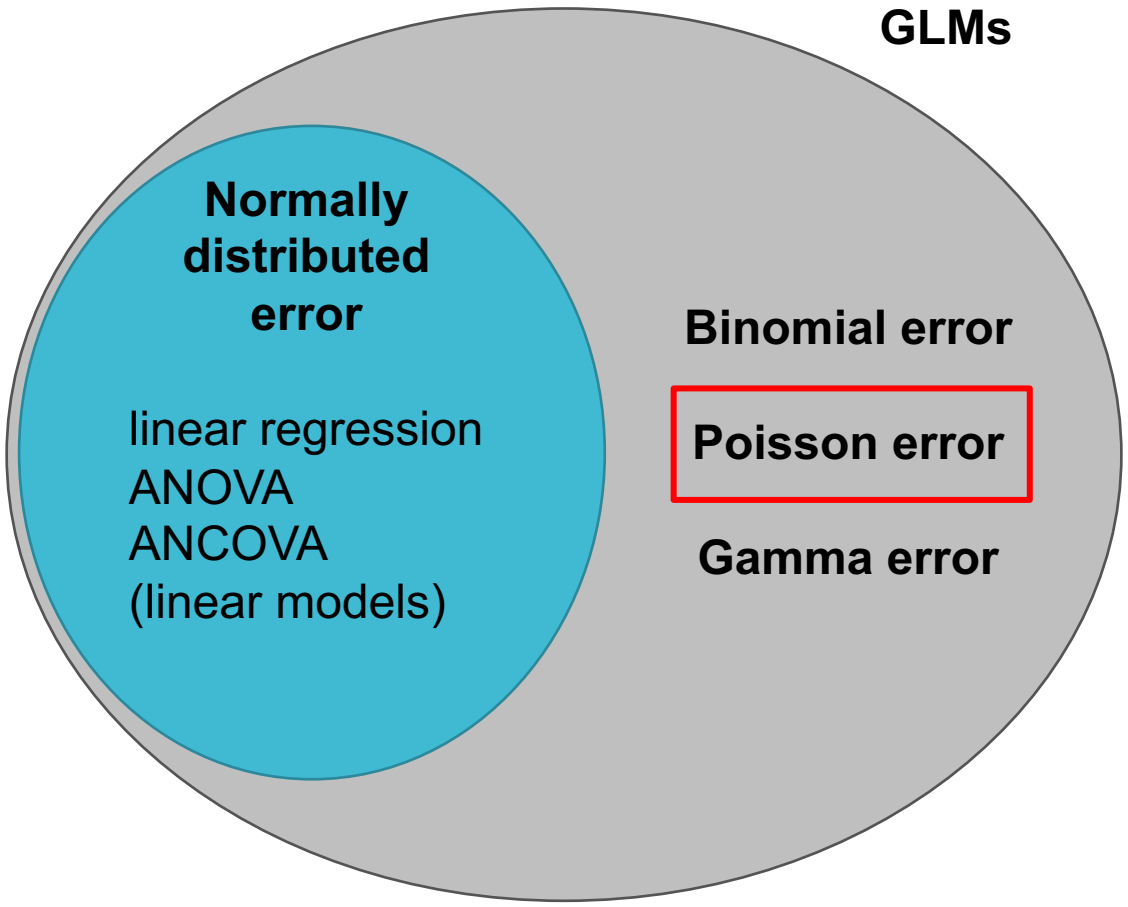
Generalised linear models

Similar to linear models but much more flexible



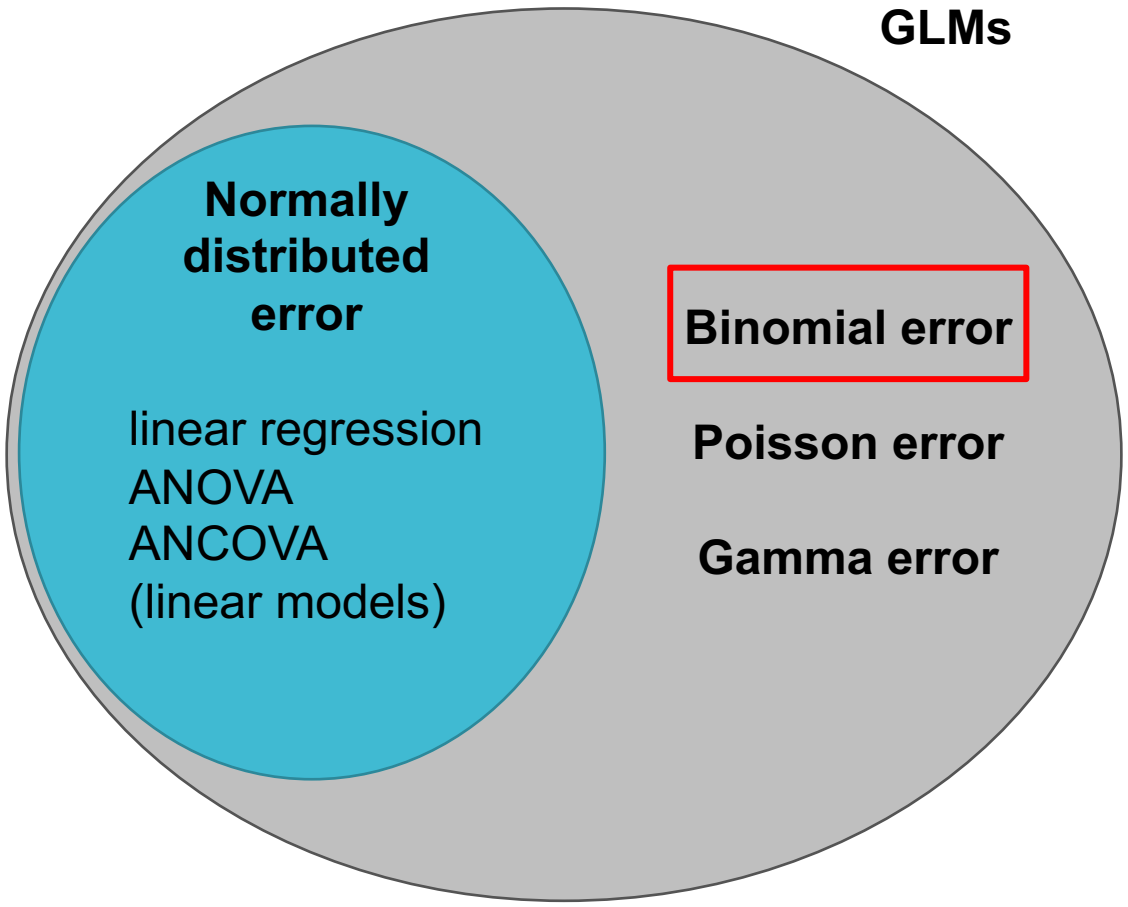
Generalised linear models

Similar to linear models but much more flexible



Generalised linear models

Similar to linear models but much more flexible

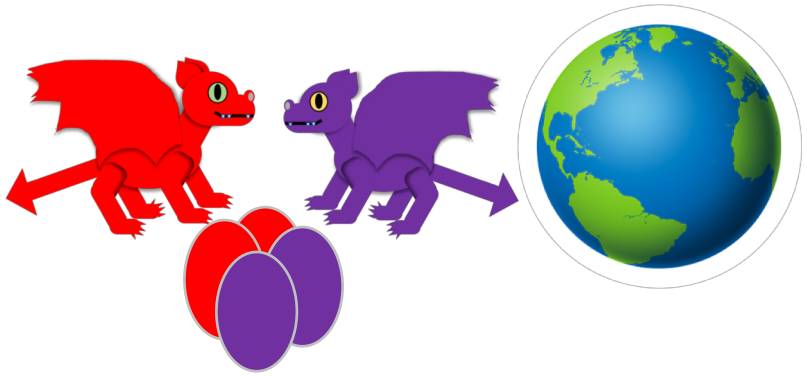


What is the Binomial GLM (logistic regression)?

