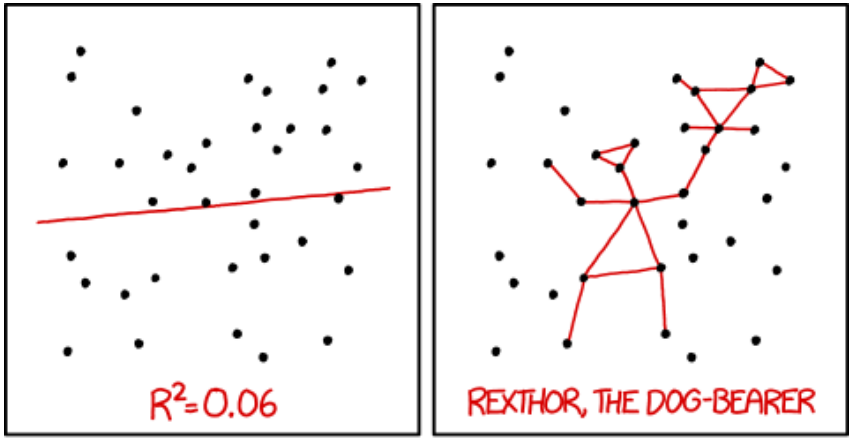
Generalised linear models

Bert van der Veen

Department of Mathematical Sciences, NTNU

So far: linear models

$$y_i = \alpha + \sum_{k=1}^{p} x_{ik}\beta + \epsilon_i \qquad (1)$$



So far: linear models

$$y_i = \alpha + x_i\beta + \epsilon_i \quad (2)$$

- y_i the data
- $\alpha + x_i\beta$ the systematic component: "linear predictor"
- ϵ_i the random component: "error"

Key assumptions

- Linearity (straight line)
- Independence of errors
- Homoscedasticity (same variance for all errors)
- Normality (distribution of errors)

Real ecological data do not usually follow these assumptions

and don't range $(-\infty, \infty)$

Generalised linear models (GLMs)

GLMs as a framework were introduced by Nelder and Wedderburn (1972) uniting many different models. With a special focus on teaching statistics.

- Linear regression
- Logistic regression
- Probit regression
- Complementary log-log regression
- Log-linear regression
- Gamma regression

Generalised linear models (2)

GLMs extend the linear model framework to address:

- Variance changes with the mean
- Range of **y** is bounded

The basis of many statistical models in Biology

Many results are now asymptotic.

Components of a GLM

- Systematic component: η
- Random component: data/distribution)
- The link function: connects these components
  - This is not a data transformation
- The variance function

But no explicit error term

GLM Likelihood

- We still use MLE for estimation
- But now a different likelihood function (in EF for fixed ϕ)

All GLMs can be formulated as:

$$\mathcal{L}(y_i; \Theta) = \exp\{\frac{y_i \eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\} \tag{3}$$

Linear regression as GLM

Previously:

$$\mathcal{L}(y_i; \Theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\} \tag{4}$$

Now:

$$\mathcal{L}(y_i; \Theta) = \exp\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\} \tag{5}$$

- for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$
- for normal distribution: $\eta_i = \mu_i$, $a(\phi) = \sigma^2$, $b(\eta_i) = -\mu^2/2$, $c(y_i, \phi) = -\frac{1}{2}\{y^2/\sigma^2 + \log(2\pi\sigma^2)\}$

The linear model

Writing the linear model:

$$y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

Is the same as:

$$\mathbb{E}(y_i|\mathbf{x}_i) = \alpha + \mathbf{x}_i\beta \tag{7}$$

as long as $\mathbb{E}(\epsilon_i) = 0$.

