



Generalised Linear Models

Why this

- Generalised Linear Models are not Bayesian
- But they are used extensively in statistics
- The ideas form the basis of many analyses
- Extensions will come later, where the Bayesian approach is very useful

Regression

- A simple linear regression:

$$y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- $E(y_i) = \alpha + \beta x_i$

- More generally a multiple regression can be written as

$$y_i \sim N(\mu_i, s^2)$$

$$\mu_i = \sum_{j=1}^p \beta_j x_{ij}$$

- The intercept is β_1 ($x_{i1}=1$)
- Note the model is additive

ANOVA

- ANOVA is the same as regression
- Use dummy variables:

Level	X_1	X_2	X_3
1	1	0	0
2	0	1	0
3	0	0	1

- Regression with these variables
- Computer can work these out from a model, and convert back afterwards

Not normal

- Regression assumes the error is normally distributed
- Often this is not the case
 - counts of individuals
 - proportion of tests that are positive
- Need a more flexible approach

Fish: a motivation

- Imagine sitting on the banks of the Seine catching fish
- A natural model:
 - Fish go past at a constant rate, λ
 - caught with a probability p
 - in time t , expect to catch $p\lambda t$
- Number caught: Poisson distributed
- Look at changes in the rate, and how they affect the number caught

Go forth and multiply?

- The rate of capture is positive
- It is natural to think about multiplicative changes
 - the rate doubles or halves
- So, if the rate in the control is $p\lambda t$, then a treatment might have a rate $r = \alpha p\lambda t$,
 - $\alpha = 2$ would double the rate
- Now take logs:
- $\log(r) = \log(\alpha) + \log(p) + \log(\lambda) + \log(t)$
- Additive!

Poisson Regression

- More generally, the model would be

$$N_i \sim \text{Po}(\mu_i) \quad \log(\mu_i) = \sum_{j=1}^p \beta_j x_{ij}$$

- So, $\log(\mu_i)$ is additive, not μ_i
- More generally, the data is a function of $f(\mu_i)$

Generalised Linear Models

- Random Component
 - $y_i \sim \text{Po}(\mu_i)$
- Systematic Component
 - $\eta_i = \sum_j x_{ij} b_j$
 - linear predictor
- Link function
 - $\eta_i = g(\mu_i)$
 - e.g. $\eta_i = \log(\mu_i)$
 - the normal distribution has an identity link (i.e. $\eta_i = \mu_i$)

Why does this work?