The binomial distribution

Data: r, number of successes in N trials

Parameters: probability (p)

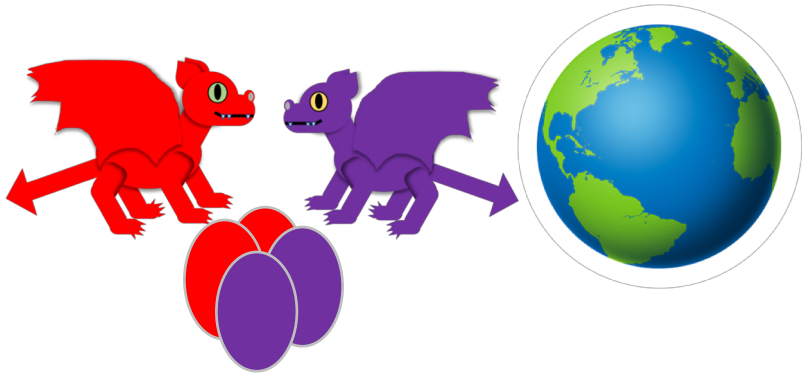


The binomial distribution

Data: r, number of successes in N trials

Parameters: probability (p)

Now using it in a GLM – **called logistic regression**



Binomial (log) likelihood

$$\log(\Pr(n = r|N,p)) = \log\left(\frac{N!}{r!\,(N-r)!}\right) + r\log p + (N-r)\log(1-p)$$

Binomial (log) likelihood

$$\log(\Pr(n = r|N,p)) = \log\left(\frac{N!}{r!\,(N-r)!}\right) + r\log p + (N-r)\log(1-p)$$

Same as in week 2!

Binomial (log) likelihood

$$\log(\Pr(n = r|N,p)) = \log\left(\frac{N!}{r!\,(N-r)!}\right) + r\log p + (N-r)\log(1-p)$$

Same as in week 2! But does it work as a GLM?

Binomial (log) likelihood

$$\log(\Pr(n = r|N,p)) = \log\left(\frac{N!}{r!\,(N-r)!}\right) + r\log p + (N-r)\log(1-p)$$

Same as in week 2! But does it work as a GLM?

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

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)

$a, \mathrm{b}, \mathrm{and}\ \mathrm{c}$ are functions – will depend on the distribution used

Binomial (log) likelihood

$$\log(\Pr(n = r|N,p)) = \log\left(\frac{N!}{r!\,(N-r)!}\right) + r\log p + (N-r)\log(1-p)$$

$$\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)$$

$$\log(\Pr(n = r|N,p)) = \log\left(\frac{N!}{r!\,(N-r)!}\right) + r\log p + (N-r)\log(1-p)$$

Same as in week 2! But does it work as a GLM?

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

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)

$a, \mathrm{b}, \mathrm{and}\ \mathrm{c}$ are functions – will depend on the distribution used

Binomial (log) likelihood

$$\log\binom{N}{r} + \frac{\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/N}$$

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

θ is the expected value (e.g. the mean) = $\log\left(\frac{p}{1-p}\right)$

y is the data = $\frac{r}{N}$

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

ϕ is the variance (dispersion)

$a, \mathrm{b}, \mathrm{and}\ \mathrm{c}$ are functions – will depend on the distribution used

Binomial (log) likelihood

$$\log\binom{N}{r} + \frac{\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/N}$$

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

θ is the expected value (e.g. the mean) = $\log\left(\frac{p}{1-p}\right)$

y is the data = $\frac{r}{N}$

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

ϕ is the variance (dispersion)

$a, \mathrm{b}, \mathrm{and}\ \mathrm{c}$ are functions – will depend on the distribution used

Binomial (log) likelihood

$$\log\binom{N}{r} + \frac{\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/N}$$

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

θ is the expected value (e.g. the mean) = $\log\left(\frac{p}{1-p}\right)$

y is the data = $\frac{r}{N}$

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

ϕ is the variance (dispersion)

$a, \mathrm{b}, \mathrm{and}\ \mathrm{c}$ are functions – will depend on the distribution used

Binomial (log) likelihood

$$\log\binom{N}{r} + \frac{\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/N}$$

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

θ is the expected value (e.g. the mean) = $\log\left(\frac{p}{1-p}\right)$

y is the data = $\frac{r}{N}$

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

ϕ is the variance (dispersion)