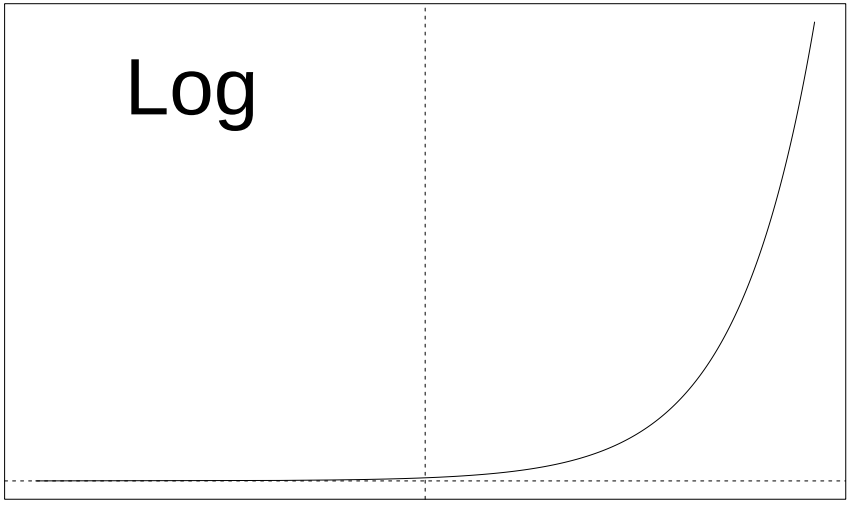
Generalised linear model

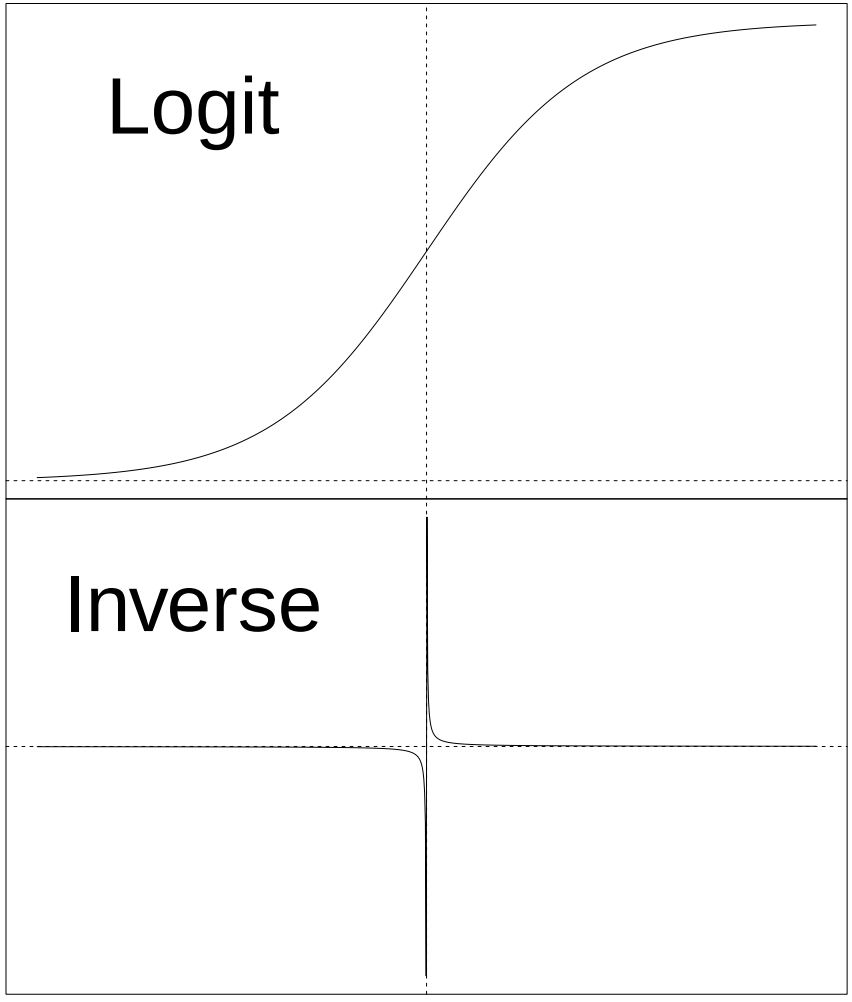
$$
\begin{aligned}
g\{\mathbb{E}(y_i|x_i)\} &= \eta_i = \alpha + x_i\beta \\
\mathbb{E}(y_i|x_i) &= g^{-1}(\eta_i) = g^{-1}(\alpha + x_i\beta)
\end{aligned}
\tag{8}
$$

$\mathrm{g}(\cdot)$ is the **link function**

The link function

- Is a smooth/monotone function
- Has an inverse $g^{-1}(\cdot)$
- Restricts the scale
- g$(\cdot)$ can be e.g.





Variance function

Variance changes with the mean:

$$\text{var}(y_i; \mu_i, \phi) = \frac{\partial^2 g(\eta_i)}{\partial \eta_i^2} a(\phi)$$

- ϕ: the dispersion parameter, constant over observations
  - Fixed for some response distributions
- $a(\phi)$ is a function of the form ϕ / w_i (McCullagh and Nelder 1989)

Assumptions

- No outliers
- Independence
- Correct distribution
- Correct link function
- Correct variance function (implied by previous two)

More on checking assumptions in GLMs tomorrow.

