Generalised Linear Models

Dr Josh Hodge

J.Hodge@imperial.ac.uk

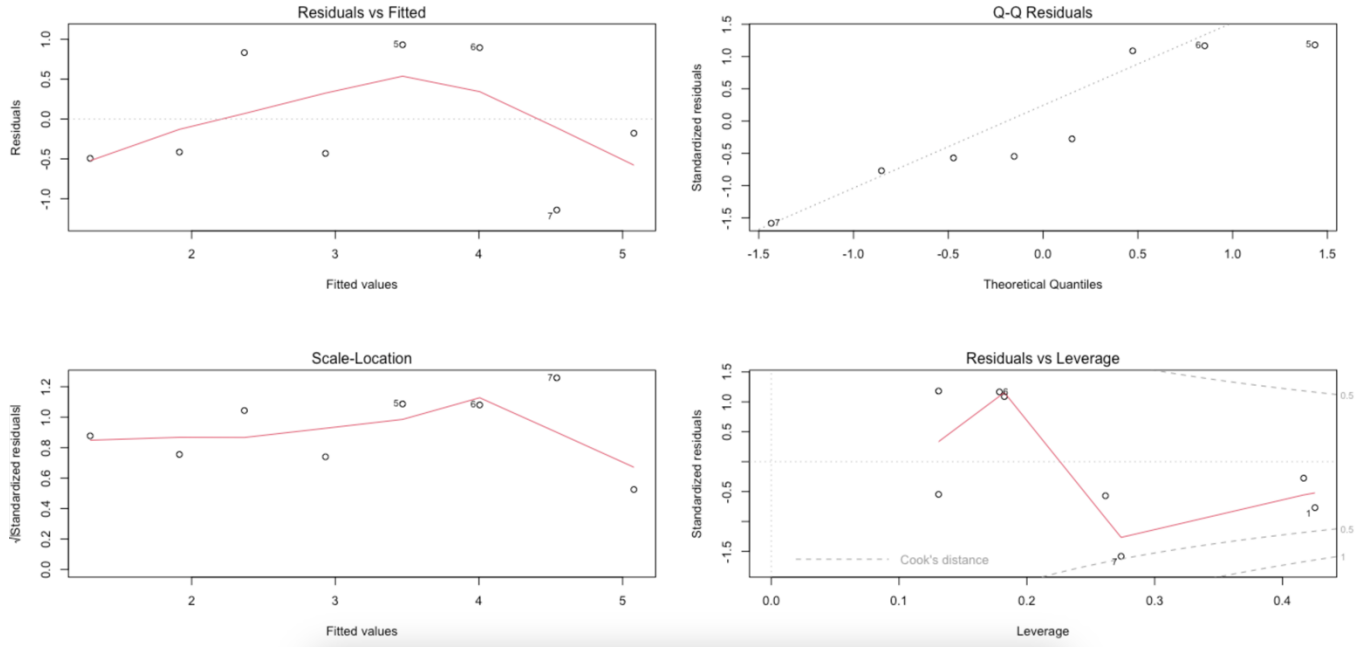
Intended Learning Outcomes

You will be able to:

- Explain and validate linear models by interpreting diagnostic plots
- Differentiate between a linear model and generalised linear models of binomial and poisson data
- Explain why generalized linear models are preferential over simple linear models
- Outline the three steps of generalized linear model

Assumptions of Linear Models

- Defined response/dependent variable
  - Different from correlation
  - Causation
- Independent variables are not collinear
- Relationship is linear
  - Examined through plotting
- Residuals are normally distributed
- Homogeneity of variances
- No outliers