$a, \mathrm{b}, \mathrm{and}\ \mathrm{c}$ are functions – will depend on the distribution used

Binomial (log) likelihood

$$\log\binom{N}{r} + \frac{\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/N}$$

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

θ is the expected value (e.g. the mean) = $\log\left(\frac{p}{1-p}\right)$

y is the data = $\frac{r}{N}$

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

ϕ is the variance (dispersion)

$a, \mathrm{b}, \mathrm{and}\ \mathrm{c}$ are functions – will depend on the distribution used

Binomial (log) likelihood

$$\log\binom{N}{r} + \frac{\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/N}$$

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

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

$y = \frac{r}{N}$

Yay, it fits the same format too!

Binomial (log) likelihood

$$\log\binom{N}{r} + \frac{\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/N}$$

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

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

$y = \frac{r}{N}$

Also – we can see our link function (again)

Binomial (log) likelihood

$$\log\binom{N}{r} + \frac{\frac{r}{N}\log\left(\frac{p}{1-p}\right) + \log(1-p)}{1/N}$$

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

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

$y = \frac{r}{N}$

Also – we can see our link function (again) called logit

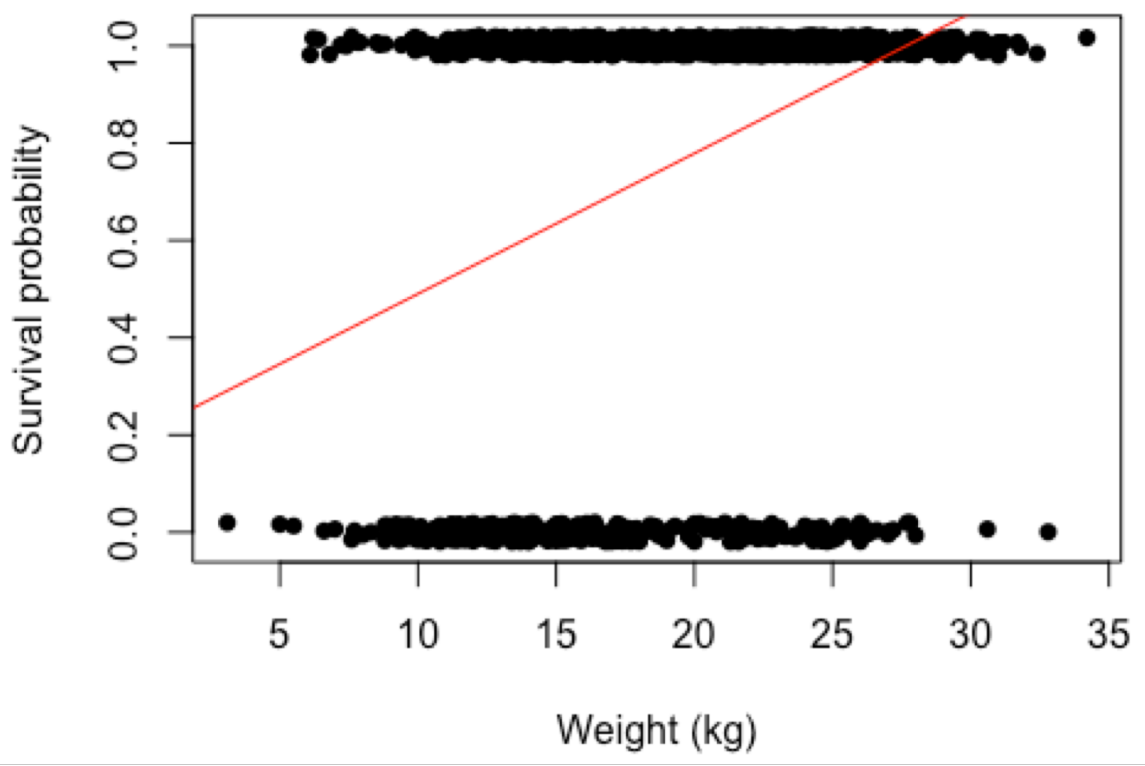
When and why to use a binomial GLM?