Fitting GLMs

Unlike LMs, parameters in GLMs need to be estimated **iteratively**.

- More difficult to fit
- Requires numerical *optimisation*
- Susceptible to local convergence

Popularisation of GLMs

Nelder and Wedderburn (1972) proposed GLMs as a class to unify different forms of regression.

- Linear regression
- Probit regression
- Logistic regression
- Log-linear regression
- Gamma regression
- Inverse Gaussian regression

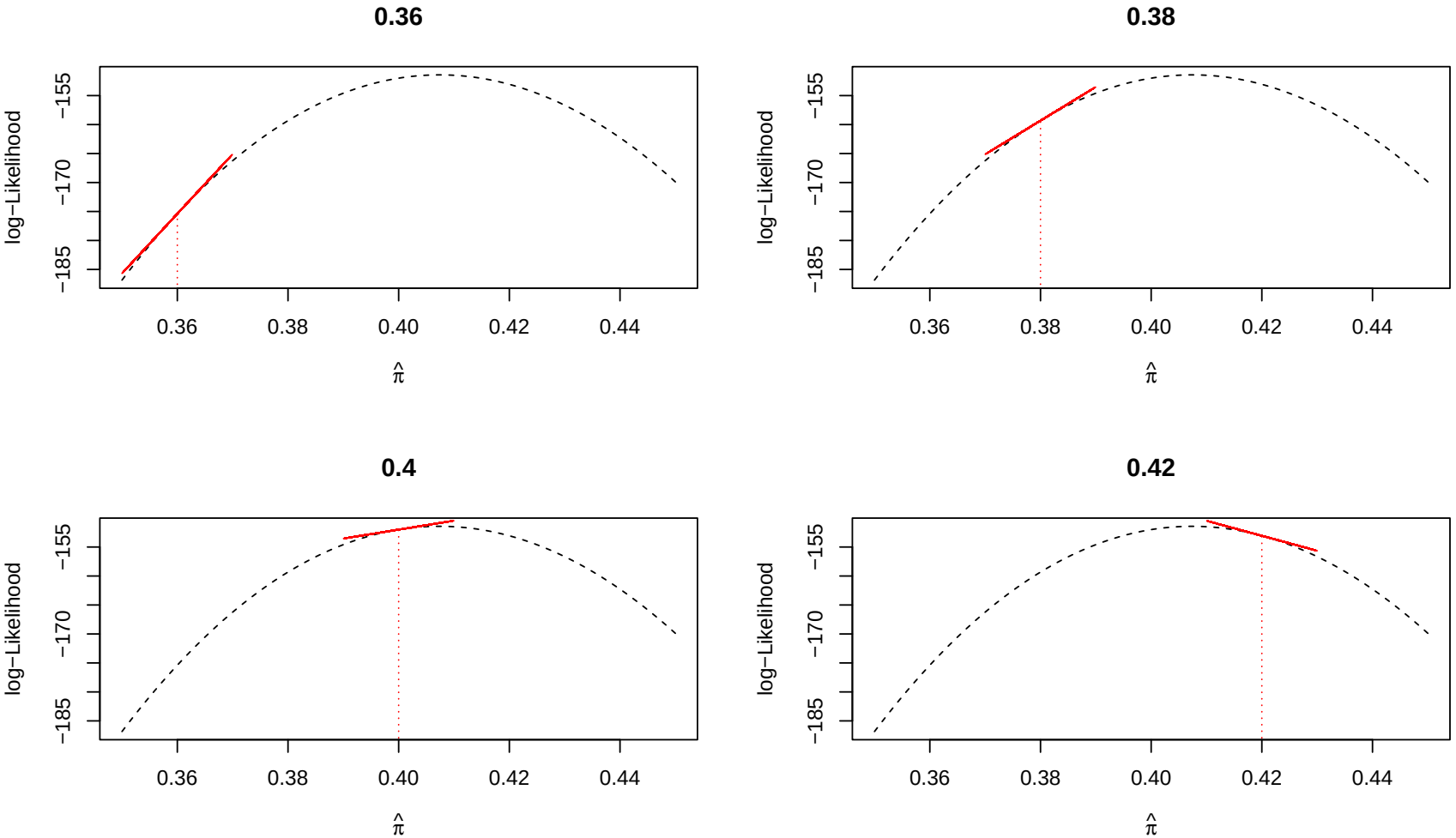
McCullagh and Nelder (1989) wrote a book that popularised the class.