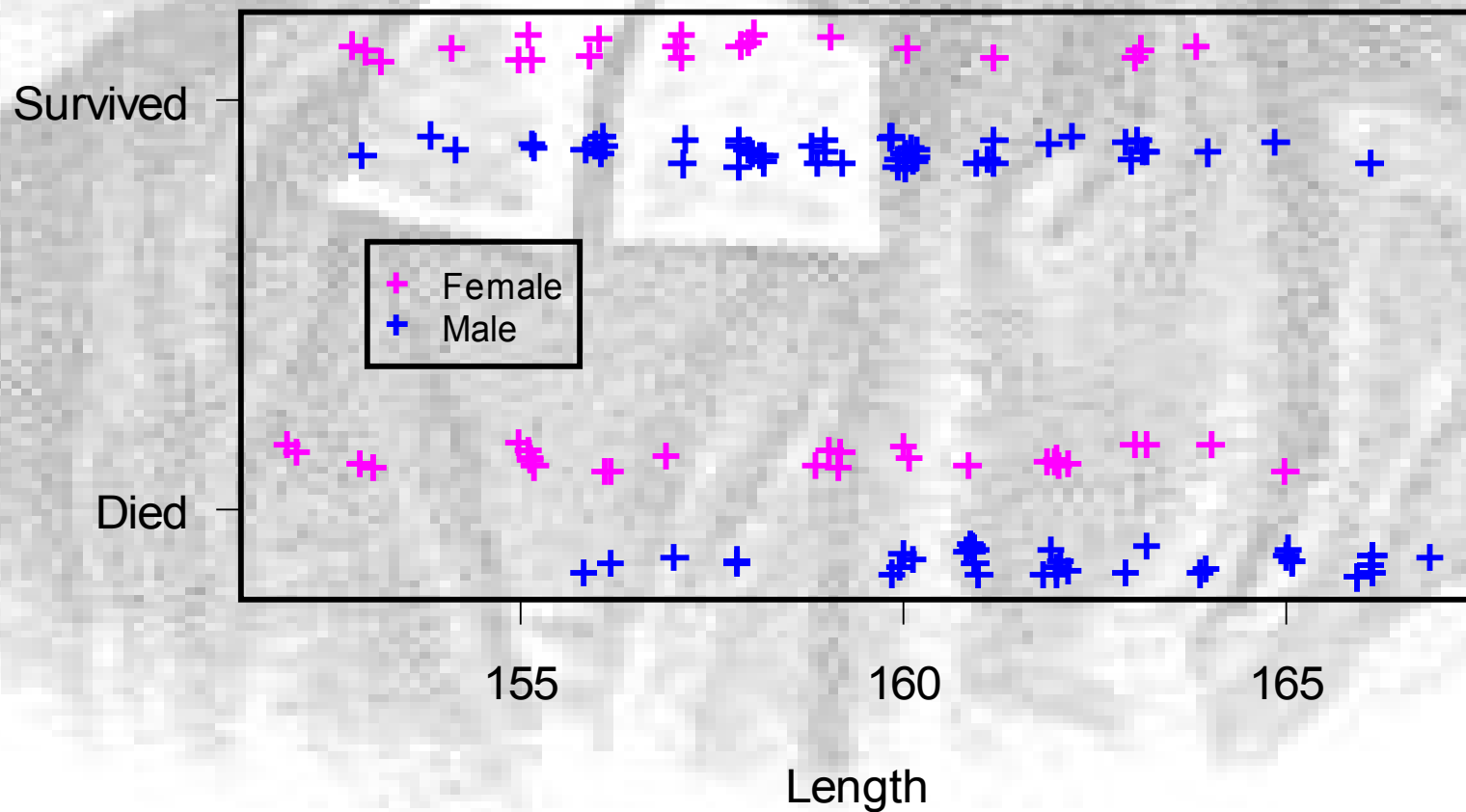
- The systematic part of the model is built in the same way for everything
 - just adding terms
 - not restricted in range
- The random component can be chosen from a range of distributions
 - e.g. normal, Poisson, binomial, gamma
 - technically, have to be from the exponential family
- The link joins the two
 - note that μ_i could be bounded

Example: Logistic Regression

- Bumpus (1899) collected sparrows that had been blown off their perches after a storm. He measured their sizes and weights, and recorded which survived.
- Question: what factors affect survival?
- Look at sex and body length

The Data

- Females are smaller
- Larger birds tended to die



The Model

- The probability scale goes from 0 to 1
- The link function must transform from $-\infty$ to ∞
- A couple of choices
- Natural: logit link

$$\eta = g(p) = \log\left(\frac{p}{1-p}\right) \quad p = g^{-1}(\eta) = \log\left(\frac{e^\eta}{1+e^\eta}\right)$$

Logit link

- All GLMs have a “natural” link function
 - drawn from statistical theory
- They usually have sensible interpretations
- For a binomial, the logit link is canonical
- Interpretation: log odds
- $O = p/(1-p)$ is the odds
 - for every failure, there are O successes
 - if O is close to 1, $\log(O) \approx O-1$
 - e.g. $O=1.1$, $\log(O)=0.095$