Assumptions of Linear Models

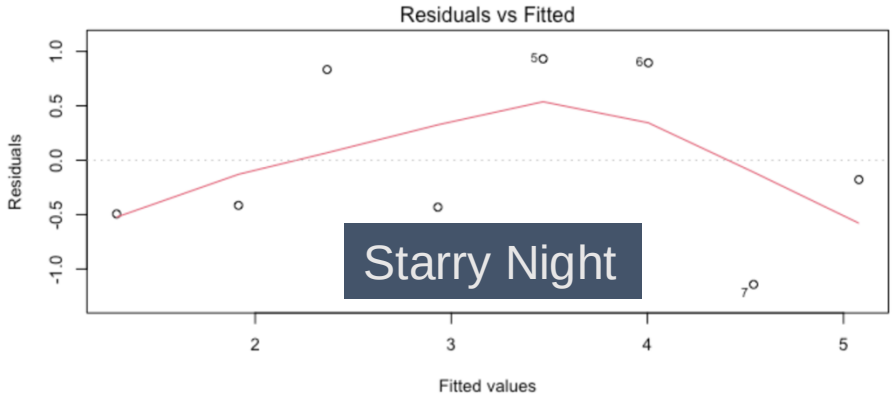
(1) & (3) Homogeneity of Variances

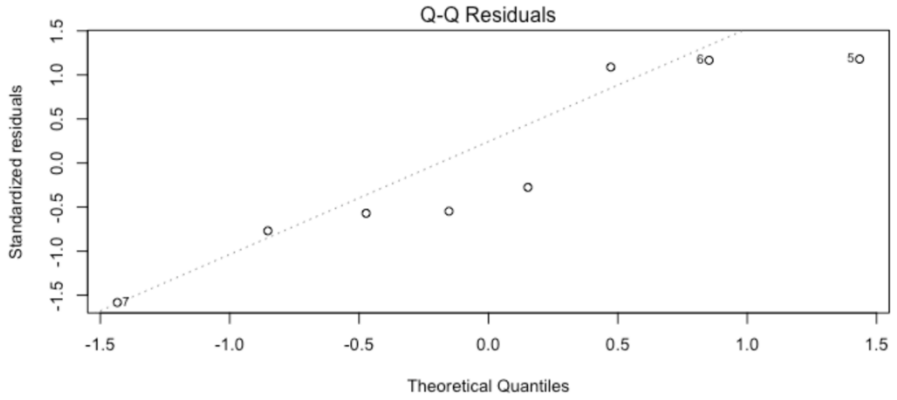
$$\varepsilon_i \sim y_i$$

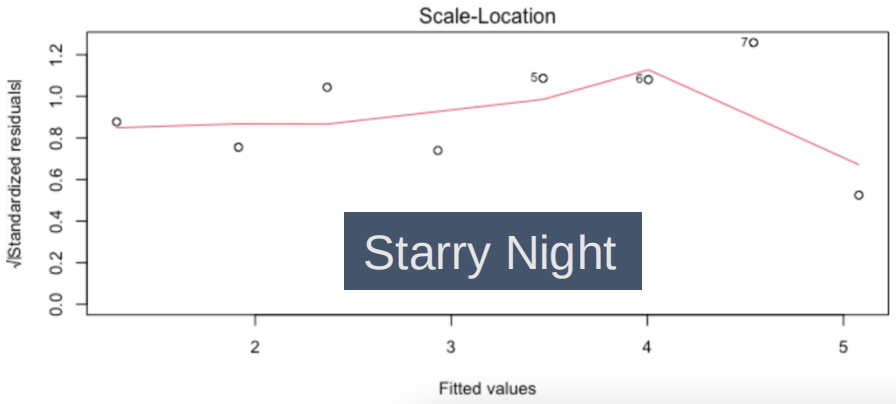
(2) Normality of the residuals

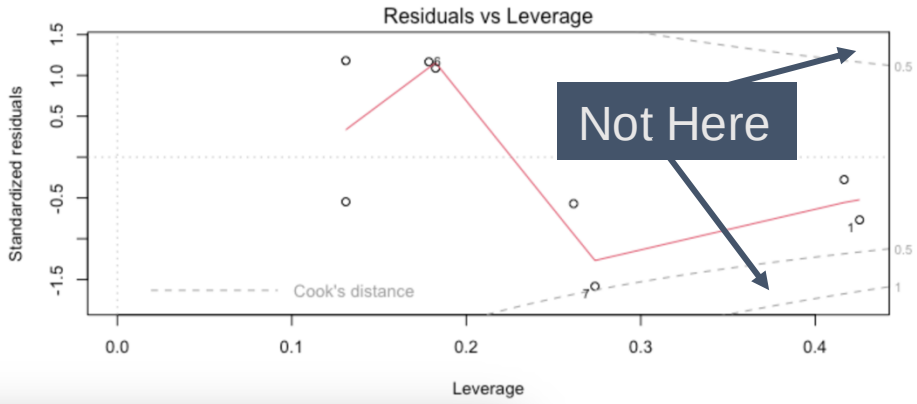
(4) Outliers

$$\varepsilon_i \sim leverage$$

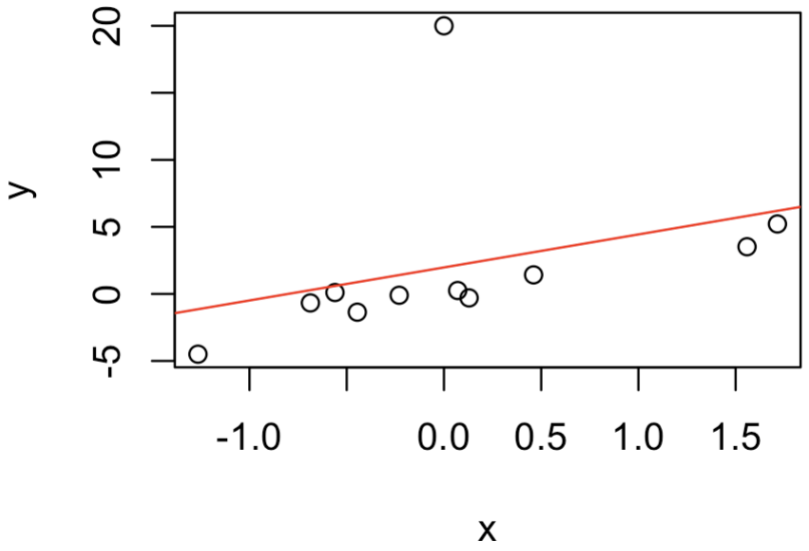


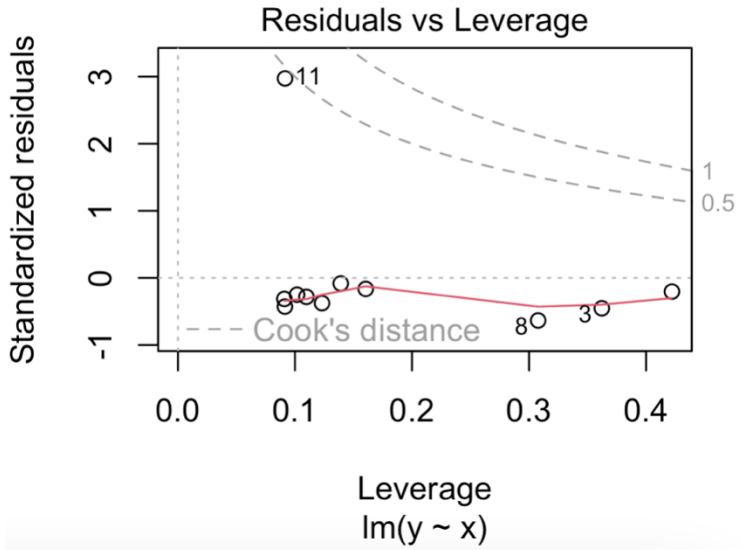


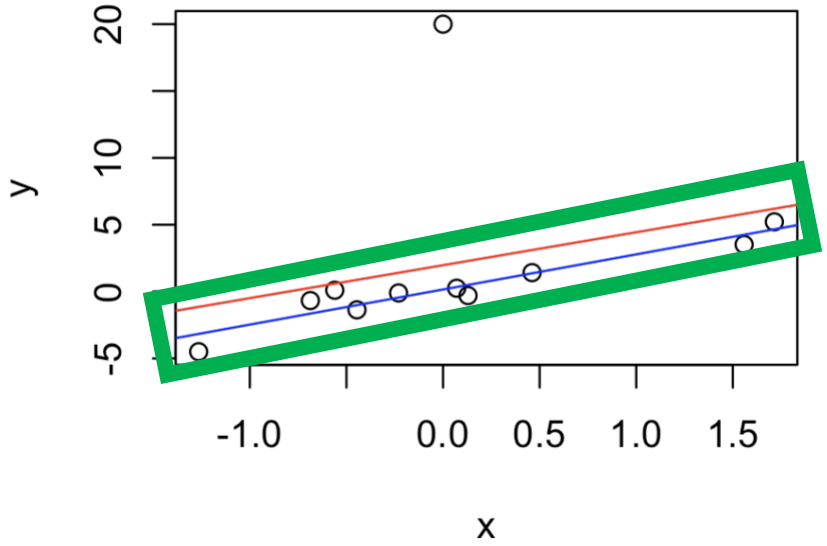




What is leverage?







LMs vs GLMS

- Response variable data type
  - Unconstrained vs Constrained
- Model fitting approach
  - Ordinary Least Squares vs Maximum Likelihood
- Assumptions
  - Means and variance