Fitting GLMs

Easy and quick in R.

Mathematically more involved than LMs.

Finding the maximum (from day 1)



We need a good algorithm to find the maximum!

Finding the maximum: GLMs

$$\Theta^{t+1} = \Theta^t + \frac{\partial^2 \log\{\mathcal{L}(\mathbf{y}_i; \Theta^t)\}}{\partial\Theta\partial\Theta^\top}^{-1} \frac{\partial\mathcal{L}(\mathbf{y}_i; \Theta^t)}{\partial\Theta}$$

- (Newton-Rhapson) Can get quite expensive to evaluate.
- Nelder and Wedderburn (1972) instead suggested an algorithm that fits a LM repeatedly.

Iteratively reweighted least squares (IRLs)

1) Start at some (decent) point (e.g., $\hat{\boldsymbol{\eta}} = \mathbf{y} + \boldsymbol{\epsilon}$) (Wood 2017)
2) Set $\mathbf{z}^t = \boldsymbol{\eta} + \frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}\{\mathbf{y} - g^{-1}(\hat{\boldsymbol{\eta}})\}/a(g^{-1}(\hat{\boldsymbol{\eta}}))$
3) weight with $\mathbf{w} = a(g^{-1}(\hat{\boldsymbol{\eta}}))/\{\frac{\partial \hat{\boldsymbol{\eta}}}{\partial g^{-1}(\hat{\boldsymbol{\eta}})}^2 \text{var}(\mathbf{y}_i)\}$
4) Fit weighted LM with $\mathbf{z}$ as pseudodata and $\mathbf{w}$ as weights
5) Repeat until convergence

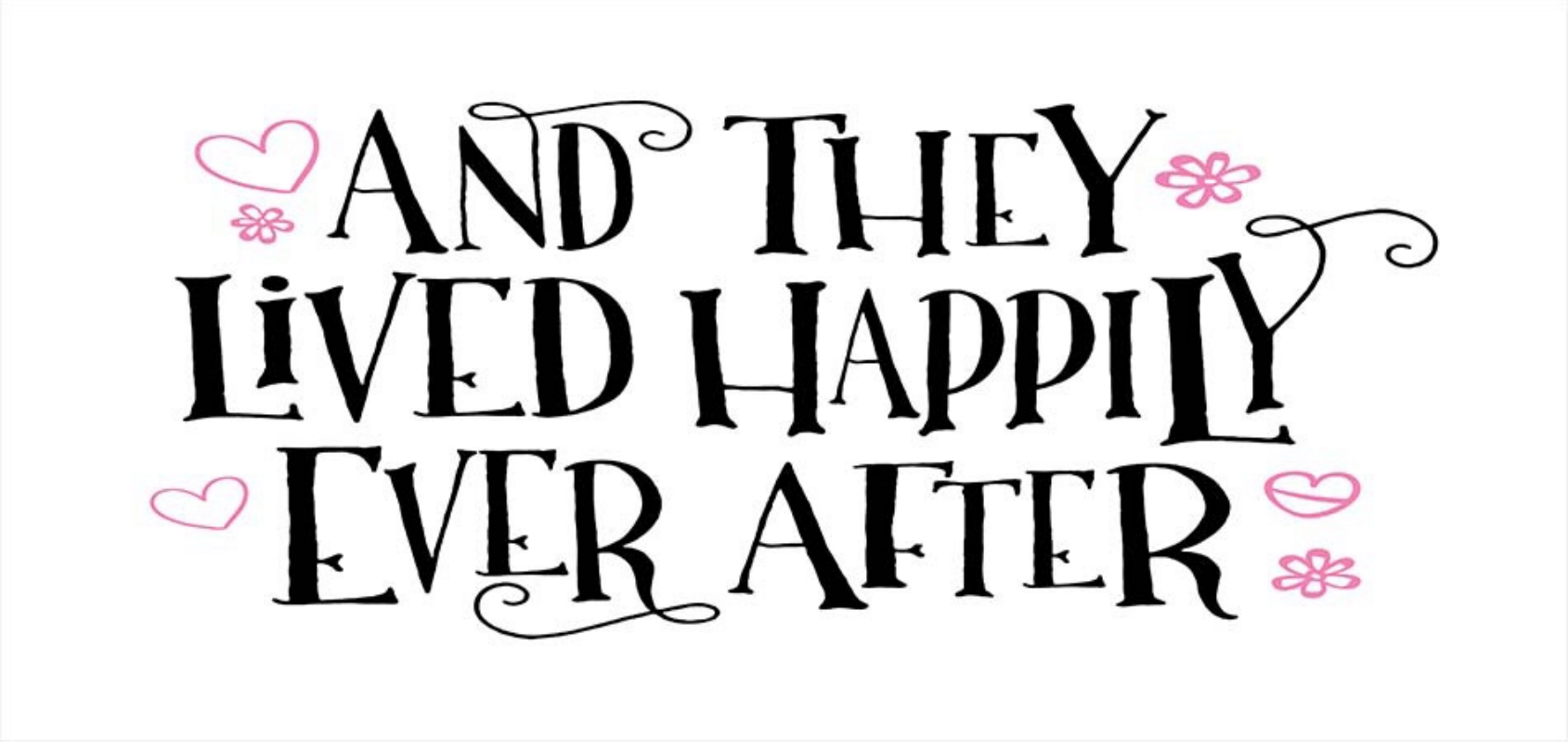
(details omitted)