Exercise 1: The data

- Data from 1986 to 1996
- Population of Soay sheep
- Today we will look at what influences survival
- Data on survival, body weight, age, year, and population size

Exercise 1: Why is a straight line bad?

- Look at the data below, a straight line is fitted. Why is this a bad idea? (Think of several reasons)
- If you get stuck, think about what survival probability you would predict based on this line for a body weight of 40kg
- What shape line do you think would fit better?



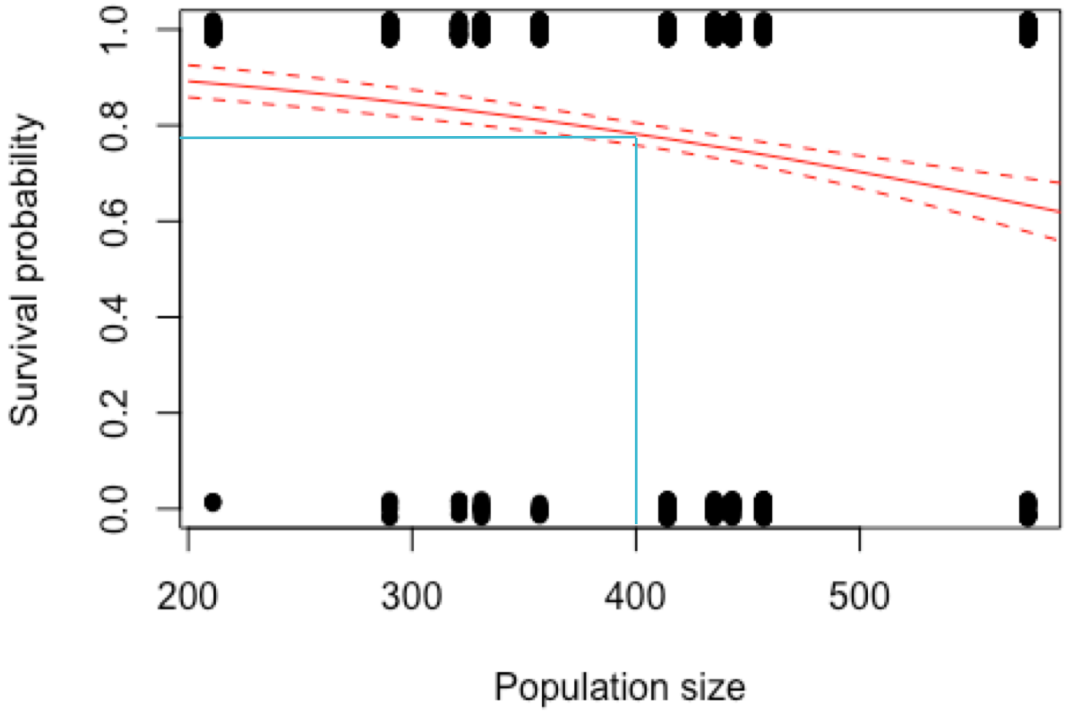
Binomial GLM

Still regression based

Aim is to predict Y values for a given X

Here Y is probability being in state 1

Creates curved lines bounded at 0 and 1



Exercise 2: Fitting a Binomial GLM

- Data at https://www.math.ntnu.no/emner/ST2304/2019v/Week12/SheepData.csv
- Fit a Binomial GLM to answer "Does body weight influence survival probability in sheep?"
- Look at result using coef() and confint()
- What do the coefficients represent? (i.e. the (Intercept) and Weight) – don't worry about the link just think where they fit into $Y_i = \alpha + \beta X_i + \varepsilon_i$

glm(Y ~ X, data, family = binomial(link=logit))

Ways of fitting a GLM

Option 1 for fitting was as a single factor (Here)

Option 2 for fitting (Alternative)

You can make two columns (one of success and one of failures)

Accounts for number of trials (Number of trials is number in population)

```
Y <- cbind(NumSurvived, NumDied)
model2 <- glm(Y ~ PopSize, family = binomial(link=logit))
```