Linear Models

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Response Variable

Intercept

Slope of Explanatory Variable

Error Term

Linear Models

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Response Variable

Intercept

Slope of Explanatory Variable

Error Term

Linear Models

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

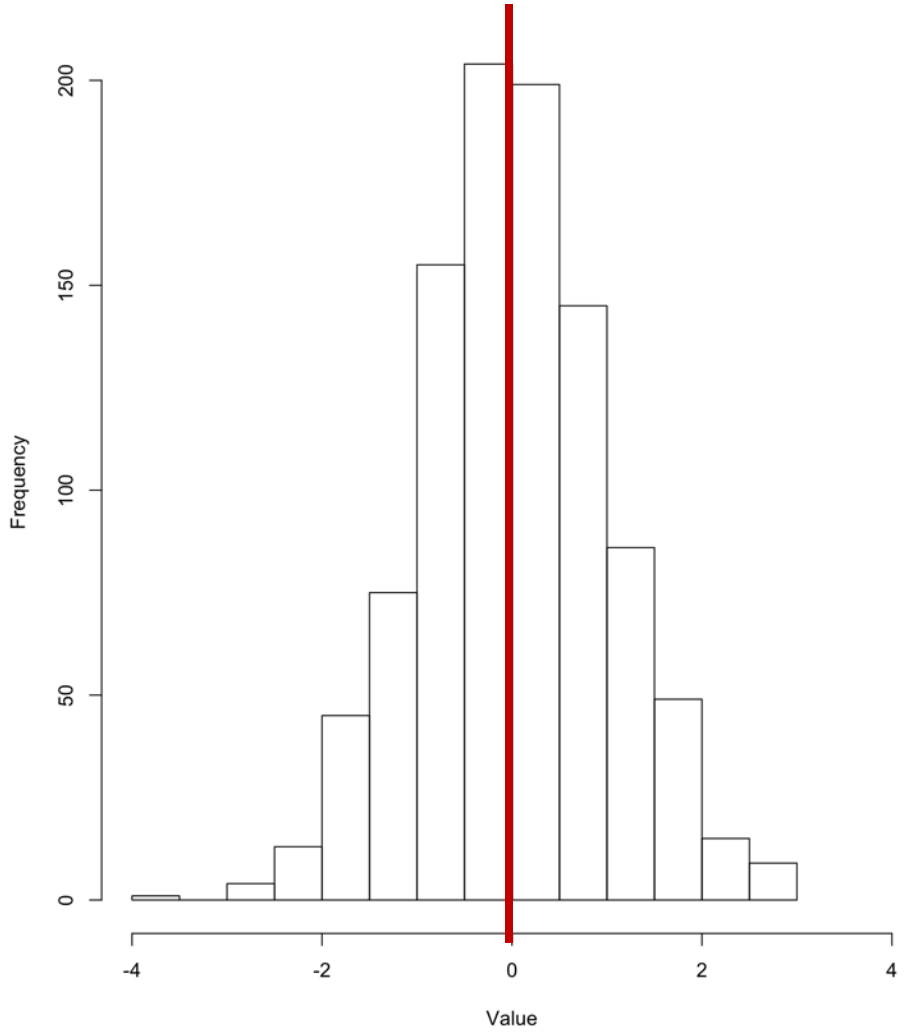
Response Variable

Intercept

Slope of Explanatory Variable

Error Term

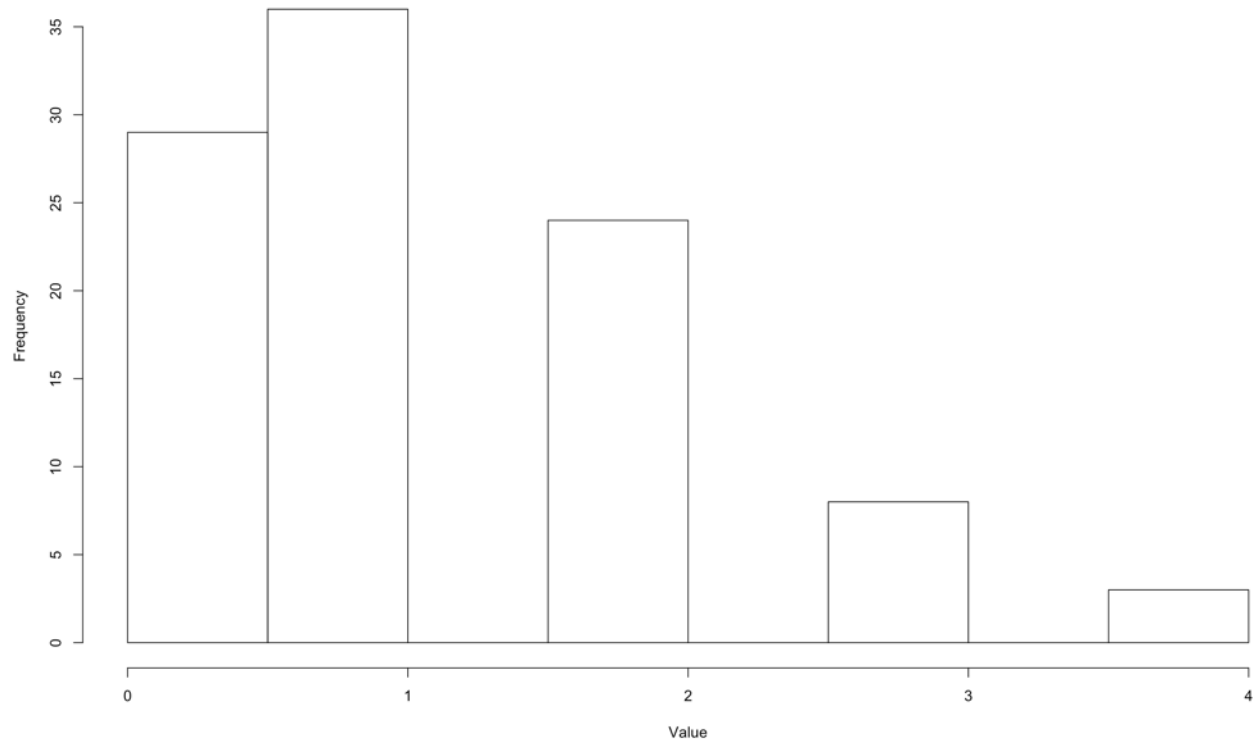
Revisiting Data Distributions- Normal



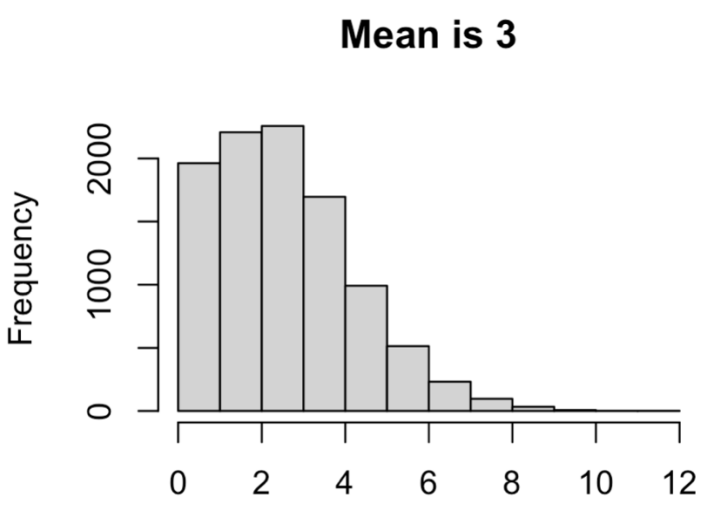
- -Infinity to Infinity
- Mean represents the centre of the data

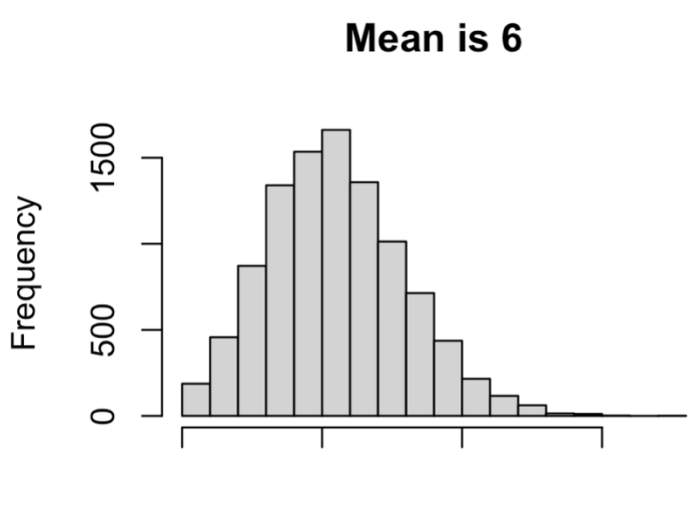
Revisiting Data Distributions- Poisson

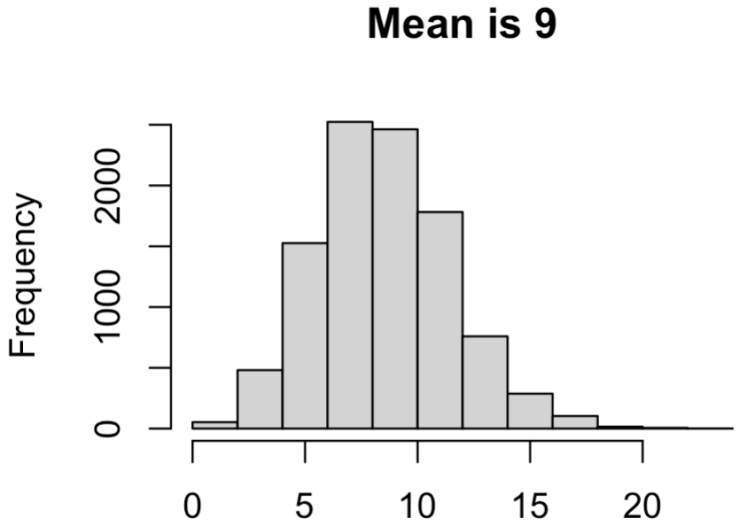
- Count Data
- **Constrained** to absolute whole numbers
- Typically right skewed
- Examples:
  - Number of Species
  - Number of Enzymes
  - Heartbeat
  - Number of Offspring
- Mean is equal to variance??