The Model

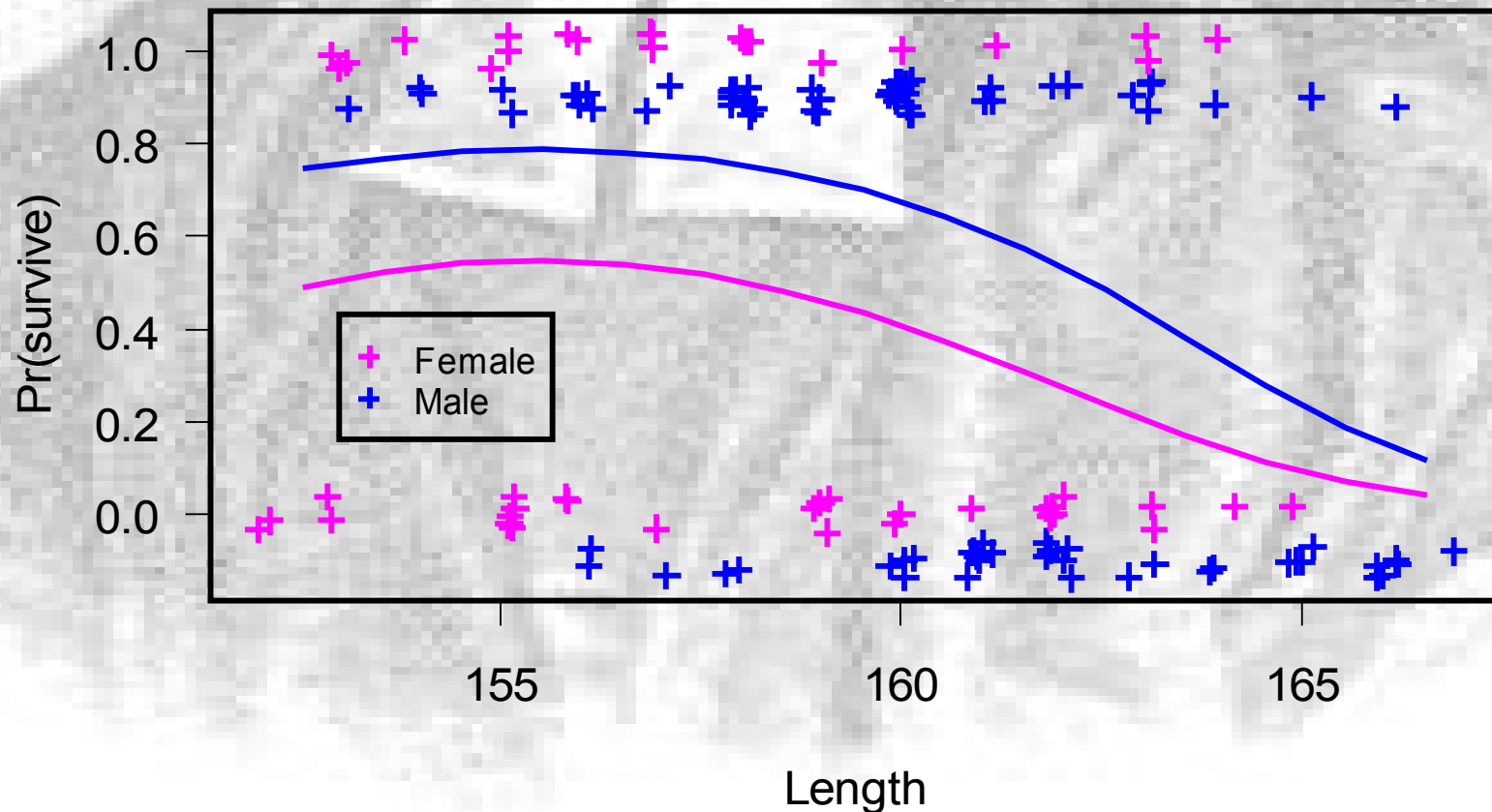
- For each individual
- $X_i \sim \text{Bin}(1, p_i)$
 - $X_i = 1$ if the individual survives
 - p_i = probability of survival

$$\log\left(\frac{p}{1-p}\right) = \alpha_{S(i)} + \beta_1 l_i + \beta_2 l_i^2$$

- $S(i)$ – sex of i^{th} individual
 - l_i – body length
- $\alpha_{S(i)}, \beta_1, \beta_2$ - parameters

Results

- Males have higher survival probability
- Survival decreases with size
 - threshold at small sizes?



Common GLMs

- ANOVA/regression
 - normal, identity link
- log-linear models
 - Poisson, log link
 - count data
- Logistic regression
 - Binomial, logit link (or probit, or cloglog)
 - success/failure trials
- Failure times/survival analysis
 - exponential, log link or Gamma, log or inverse link

Link to Bayesian Analysis

- GLMs are >35 years old
- They separate the components of the model into pieces
- They are not Bayesian, and usually it is not worth using a Bayesian approach to fit them
 - ML is quicker, and the results are usually the same
- But they form the basis of a lot of Bayesian models
 - the same structure is used in a larger framework