	NumSurvived	NumDied	PopSize
[1,]	50	161	211
[2,]	73	217	290
[3,]	124	197	321
[4,]	98	233	331
[5,]	136	221	357
[6,]	109	305	414
[7,]	99	336	435
[8,]	128	315	443
[9,]	50	407	457
[10,]	174	401	575

More on link functions – Logit link

Link functions

We have seen where we can find the link function (help, equation, google)

Different link functions for different distributions

Link functions have a name and an equation

Used the log link last week

This week - logit

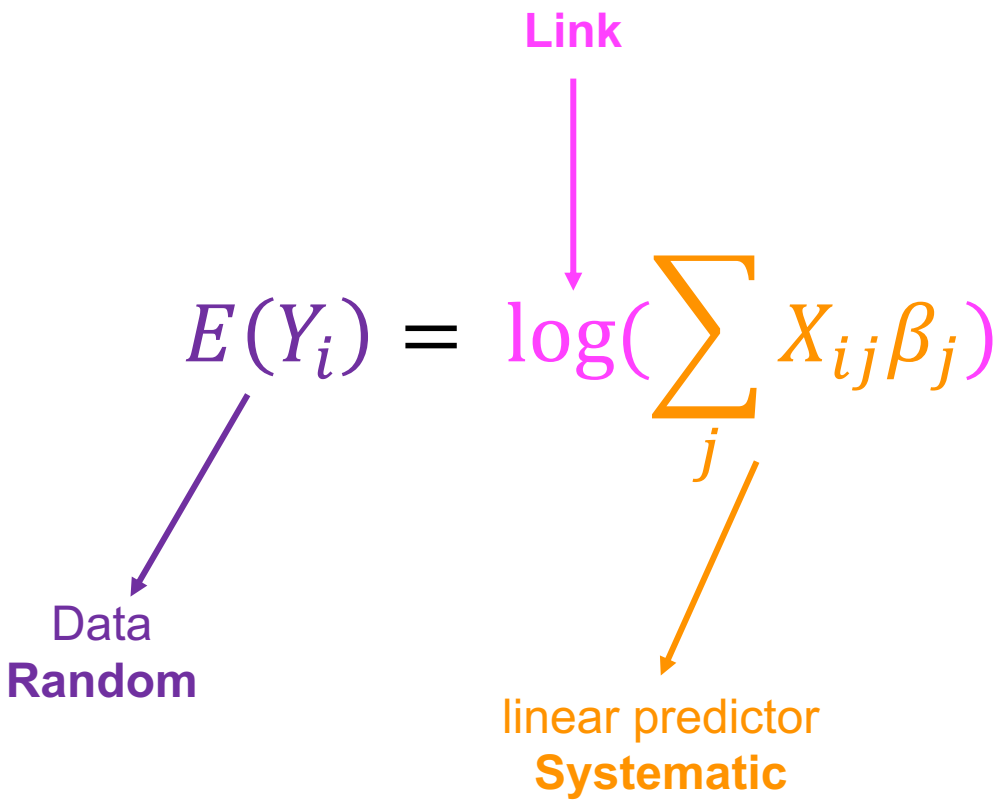
Link functions

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

Data
Random

linear predictor
Systematic

Link functions



Logit link

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

The inverse is

$$p = \frac{e^{\mu}}{1+e^{\mu}} = \frac{1}{1+e^{-\mu}}$$

Logit link

From model

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

The inverse is

$$p = \frac{e^{\mu}}{1+e^{\mu}} = \frac{1}{1+e^{-\mu}}$$

Back to original scale

Logit link

From model

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

The inverse is

Back to original scale

$$p = \frac{e^{\mu}}{1+e^{\mu}} = \frac{1}{1+e^{-\mu}}$$

μ = log odds, p = probability

Logit link

From model

$$\mu = \log \frac{p}{1-p}$$

The inverse is

$$p = \boxed{\frac{e^{\mu}}{1+e^{\mu}}} = \boxed{\frac{1}{1+e^{-\mu}}}$$

Back to original scale

μ = log odds, p = probability

Interpreting the logit

E.g. Betting:

Odds of **10:1** (if you bet 1kr you win 10kr)