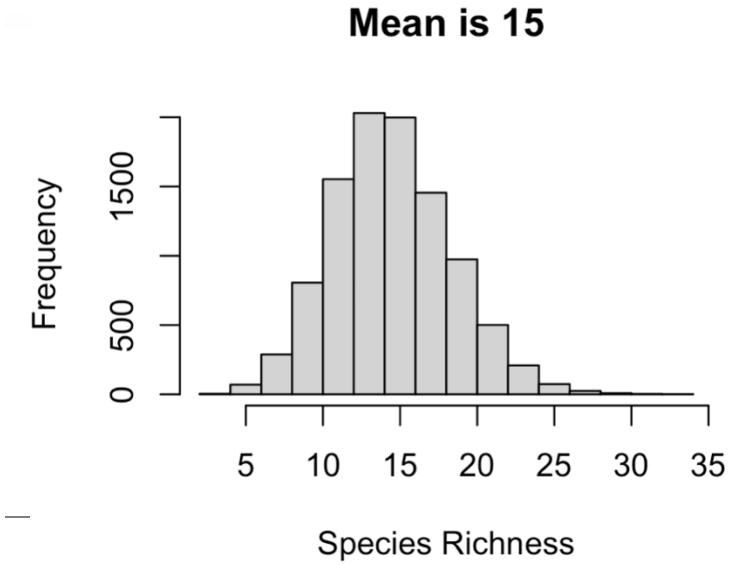


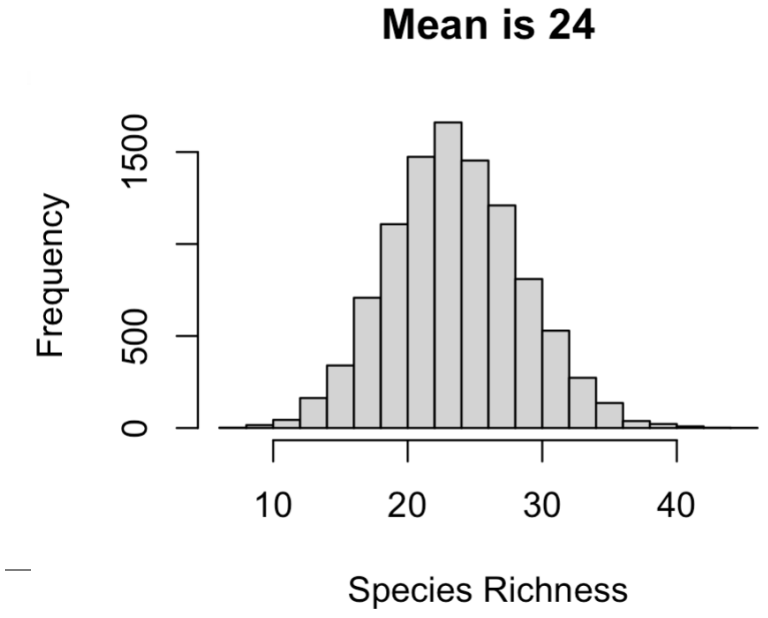
Mean is equal to variance

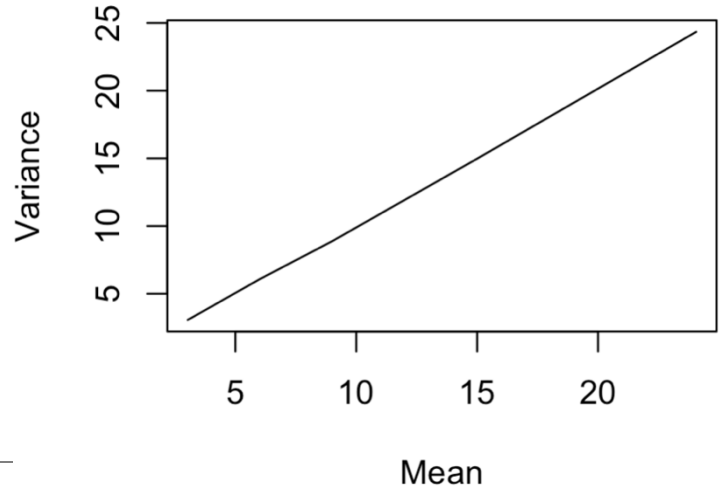






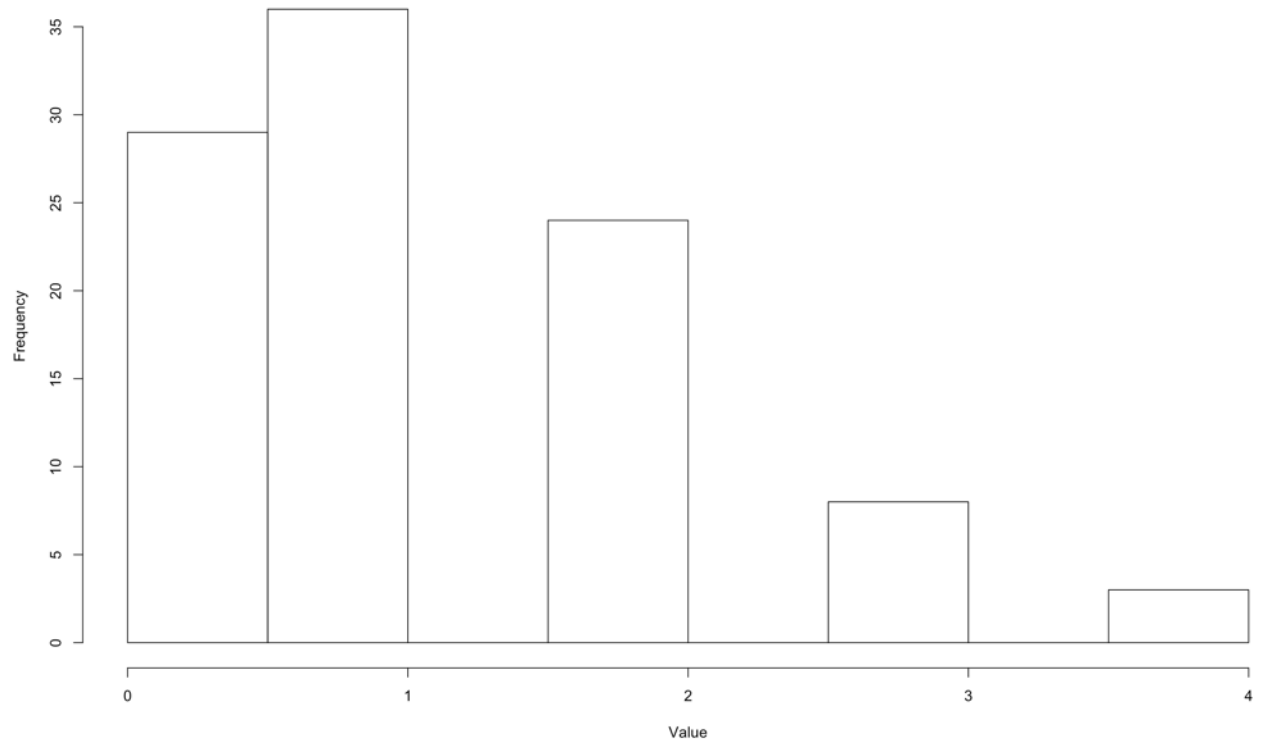




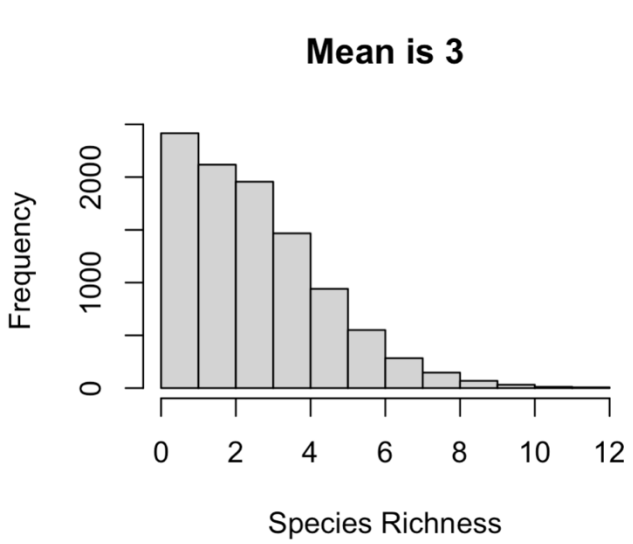


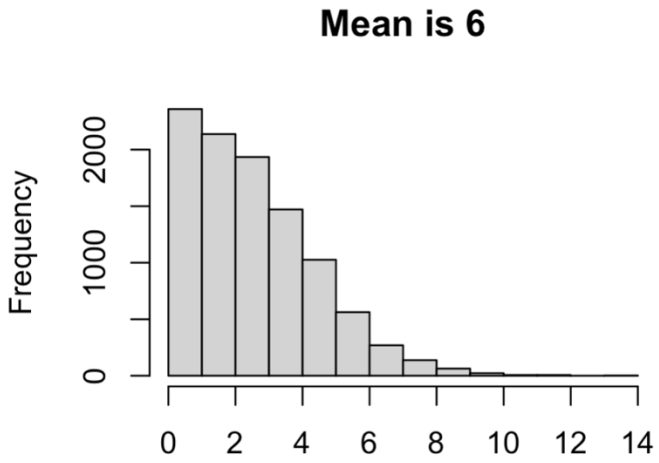
Revisiting Data Distributions- Negative Binomial