And that is the day researchers started liking GLMs.

Iteratively reweighted least squares (IRLs)

Prevents having to do numerical optimisation.

The end?



Why is stuff this important?

1) A basic (mathematical) understanding helps apply methods correctly.
2) GLMs may not always converge to the MLE. Then, you will get warnings/errors.
3) If you understand them, you might know what to do! Similar problems in a lot of more complex models (e.g., GLMMs).

Often used GLMs in ecology

- Binomial: occurrence/counts. Presence of species, number of germinated seeds out of a total
- Poisson: counts. Number of fish caught
- Negative binomial (fixed dispersion): counts. Overdispersed fish.
- Gamma: (positive) continuous. Body size
- Ordinal (cumulative link). Plant cover classes

in R

- Similar to the lm() function!
- Now the glm() function

A linear regression:

```
model <- glm(y ~ x, family = gaussian(link = identity), data = data)
```

A glm:

```
model <- glm(y ~ x, family = poisson(link = log), data = data)
```

Output

```
##
## Call:
## glm(formula = y ~ X, family = "poisson")
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5965     0.1820   3.278 0.001045 **
## X             0.6554     0.1810   3.621 0.000294 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 45.250  on 19  degrees of freedom
## Residual deviance: 31.486  on 18  degrees of freedom
## AIC: 78.867
##
## Number of Fisher Scoring iterations: 5
```

Back-transformation

- Confidence intervals can be back-transformed to the response scale
  - as long as we have "monotonicity" of the link function
- Standard errors cannot be back-transformed!

Prediction

$$\hat{\mu}_i = g^{-1}(\alpha + \mathbf{x}_i\boldsymbol{\beta}) \tag{9}$$

```
predict(model, type = "response")#Type = link alternatively
```

Newdata

```
predict(model, newdata = X, type = "response")#Type = link al
```

Intervals for $\widehat{\mu}_i$

- Confidence intervals
- Prediction intervals

More on predicting tomorrow in the practical

Deviance

- LMs used RSS to quantify fit
- GLMs use deviance to quantify **lack of fit**
- Deviance $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is different for every GLM.

Normal: $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$ Poisson: $2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i)$

binomial: $2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i) + (N - y_i)\log\{(N - y_i)/(N - y_i)\}$

gamma: $2\sum_{i=1}^{n} -\log\{y_i/\hat{\mu}_i\} + (y_i - \hat{\mu}_i)/\hat{\mu}_i$

and so on.

Deviance

- Measures (twice) difference to a model that perfectly fits the data ("saturated model")
  - equal to RSS for linear regression
- Can be used for hypothesis testing, calculate $R^2_{deviance}$, and "deviance residuals"

Asymptotic requirements

- Binomial responses: $N\pi \geq 3$ and $N(1-\pi) \geq 3$
  - Deviance for $N = 1$ has no concept of residual variability
- Poisson responses: $\lambda \geq 3$
- Gamma distribution: $\phi \leq 3$

(Dunn and Smyth 2021)

Residuals

- GLMs lack an explicit error term, but it is there!
- So we can still check assumptions by residuals, though they are differently defined
- There are different types; Pearson, Deviance, Anscombe, Quantile residuals..
- We usually -hope- that they are approximately normally distributed
- Residual checking in GLMs can be difficult

(omitted details)

Recap

- Remember to bring all components together
- Parameter estimates, uncertainty, multiple predictors, interaction, model selection
- GLMs for when assumptions of LMs fail (which is very often)
- We covered components of GLMs here
- And how they are fitted in R
- Deviance and residuals