Probability for this = Success/(Success+Failure)

1/(1+10) = 1/11 = 0.09

* image from Wikipedia

Interpreting the logit

```
  (Intercept)         Total
-0.7191669164 -0.0006582301
```

Interpreting the logit

(Intercept)	Total
-0.7191669164	-0.0006582301

$$\alpha + \beta X_i$$

Interpreting the logit

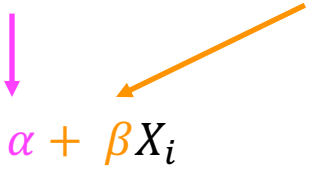
(Intercept)	Total
-0.7191669164	-0.0006582301

$\alpha + \beta X_i$

$$\frac{1}{1 + e^{-\mu}}$$

Interpreting the logit

(Intercept)	PopSize
2.945	-0.004



$$\alpha + \beta X_i$$

$$\frac{1}{1 + e^{-\mu}} = p$$

E.g.

For X (PopSize) = 300

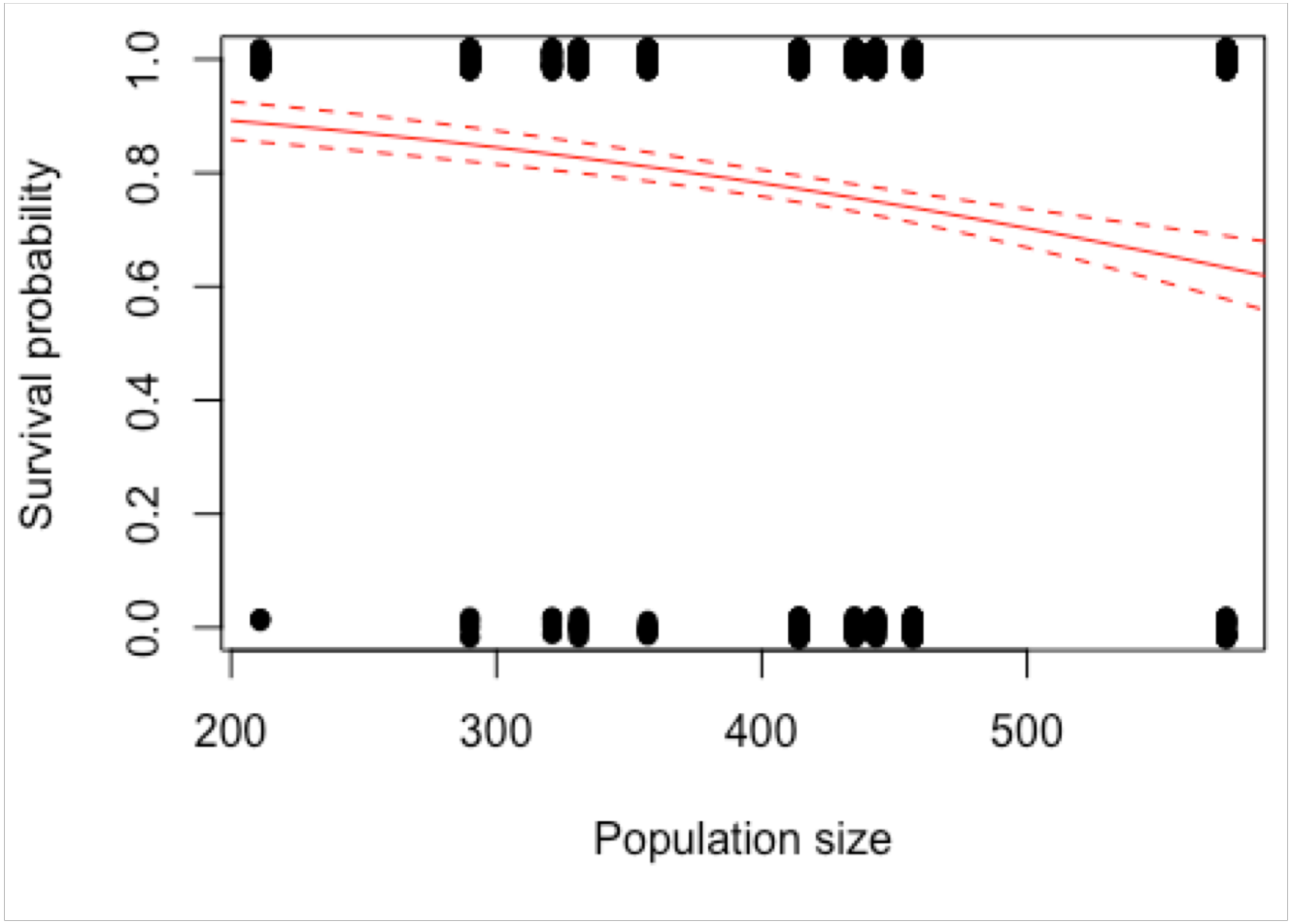
$$\frac{1}{1 + e^{-(2.945 + (-0.004 * 300))}} = 0.85$$