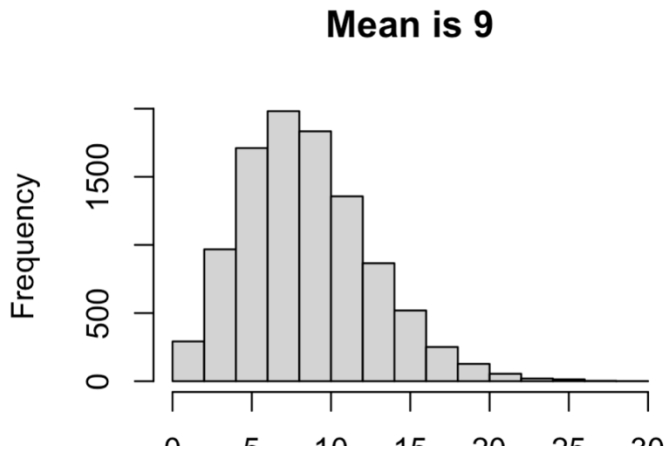
- Count Data
- **Constrained** to absolute whole numbers
- Typically right skewed
- Examples:
  - Number of Species
  - Number of Enzymes
  - Heartbeat
  - Number of Offspring
- Mean is **not** equal to variance

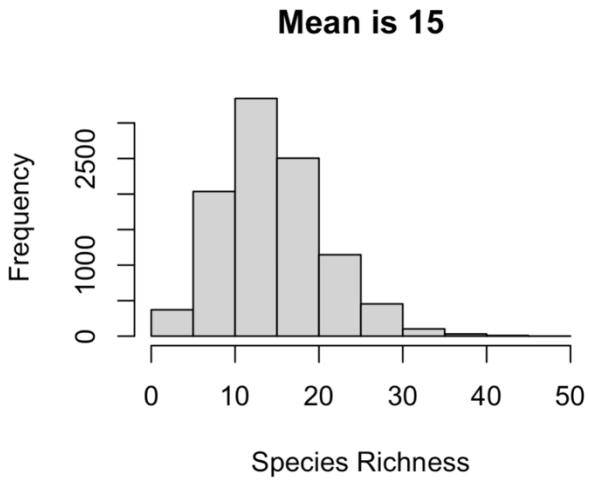


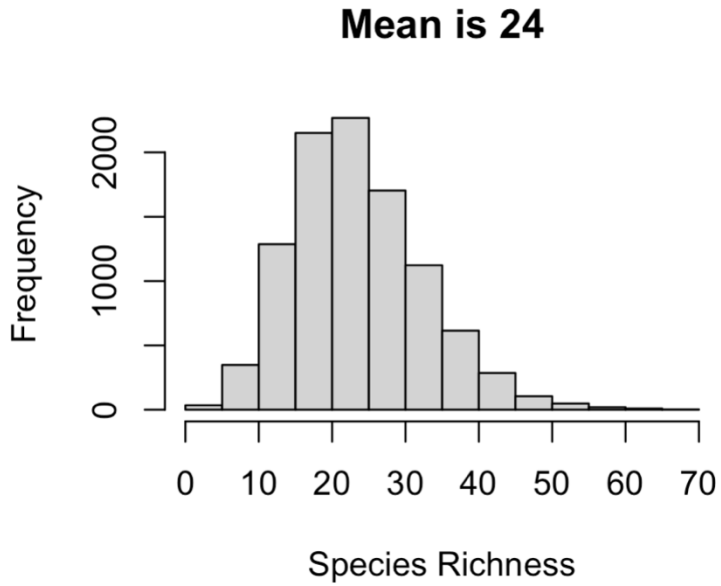
Revisiting Data Distributions- Negative Binomial

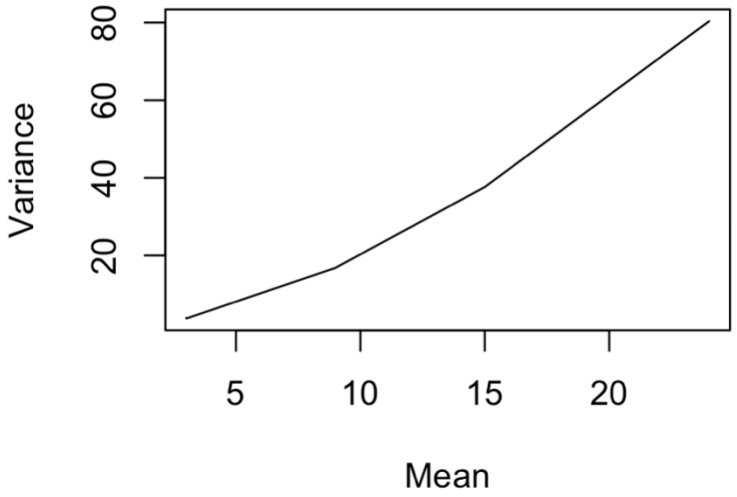








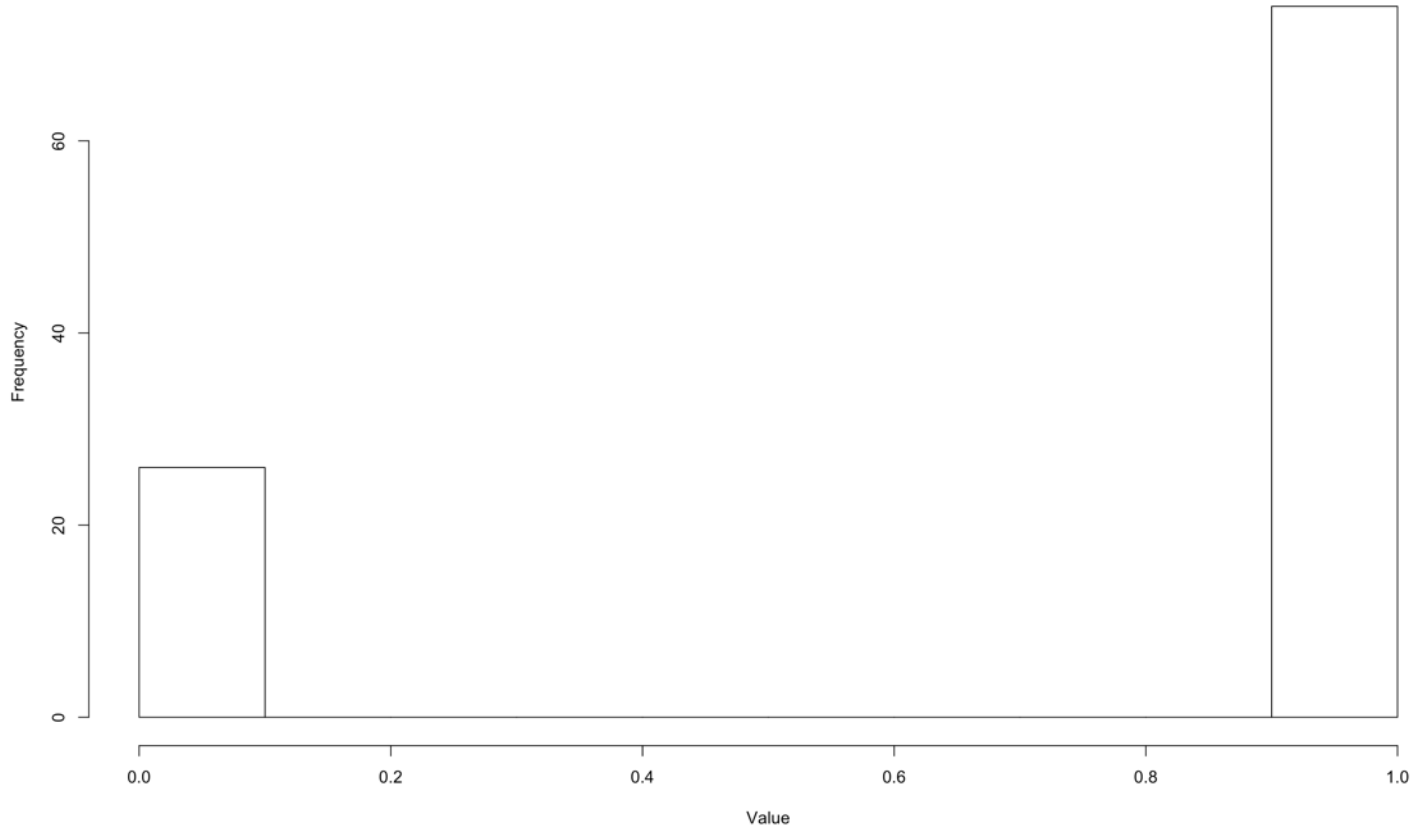




Revisiting Data Distributions- Binomial

- **Constrained** between 0 and 1

- Examples:
  - Proportions
  - Presence/Absence Data



LMs vs GLMS

- Response variable data type
  - Unconstrained vs Constrained

- Model fitting approach
  - Ordinary Least Squares vs Maximum Likelihood Estimation

- Assumptions
  - Means and variance

OLS vs MLE

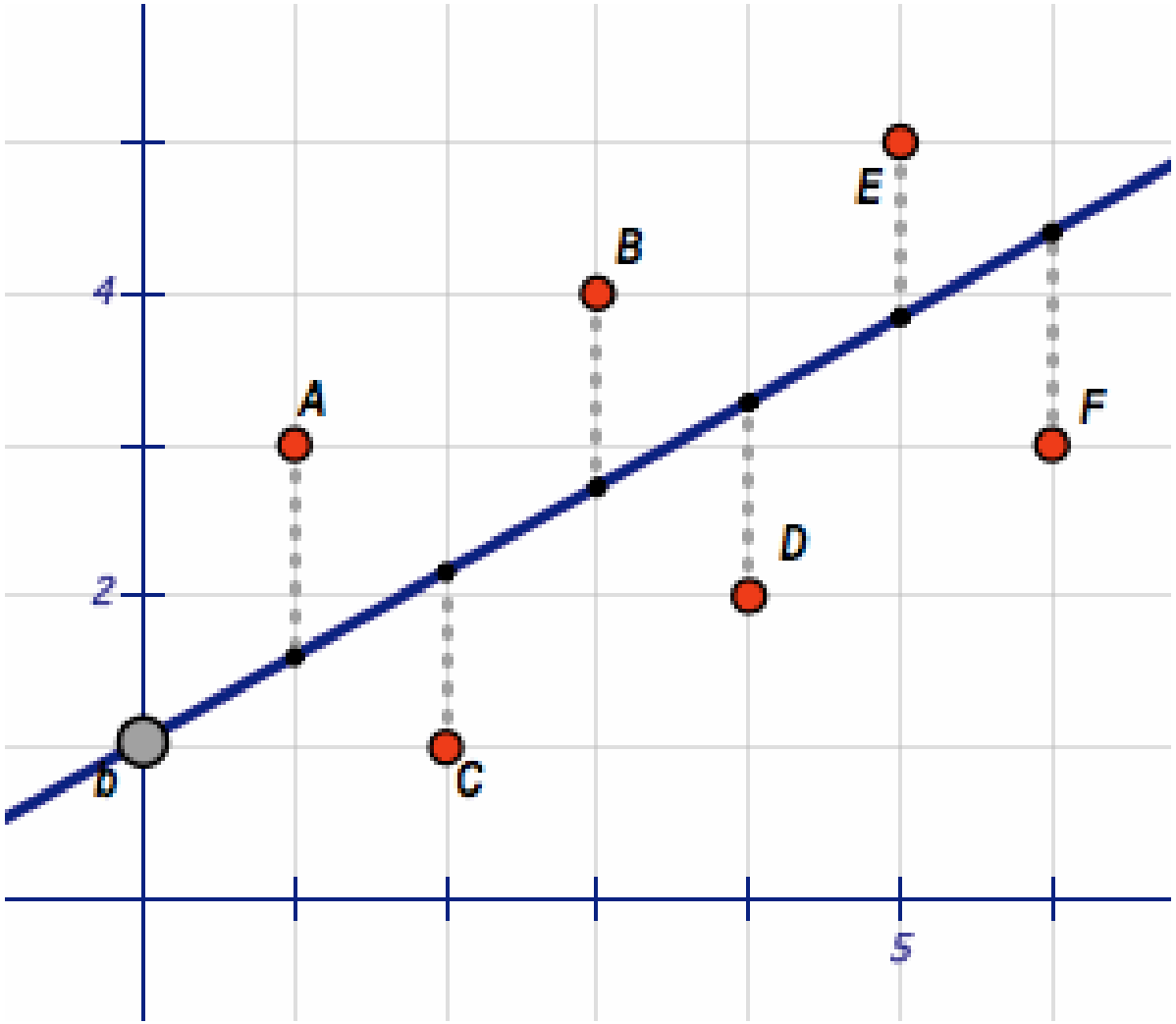
- Model fitting approach
  - Ordinary Least Squares vs Maximum Likelihood Estimation