For X (PopSize) = 400

$$\frac{1}{1 + e^{-(2.945 + (-0.004 * 400))}} = 0.79$$

Interpreting the logit

or plot it!



Interpreting (and plotting) results from a Binomial GLM

Exercise 3: Interpretation

- Look at the coefficient values, what can these tell us about the relationship between survival probability and weight?
- Use the inverse link equation to work out the probability of survival 0 weight (the intercept)
- Use the inverse link to work out the change in probability of survival between the mean body weight (20kg) and one standard deviation above the mean population size (25kg)

Exercise 3: Help

- Equation for inverse of logit link

$$\text{Your prediction} = \frac{1}{1+e^{-\mu}}$$

Remember: $\mu = \alpha + \beta X_i$

- How to write it in R

```
prediction = 1/(1+exp(-(Intercept + (Slope*X))))
```

- You need to fill in your own intercept, slope, and X values

Exercise 4: Plotting

- Can be easier to interpret by plotting
- Cannot use abline() here
- Instead we predict Y values given our X values
- Follow code below to plot your results – what can you interpret?
- Why might you expect this pattern? (Biology!)

```
# Make some 'new' X values to predict for
newdata <- data.frame(Weight = seq(0, 35, 1))
# predict, including standard error (se.fit)
predictions <- predict(model1, newdata, type="response", se.fit=T)
# plot the predicted values
points(seq(0, 35, 1), predictions$fit, type='l', col=2)
# and plot the confidence intervals
# ± 2 * standard error predictions
points(seq(0, 35, 1), predictions$fit+(2*predictions$se.fit), type='l', col=2, lty=2)
points(seq(0, 35, 1), predictions$fit-(2*predictions$se.fit), type='l', col=2, lty=2)
```

Rest of this week

Tomorrow will go through example together

Exercise will be whole data analysis yourselves

Lecture Outline

What is the binomial GLM (logistic regression)?

When and why to use a binomial GLM?

More on the logit link function

Interpreting (and plotting) results from a binomial GLM

Lecture Outline

What is the binomial GLM (logistic regression)?
Model we can use for binary responses. Bounded 0 to 1 and non-linear

When and why to use a binomial GLM?

More on the logit link function

Interpreting (and plotting) results from a binomial GLM

Lecture Outline

What is the binomial GLM (logistic regression)?
Model we can use for binary responses. Bounded 0 to 1 and non-linear

When and why to use a binomial GLM?
For binary response to predict within the possible outcomes

More on the logit link function

Interpreting (and plotting) results from a binomial GLM

Lecture Outline

What is the binomial GLM (logistic regression)?
Model we can use for binary responses. Bounded 0 to 1 and non-linear

When and why to use a binomial GLM?
For binary response to predict within the possible outcomes

More on the logit link function
Produces log odds, now know the equation for it and the inverse

Interpreting (and plotting) results from a binomial GLM

Lecture Outline

What is the binomial GLM (logistic regression)?
Model we can use for binary responses. Bounded 0 to 1 and non-linear

When and why to use a binomial GLM?
For binary response to predict within the possible outcomes

More on the logit link function
Produces log odds, now know the equation for it and the inverse

Interpreting (and plotting) results from a binomial GLM
Plotting can be easier to interpret, need to predict not abline()