Fits a line that minimizes the residual sum of squares

Fits a line that maximises the log-likelihood

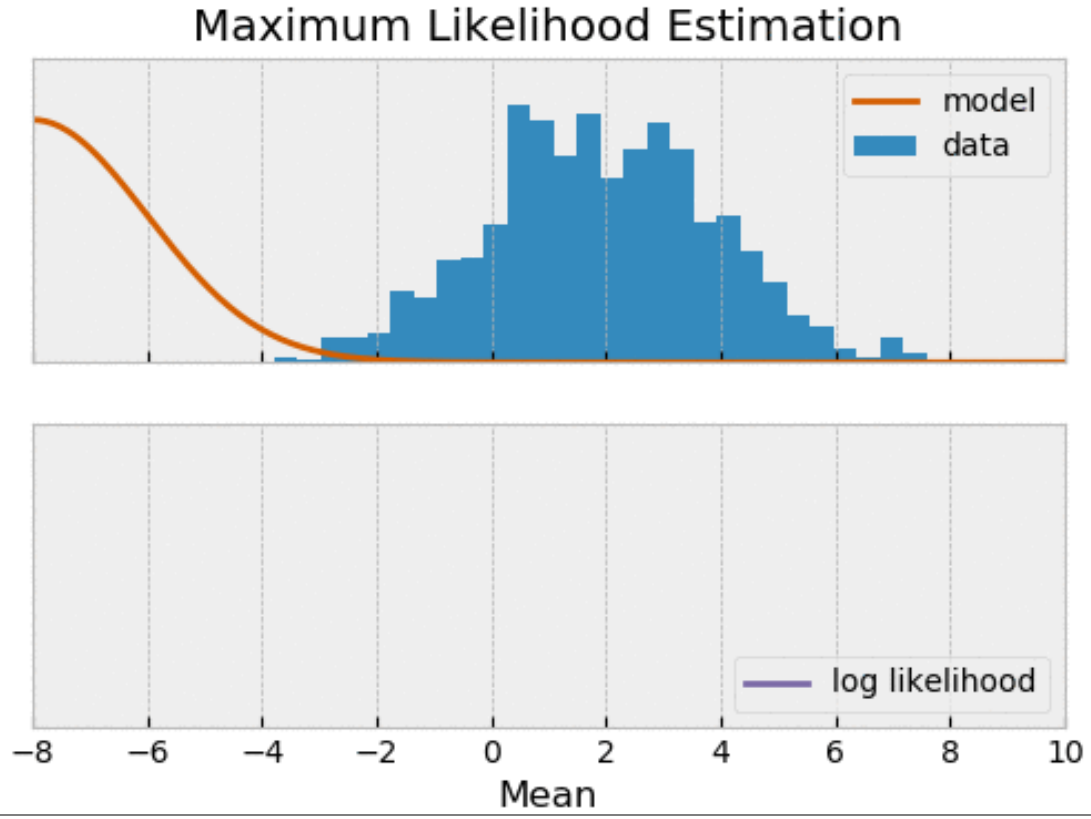
OLS vs MLE

- Model fitting approach
  - Ordinary Least Squares



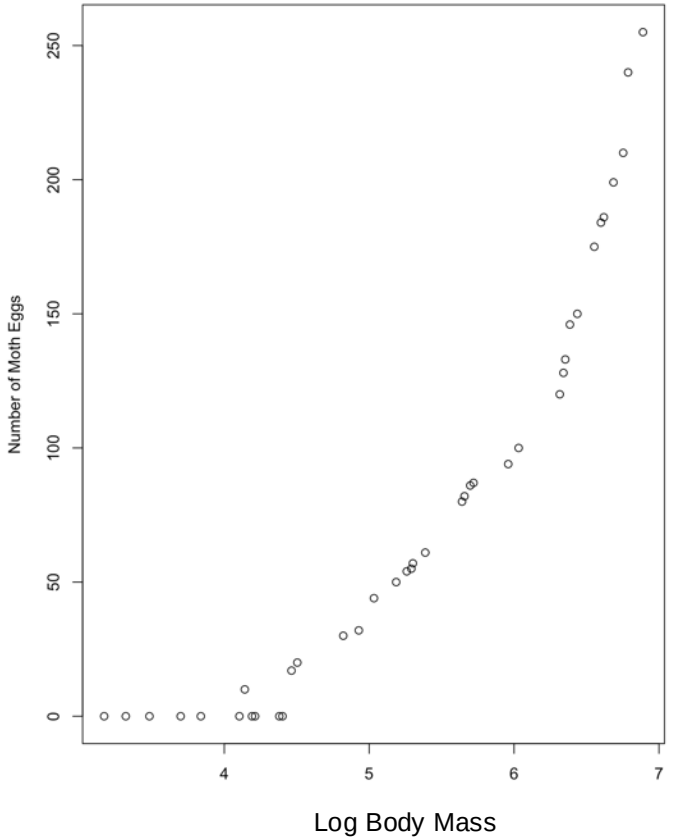
OLS vs MLE

- Model fitting approach
  - Maximum Likelihood Estimation

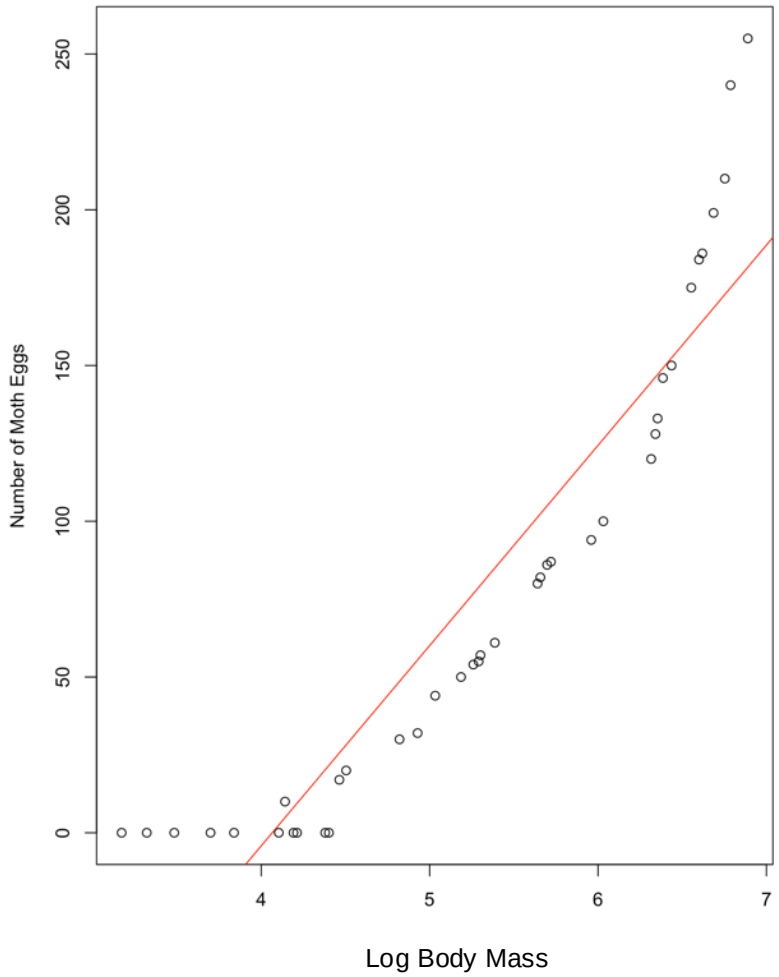


Example

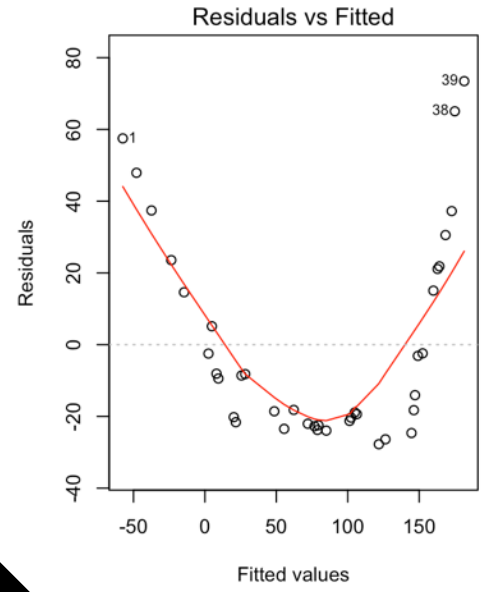
- Number of eggs laid and female body size of vapourer moth

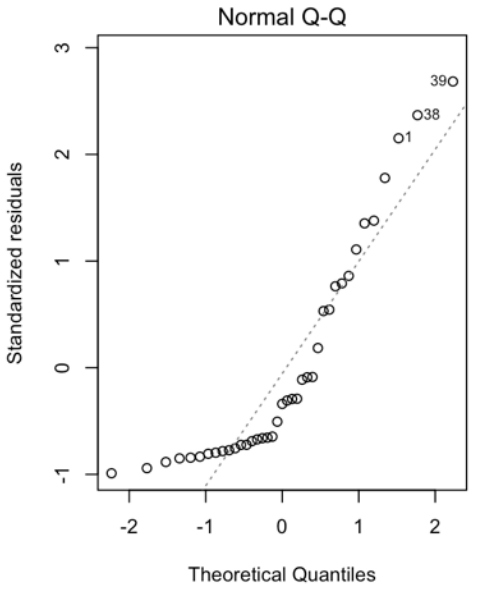


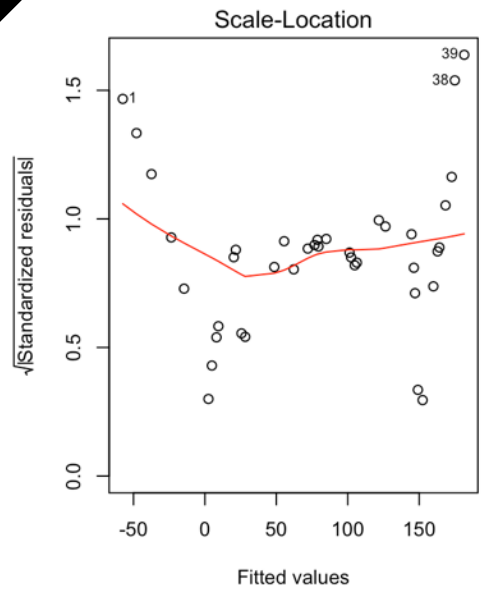
Example

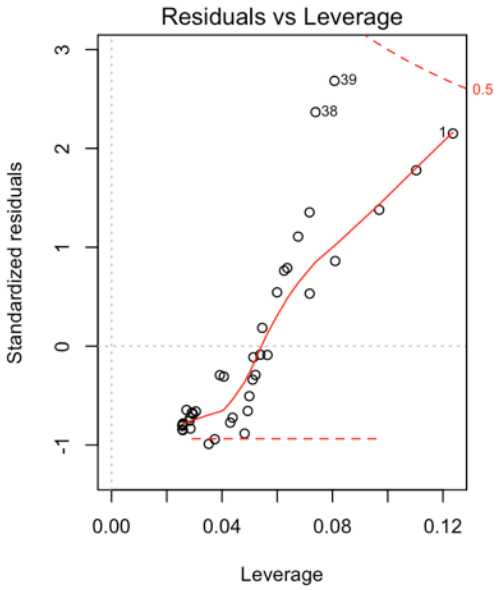


Diagnostics



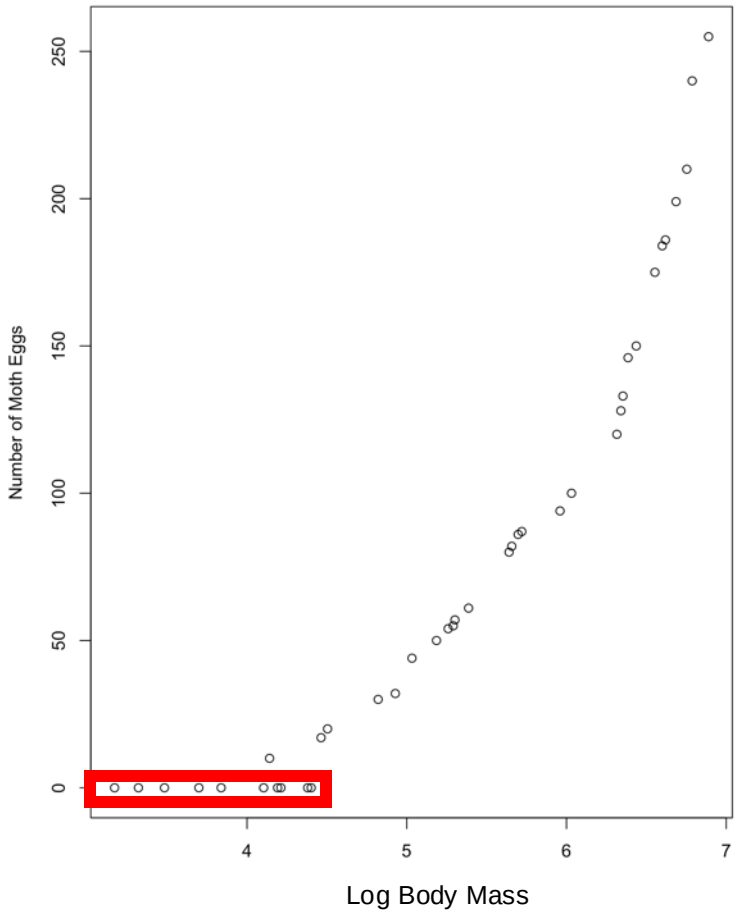




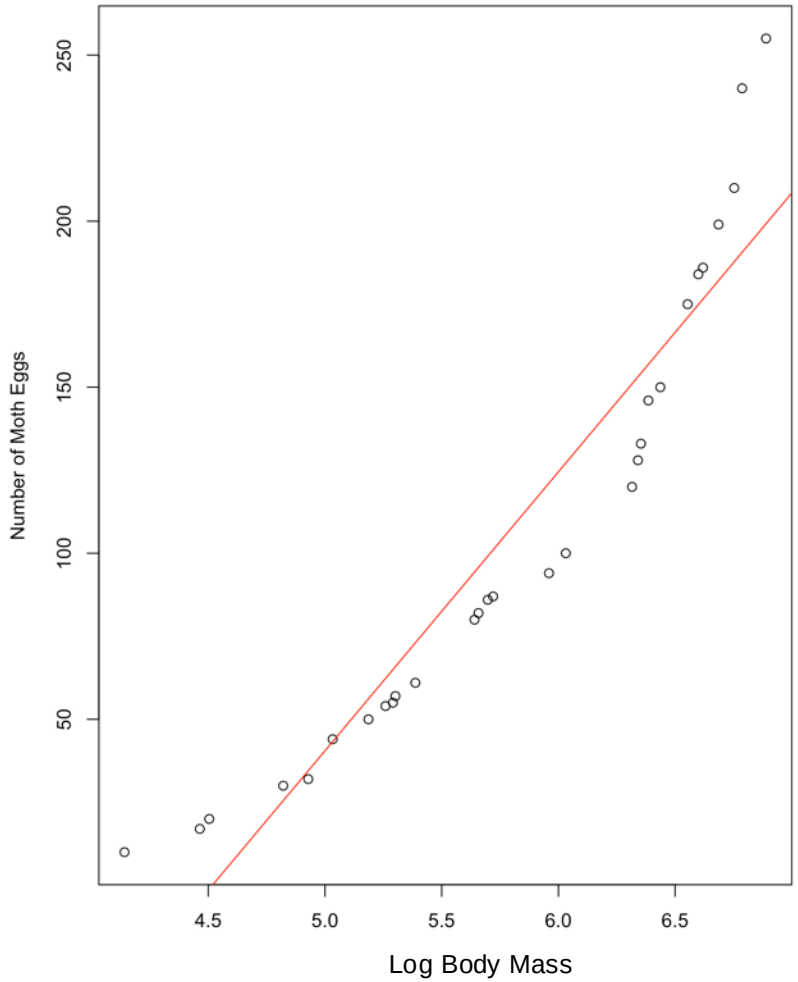


Solutions?

- Log transformations of count data



Lose ~26% of Data



Generalised Linear Models

- Any GLM consists of three steps:

1. Choosing a distribution for the response variable

2. Specifying the linear function of covariates and/or fixed factors

3. Choosing a link between the predictor function and the mean of the distribution

Generalised Linear Models

1. Distribution of response = normal
2. Predictor function = $\beta_0 + \beta_1 x_i + \varepsilon_i$
3. Link between the predictor and the mean of the distribution:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

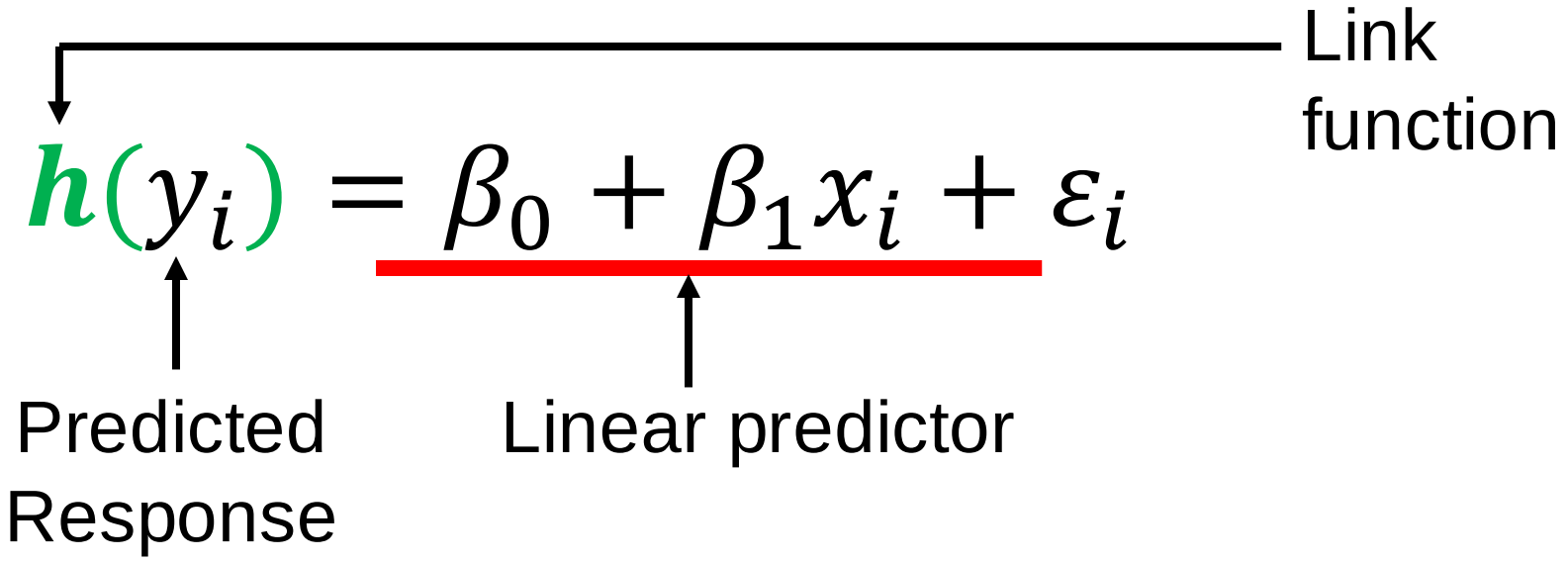
Predicted Response

Linear predictor

Link Functions

- Generalised linear model relates the response distribution to the linear predictor via a link function

- There are different link functions for different distributions



Link Functions

Normal	Poisson (Counts)	Binomial
Leaf Mass	Number of Species	0,1,1,0
Bill Width	Number of Enzymes	Female, Male, Female, Male
Height	Number of Offspring	Absent, Present, Present, Absent
Weight	0, 1, 2, 3, 7	Heads, Heads, Heads, Tails
Identity	**Log-linear (natural)**	**Logit**

Summary

- Generalised linear models estimate the linear estimates via maximum likelihood estimation and are able to handle constrained data types

- The response variable is related to the linear predictor via a link function and these are specific to distribution families

- We first need to establish the distribution of our response variable and the linear predictor. With both of these, we can select an appropriate link function